A most powerful moment-based test for the distribution of random coefficients*

Hippolyte Boucher[†] Gökçe Gökkoca[‡] Max Lesellier[§]

May 1, 2025

Abstract

Random coefficients are a powerful tool to incorporate unobserved heterogeneity in econometric models. While for estimation purposes, it is common to specify a parametric family for the distribution of random coefficients, an inaccurate specification can significantly bias the empirical predictions, especially when studying the distributional effects of a policy. In this paper, we develop a simple yet powerful moment-based specification test for the distribution of random coefficients. The moment conditions chosen for the test are designed to maximize the power of the test when the distribution of random coefficients is misspecified. We show that our test is applicable to a wide class of random coefficient models and can be used to discriminate between competing specifications. We conduct extensive Monte Carlo simulations to assess the performance of our test and we illustrate its empirical relevance by selecting the best-fitting specification for price sensitivity in demand estimation for the German automobile market.

Keywords: specification tests, random coefficients, discrete choice models
JEL codes: C11, C12, C35, C52

*The authors would like to thank Christian Bontemps, Nour Meddahi and Mathias Reynaert for their guidance and support as well as Steve Berry, Pierre Dubois, Jean-Marie Dufour, Eric Gautier, Cristina Gualdani, Koen Jochmans, Vishal Kamat, Pascal Lavergne, and Thierry Magnac for helpful comments and discussions. We also thank conference and seminar participants at the various venues where this paper was presented. Finally, we want to thank Kevin Remmy for providing us with the data set used in the empirical application.

[†]Airbus, hippolyte.boucher.mail@gmail.com
[‡]Analysis Group, gokce.gokkoca@analysisgroup.com
[§]University of Montreal, max.antoine.lesellier@umontreal.ca

1 Introduction

Random coefficients (RCs) are a powerful tool that allow researchers to introduce unobserved heterogeneity in a non-additive and flexible way while retaining the tractability of a parametric model. For instance, the econometrician may use RCs to model preference heterogeneity in discrete choice models or capture disparities in how agents enroll in treatment in models with endogenous selection. Examples of models using random coefficients include the mixed linear model, the mixed logit, and the aggregate random coefficient logit from Berry et al. (1995), commonly known as the BLP demand model. The standard estimation procedure requires the econometrician to specify a parametric family for the distribution of random coefficients and then estimate the parameters associated with this parametric specification. It is now widely recognized that the shape of the distribution of RCs matters greatly for key empirical predictions and policy evaluations, particularly when assessing the distributional effects of a policy. For instance, Miravete et al. (2023) shows that standard specifications chosen to model heterogeneity in the willingness to pay in the BLP demand model imposes strong restrictions on the demand curvature, thereby limiting the range of implied cost pass-through rates. In spite of its importance, the choice of the parametric distribution is, in the vast majority of cases, arbitrary, and can vary significantly from one application to the other.

This paper provides a simple yet powerful specification test for the distribution of random coefficients, applicable across a wide range of models, to mitigate the risk of misspecification. Additionally, we show how our test can serve as the foundation for a formal model selection procedure. Moreover, the general machinery developed in this paper, which maximizes the power of a moment-based test against a known form of misspecification, is applicable across a range of contexts beyond random coefficient models.

The specification test. The starting point for our test is to show that the class of random coefficient models studied in this paper is characterized by a vector of conditional moment restrictions, which are satisfied only when evaluated at the true distribution of RCs. These conditional moment restrictions can be transformed into unconditional ones, which form the basis of our specification test. This paper develops a systematic method for selecting the set of moment conditions, or equivalently the set of instruments, that maximizes the power of our test against potential misspecification in the distribution of RCs. We apply Bahadur's criterion to quantify the power of a test. The construction of these moment conditions proceeds in two steps that we lay out below.

First, we focus on a theoretical scenario in which the econometrician wants to test the specification against a known and fixed alternative. We consider both a simple null hypothesis, $\overline{H}_0 : F = F_0$, and a composite null hypothesis, $H_0 : F \in \mathcal{F}_0$. In each case, we derive the first-best instrument that maximizes the power of our moment-based test. We call this instrument the most powerful instrument (henceforth, MPI). In the composite hypothesis case, we must estimate the unknown parameters of the distribution in the first step. To ensure our test statistic is asymptotically robust to first-stage estimation, we propose a novel orthogonalization strategy based on transforming the instruments. This approach is agnostic about the first-stage estimator and may be relevant beyond the scope of this paper. Moreover, we highlight an important connection between the MPI and the classical optimal instruments used for efficient estimation purposes.

Second, we turn to the general case in which the econometrician remains agnostic about the alternative. We provide a feasible approximation of the MPI that can be derived without the knowledge of the fixed alternative. To derive our feasible approximation we show that the MPI can be expressed as a known function integrated over the unknown distribution of RCs. We propose discretizing the integral as a means to recover a vector of feasible instruments, which we refer to as interval instruments. We theoretically justify this approach by showing that the interval instruments are a linear approximation of the MPI.

In addition to maximizing power against misspecification in the distribution of RCs, our test possesses the following attributes that make it easy to implement and appealing to practitioners. First, the test statistic is asymptotically pivotal under the null hypothesis, which implies that the critical value is straightforward to derive and does not involve resampling methods. Second, our test only needs the residuals of the first-stage estimation and imposes minimal assumptions, demanding only that the estimator be $\sqrt{N}$-consistent under the null hypothesis.

The model selection procedure. The test developed in this paper can be used in applied work to discriminate between different specifications for the distribution of

RCs. We adopt the framework established in Rivers and Vuong (2002), which involves comparing two lack-of-fit criterion functions derived under the competing specifications. We show that the value of our test statistic derived under each specification with the approximation of the MPI can be successfully used as an objective, invariant, and straightforward metric for assessing the validity of the conditional moment restrictions.

Simulation experiments. We evaluate the small-sample properties of our test and selection procedure using a set of Monte Carlo experiments. First, we compare the performance of our test with other available specification tests, such as the classical Sargan-Hansen overidentification test. We show that our test has the correct empirical size and that it significantly outperforms the rest of the tests in terms of power under alternative distributions. A second set of Monte Carlo experiments indicates that our model selection procedure reliably selects the correct specification.

Empirical application. We estimate a demand system for automobiles in Germany over the period 2012 to 2018. We estimate the model under different specifications for the distribution of the RC associated with price. We show that different specifications lead to substantially different results for the key empirical predictions of the model. In line with Miravete et al. (2023), our results indicate that the degenerate and normal specifications yield lower demand curvature and pass-through rates. Additionally, we find these specifications also result in lower price elasticities. Next, we apply our test and selection procedure to identify the specification best supported by the data, and we find that all the specifications are rejected except for the triweight, a kernel with cubic decay.

Related literature. Our paper contributes to several strands of the literature. First, it relates to the extensive body of work on specification tests, including Sargan (1958), Hausman (1978), Hansen (1982), Bierens (1982), Newey (1985), Newey and McFadden (1994), Godfrey (1996), Zheng (1996), Bontemps and Meddahi (2005), and Huber and Ronchetti (2009). More specifically, it contributes to the literature on tests targeting the distribution of random coefficients (RCs), such as McFadden and Train (2000) and Fosgerau and Bierlaire (2007). These two papers aim to detect the presence of additional heterogeneity in the mixed logit model.[1] In comparison, our test avoids

[1]McFadden and Train (2000) propose a Lagrange Multiplier test to detect additional heterogeneity in the mixed logit model. Similarly, Fosgerau and Bierlaire (2007) develop a likelihood ratio test based

additional estimation, maximizes statistical power, and applies broadly to models with random coefficients beyond the mixed logit framework.

Second, our paper contributes to the literature on the comparison of tests, including Neyman and Pearson (1933), Bahadur (1960), Geweke (1981), Lehmann et al. (1986), Gourieroux and Monfort (1995). We show that classical optimality criteria from this literature can guide the choice of the moment conditions in the context of a moment-based test.

Third, we contribute to the literature on orthogonalization methods designed to make test statistics robust to the statistical noise arising from first-stage estimation, such as Neyman (1959), Bontemps and Meddahi (2005), and Chernozhukov et al. (2018). Our approach is widely applicable and consists of orthogonalizing the instruments rather than the unconditional moments directly. Finally, we contribute to the literature on test-based model selection: Cox (1961), Davidson and MacKinnon (1981), Vuong (1989), Rivers and Vuong (2002), Smith (1992), Hall and Pelletier (2011). In particular, we advocate for the use of our test statistic as a goodness-of-fit criterion within the Rivers and Vuong (2002) framework.

Structure of the paper. The remainder of the paper is organized as follows. In Section 2, we present the class of random coefficient models studied in this paper and provide several examples of models that fit our framework. In Section 3, we present our specification test and derive the most powerful instrument for both the simple and the composite hypothesis cases. In Section 4, we provide a feasible approximation of the MPI and we discuss the theoretical properties of this approximation. In Section 5, we show how our test can be used to discriminate between different model specifications for the distribution of RCs. In Section 6, we conduct Monte Carlo simulations to evaluate the finite sample performance of our test and model selection procedure. In Section 7, we present our empirical application. Finally, we conclude the paper in Section 8.

2 Setup, examples, and identification

First, let us introduce the general setting that we study in this paper. We observe an i.i.d. sample $\{W_i, Z_i\}_{i=1}^N$, where $Z_i \in \mathcal{Z} \subset \mathbb{R}^{d_Z}$ denotes a vector of exogenous variables

on estimating both the null and a more flexible alternative specification using Legendre polynomials, though they highlight the practical difficulties of estimating the latter.

distributed according to a probability distribution P, and W_i is a vector of random variables that may include a subset of the components of Z_i, as well as the dependent variables $Y_i \in \mathcal{Y} \subset \mathbb{R}^{d_Y}$. We consider a generic model in which, at the true parameter value $\theta = (\beta, F)$, the following p conditional moment restrictions hold:

$$\mathbb{E}[m(W_i, \theta)|Z_i] = 0_p \ a.s. \tag{2.1}$$

where $m : \mathbb{R}^{d_W} \times \Theta \to \mathbb{R}^p$ is a known measurable function. For the sake of exposition, we remove the dependence of the expectation in P. The parameter of interest θ consists of a finite dimension parameter β and an infinite dimensional parameter F, which corresponds to the distribution of random coefficients. The objective of the econometrician is to estimate $\theta = (\beta, F)$. In practice, the non-parametric estimation can be extremely challenging (see Compiani (2022), in the BLP demand model), leading researchers to adopt parametric assumptions of the form $F \in \mathcal{F}_0 = \{F_0(\cdot|\lambda) : \lambda \in \Lambda_0\}$. Nonetheless, these parametric assumptions often lack grounding in economic theory, potentially distorting the model's predictions if they fail to accurately represent the underlying shape of unobserved heterogeneity. Therefore, the objective of the paper is to construct a simple and powerful test to test a hypothesis of the form:

$$H_0 : F \in \mathcal{F}_0.$$

Before introducing our test and exploring its theoretical properties, we first present several popular models that align with our framework.

Example 1: the mixed linear model. First, we consider the mixed linear model, which is particularly well-suited to model clustered data or heterogeneous treatment effects. In particular, it allows for the estimation of not only the average treatment effect, but also higher moments and quantiles of the treatment effect. Let Y_i denote the dependent variable and X_i the explanatory variables for individual $i = 1, \ldots, N$.

$$Y_i = X_{1i}'\beta + X_{2i}'v_i + \varepsilon_i \quad \text{with} \quad \mathbb{E}(\varepsilon_i|X_i) = 0 \ \ a.s. \ ,$$

$(X_i, \varepsilon_i) \perp\!\!\!\perp v_i$ where the RC v_i has distribution F and ε_i has a known distribution G_ε.[2] The econometrician aims to estimate $\theta = (\beta, F)$ and typically relies on classical methods such as maximum likelihood estimation, which require parametric assumptions on F. For instance, it is common to impose that v_i is normally distributed or that it follows a categorical distribution.

In this example, $W_i = (Y_i, X_i)$ and $Z_i = X_i$. For a given vector $b \in \mathbb{R}^p$, we construct the following function m_1:

$$m_1 : (W_i, \tilde{\theta}) \mapsto 1\{Y_i \leq b\} - \int G_\varepsilon(b - X'_{1i}\tilde{\beta} - X'_{2i}v)d\tilde{F}(v).$$

One can easily verify that at the true parameter $\theta = (\beta, F)$, the function m_1 satisfies (2.1):

$$\mathbb{E}[m_1(W_i, \theta)|X_i] = \mathbb{E}[1\{Y_i \leq b\}|X_i] - \int G_\varepsilon(b - X'_{1i}\beta - X'_{2i}v)dF(v) = 0_p \ a.s..$$

Example 2: the mixed logit model. Introduced by Boyd and Mellman (1980) and Cardell and Dunbar (1980), the mixed logit model has become a cornerstone in the discrete choice literature. It offers flexibility, computational tractability, and more realistic substitution patterns than the classical multinomial logit by relaxing the independence of irrelevant alternatives assumption.[3]

The researcher observes the choices made by individuals $i = 1, \ldots, N$ among $j = 0, \ldots, J$ options (where $j = 0$ corresponds to the outside option). Let U_{ij} be the indirect utility stemming from the choice of option $j \neq 0$ by individual i:

$$U_{ij} = X'_{1ij}\beta + X'_{2ij}v_i + \varepsilon_{ij}.$$

X_{ij} are observed covariates that characterize choice j for individual i (for instance, the distance to school j in a school choice context), v_i is the RC that captures preference

[2] In the statistical literature, RCs are also known as random effects. For a comprehensive review, see McCulloch and Searle (2000)

[3] As shown by McFadden and Train (2000), the mixed logit model is sufficiently flexible to approximate any choice probabilities generated by a random utility maximization process arbitrarily closely.

heterogeneity for the characteristics included in X_{2ij}. Again, $v_i \sim F$ and is assumed to be independent of all the other variables. Finally, ε_{ij} is an idiosyncratic taste shock that is distributed as extreme value type one (EV1) and independent from all other variables and across (i, j).

As previously, the parameter of interest is $\theta = (\beta, F)$. Each individual i chooses the option j that yields the highest indirect utility U_{ij} (i.e. $U_{ij} > U_{ik}\ \forall k \neq j$). Let Y_{ij} be the decision by i to choose j. We have:

$$\mathbb{P}_\theta(Y_{ij} = 1|X_i) = \mathbb{P}_\theta(U_{ij} \geqslant U_{ik}\ \forall k\ |X_i) = \int \frac{\exp\{X'_{1ij}\beta + X'_{2ij}v\}}{1 + \sum_{l=1}^{J} \exp(X'_{1il}\beta + X'_{2il}v)} dF(v)$$

where the second equality follows from the properties of the EV1 distribution and the independence between v_i and X_i. The model is typically estimated by Simulated Maximum Likelihood (SML) or Simulated Method of Moments (SMM) (see Train (2001)), with parametric assumptions on F, often normal.

This setup fits the general model defined previously. In this example, the unit of observation is the individual and $W_i = (Y_i, X_i)$, $Z_i = X_i$ with $Y_i = (Y_{i1}, ..., Y_{iJ})'$ and $X_i = (X_{i1}, ..., X_{iJ})'$. Moreover, we define the following function m_2 in $\mathbb{R}^J$:

$$m_2 : (W_i, \tilde{\theta}) \mapsto Y_i - \rho(X_i, \tilde{\theta}) \text{ with } \rho(\cdot, \tilde{\theta}) :\ X_i \mapsto \int \frac{\exp\{X_{1i}\tilde{\beta} + X_{2i}v\}}{1 + \sum_{l=1}^{J} \exp\{X'_{1il}\tilde{\beta} + X'_{2il}v\}} d\tilde{F}(v).$$

By construction, at the true parameter θ, equation (2.1) is satisfied:

$$\mathbb{E}[m_2(W_i, \theta)|Z_i] = \mathbb{E}[Y_i|X_i] - \rho(X_i, \theta) = 0_J\ a.s..$$

The function m_2 is also known as the generalized residuals.

Example 3: the BLP demand model. The differentiated product demand model initiated by Berry (1994) and Berry et al. (1995) is the standard for estimating demand functions in markets with differentiated products when only macro-level data (market shares) are observed. It accounts for price endogeneity and preference heterogeneity. In the standard setting, there are T markets indexed by t and J_t products indexed by j. Each consumer i chooses a product $j \in \{0, 1, ..., J_t\}$. Product j is characterized by a vector of characteristics X_{jt}. Consumer i derives an indirect utility U_{ijt} from purchasing good $j \neq 0$ in market t:

$$U_{ijt} = X'_{1jt}\beta + \xi_{jt} + X'_{2jt}v_i + \varepsilon_{ijt}. \tag{2.2}$$

Preferences for X_{1jt} are homogeneous, while those for X_{2jt} vary through random coefficients distributed as F. ε_{ijt} is a preference shock that follows an EV1 distribution and ξ_{jt} is an unobserved demand shock on product j in market t. As previously, $v_i \perp\!\!\!\perp (X_t, \xi_t, \{\varepsilon_{ijt}\}_{j=1,...J_t})$. Each consumer i chooses the option j that yields the highest indirect utility U_{ijt}. Let Y_{ijt} the decision to choose $j \neq 0$. We have:

$$\mathbb{P}_\theta(Y_{ijt} = 1 | X_t, \xi_t) = \mathbb{P}_\theta(U_{ijt} \geq U_{ikt} \ \forall k | X_t, \xi_t) = \int_{\mathbb{R}^{K_2}} \frac{\exp\left\{X'_{1jt}\beta + \xi_{jt} + X'_{2jt}v\right\}}{1 + \sum_{k=1}^{J} \exp\left\{X'_{1kt}\beta + \xi_{kt} + X'_{2kt}v\right\}} dF(v). \tag{2.3}$$

The second equality follows from the EV1 assumption on the idiosyncratic shocks. As previously, $\theta = (\beta, F)$ is the parameter of interest. Furthermore, a classical exogeneity condition is usually assumed on the demand shock: $\forall j \ \ \mathbb{E}[\xi_{jt}|Z_t] = 0 \ a.s.$, with Z_t a set of exogenous variables. The estimation is done via non-linear GMM. Again, a parametric assumption must be made on F for estimation.

The BLP demand model also fits within our general framework. The unit of observation is the market t. Let Y_t denote the observed market shares, and define $W_t = (Y_t, X_t)$ and $Z_t = \{Z_{jt}\}_{j=1,...,J}$. One can then define the function m_3 in $\mathbb{R}^J$ as:

$$m_3 : (W_t, \tilde{\theta}) \mapsto \rho^{-1}(Y_t, X_{2t}, \tilde{F}) - X_{1t}\tilde{\beta}$$

where $\rho^{-1}(Y_t, X_{2t}, \tilde{F})$ is the unique solution $\tilde{\delta}$ to the following system of J equations:

$$Y_t = \rho(\tilde{\delta}, X_{2t}, \tilde{F}) \ \ \text{with} \quad \rho(\cdot, X_{2t}, \tilde{F}) : \delta \mapsto \int_{\mathbb{R}^{K_2}} \frac{\exp\left\{\delta + X_{2t}v\right\}}{1 + \sum_{k=1}^{J} \exp\left\{\delta_k + X'_{2tk}v\right\}} d\tilde{F}(v)dv.$$

Berry (1994) shows that for any (Y_t, X_{2t}) and for any distribution of random coefficients $\tilde{F}$, there exists a unique solution to the previous system of equations. In words, the inverse demand function is the only mean utility $\tilde{\delta}$ that can generate the observed market shares given $\tilde{F}$. Unfortunately, there is no closed-form expression for the inverse demand function, which must be recovered numerically using a fixed-point algorithm. By construction, we have that: $\xi_t = \rho^{-1}(Y_t, X_{2t}, F) - X_{1t}\beta$. Thus, given the exogeneity assumption on the demand shock ξ_t, we have that:

$$\mathbb{E}[m_3(W_t, \theta)|Z_t] = 0_J \ a.s..$$

Other examples. Dynamic discrete choice models, such as the one pioneered by Rust (1987) and expanded by Hotz and Miller (1993) and Arcidiacono and Miller (2011), use random coefficients to model permanent unobserved heterogeneity between individuals.[4] These models are commonly used to study schooling investment or labor participation decisions. For constructing the function m in a generic dynamic discrete choice model, see Fox et al. (2011). Furthermore, RCs are a natural way to relax the monotonocity assumption in modeling selection into treatment (Heckman and Vytlacil (2005)).[5] The residual function m in the selection stage can be derived as in Example 1.

Identification. For each of the random coefficient models presented above, we have exhibited a function m such that the conditional moment restriction in (2.1) is satisfied when evaluated at the true parameter θ. For our test to be able to discriminate between various specifications, it is critical for m to also be informative about θ. Ideally, in the best-case scenario, m should identify θ non-parametrically.

Definition 1 *We say that m identifies non-parametrically the parameters β and the distribution of random coefficients F if:*

$$(\tilde{\beta}, \tilde{F}) = (\beta, F) \iff \qquad \mathbb{E}[m(W_i, \tilde{\theta})|Z_i] = 0_p \quad a.s..$$

For Examples 1 and 2, the literature has established sufficient conditions under which the functions m nonparametrically identify the parameter of interest θ.[6] In

[4] Traditionally, these random coefficients are discrete but recent work has explored continuous random coefficients (eg: Bunting (2022)).

[5] For policies with heterogeneous treatment effects, the treatment parameter has a Local Average Treatment Effect (LATE) interpretation when instruments shift participation in only one direction (Angrist and Imbens (1995)). Vytlacil (2002) shows that this assumption is equivalent to a selection equation with a latent index structure.

[6] In the linear mixed model (Example 1), sufficient conditions for identification of random coefficients with discrete distributions and with continuous distributions can respectively be found in Gao and Pesaran (2023) and Fox et al. (2016); in the mixed logit (Example 2), sufficient conditions for the identification of θ are derived in Fox et al. (2012).

Appendix A.1.1, we derive a set of sufficient conditions elucidating the identification of θ by the function m_3 in the BLP demand model (Example 3). Our identification proof leverages a recent and important identification result in Wang (2022).

The fact that m identifies θ has two key implications. First, it entails that conditional moment restriction in (2.1) can be used to consistently estimate the parameter θ by GMM (or by the simulated method of moments). Second, it implies that the conditional moment restrictions in (2.1) can be used as the basis for a specification test, and in particular, it is possible to construct a consistent test for the distribution of random coefficients using this restriction. In this paper, we focus on the second implication. However, it is worth noting that the test studied in this paper remains valid even if θ is not nonparametrically identified.

3 The infeasible most powerful instrument

The econometrician aims to test $H_0 : F \in \mathcal{F}_0$ against a known and fixed alternative $\overline{H}_a : (\beta, F) = (\beta_a, F_a)$. Within this framework, we construct the most powerful moment-based test. The intuition for our test is simple: if the model under H_0 is misspecified, then, assuming that m identifies θ, the conditional moment restriction in (2.1) will no longer be satisfied. Our objective is to find the "best instrument" (or equivalently the "best moment") to pin down this violation of (2.1). We refer to this instrument as the most powerful instrument (MPI). It is in general infeasible as the econometrician does not know the alternative (or does not want to specify it), nonetheless, it provides a useful first-best reference point. In Section 4, we provide a feasible approximation to it. We derive the MPI in two different cases: first, we start with the simple hypothesis case where the econometrician aims to test $\overline{H}_0 : (\beta, F) = (\beta_0, F_0)$ and then we move to the composite hypothesis case $H_0 : F \in \mathcal{F}_0$. Finally, we show how the MPI relates to the classical optimal instruments, derived for efficient estimation purposes. In the remainder of the paper, $\|\cdot\|_2$ denotes the L_2 norm.

3.1 Simple hypothesis

We first focus on the simplest possible case in which the econometrician wants to test a simple hypothesis of the form $\overline{H}_0 : (\beta, F) = (\beta_0, F_0)$. 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 3.2. First, we introduce a moment-based test for $\overline{H}_0$ and we show its asymptotic validity.

3.1.1 A classical moment-based test

Our approach to testing $\overline{H}_0 : (\beta, F) = (\beta_0, F_0)$ relies on a standard moment-based test. We transform the p conditional moment restrictions in (2.1) into L unconditional moment restrictions. To do so, we introduce a family of L functions of Z_i, denoted as $(h_l)_{l=1,\dots,L}$, such that for all l, $h_l : \mathcal{Z} \to \mathbb{R}^p$. The transformed variables $h_l(Z_i)$ are referred to as instruments, and we represent the horizontal stacking of the L vectors $h_l(Z_i)$ as $h(Z_i) \in \mathcal{M}_{L,p}(\mathbb{R})$.[7] For any testing instrument h, equation (2.1) implies:

$$\overline{H}_0 : (\beta, F) = (\beta_0, F_0) \implies \overline{H}'_0 :\ \mathbb{E}\left[h(Z_i)m(W_i, \theta_0)\right] = 0_L.$$

We propose to test $\overline{H}_0$ indirectly through its implication $\overline{H}'_0$, which is a set of unconditional moment conditions. The test statistic for $\overline{H}'_0$ is a squared norm of the empirical moments and is written as follows:

$$S_N(h, \theta_0) = N \left(\frac{1}{N} \sum_{i=1}^{N} h(Z_i)m(W_i, \theta_0) \right)^T \hat{\Omega}_0^{-1} \left(\frac{1}{N} \sum_{i=1}^{N} h(Z_i)m(W_i, \theta_0) \right),$$

with $\hat{\Omega}_0$ the empirical counterpart of $\Omega_0 = \mathbb{E}\left[h(Z_i)m(W_i, \theta_0)m(W_i, \theta_0)^T h(Z_i)^T\right]$. Next, we study the asymptotic properties of our test as N goes to infinity. For our test statistic to have a known asymptotic distribution under $\overline{H}'_0$, we make the following regularity assumptions on $\{h_l(\cdot)\}_{l=1,\dots,L}$.

Assumption 1 (Regularity conditions on the instruments) *The instrument h belongs to the class $\mathcal{H}_0$ if, for any true parameter θ, they satisfy:*

1. *$\exists C$ such that $\mathbb{E}[\|h(Z_i)m(W_i, \theta_0)\|_2^2] < C$,*
2. *$\Omega_0 \equiv \mathbb{E}[h(Z_i)m(W_i, \theta_0)m(W_i, \theta_0)^T h(Z_i)^T]$ is full rank.*

Namely, we assume that the product between the testing instruments and the moment function possesses a fully-ranked and bounded moment of order 2.[8] Here, the

[7] The structure of h offers considerable flexibility. Specifically, it does not require uniformity in the number of instruments for each component of m, as certain components of h can be set to zero.

[8] Note that for the validity of our test, it is sufficient for the conditions to hold at $\theta = \theta_0$. Nevertheless, extending the assumption to cases where $\theta \neq \theta_0$ simplifies the analysis.

class of valid transformations is indexed by 0 to emphasize that its composition depends on the value of the tested parameter θ_0.[9] The following proposition establishes that if the testing instruments are functions of Z_i then, under the previously outlined regularity conditions, our test procedure is asymptotically valid for $\overline{H}_0$ and consistent under $\overline{H}'_a : \mathbb{E}[h(Z_i)m(W_i, \theta_0)] \neq 0_L$.

Proposition 3.1 *Under the assumption that $h \in \mathcal{H}_0$, the following properties hold:*

- *Under* $\overline{H}_0 : (\beta, F) = (\beta_0, F_0), \qquad S_N(h, \theta_0) \underset{N\to+\infty}{\overset{d}{\longrightarrow}} \chi^2_L,$
- *Under* $\overline{H}'_a :\ \mathbb{E}\left[h(Z_i)m(W_i, \theta_0)\right] \neq 0, \quad \forall q > 0, \quad \mathbb{P}(S_N(h, \theta_0) > q) \underset{N\to+\infty}{\longrightarrow} 1.$

We test $\overline{H}_0$ via its implication $\overline{H}'_0 : \mathbb{E}\left[h(Z_i)m(W_i, \theta_0)\right] = 0_L$. Consequently, the power and consistency of our test under the alternative $\overline{H}_a$ depend critically on the choice of the testing instrument $h(Z_i)$. The next subsection is devoted to studying this aspect in detail.

3.1.2 The most powerful instrument (MPI)

Power criterion. We define the most powerful instrument by employing Bahadur's non-local approach (see Bahadur (1960)), which recommends selecting the test that minimizes the significance level α (or equivalently, the probability of a Type I error) required to attain a given power against a fixed alternative for a specified sample size. In Appendix A.2, we discuss the primary alternative for evaluating the power of a test, based on local alternatives, and explain why Bahadur's approach is better suited to the testing problem at hand.

A common implementation of Bahadur's non-local approach is to compare the asymptotic slopes of the tests. In Appendix A.2, we review this method and provide a formal interpretation of the asymptotic slope. Since our test statistic follows a chi-square distribution under $\overline{H}_0$, it follows from Geweke (1981) that the asymptotic slope for the test with instrument $h(Z_i)$ is equal to:

$$c(h, \theta_0) = \operatorname{plim} \frac{1}{N} S_N(h, \theta_0) = \mathbb{E}\left[h(Z_i)m(W_i, \theta_0)\right]^T \Omega_0^{-1} \mathbb{E}\left[h(Z_i)m(W_i, \theta_0)\right]. \qquad (3.4)$$

[9] For greater rigor, we could have indexed it by the data generating distribution P, but we refrain from doing so to simplify the exposition.

A few remarks are in order. First, under $\overline{H}_0$, the slope is naturally equal to 0 for all valid instruments $h \in \mathcal{H}_0$. Second, under $\overline{H}_a$, the asymptotic slope quantifies the rate at which the test statistic diverges, expressed in terms of population moments: divergence $\approx N \times c(h, \theta_0)$. Third, in Section 4.2, we show that asymptotic slope of the test associated with $h(Z_i)$ can also serve as a metric to evaluate how effectively the space spanned by $h(Z_i)^T$ predicts the "residuals" $m(W_i, \theta_0)$ under the alternative $\overline{H}_a$. In the next proposition, we derive an analytical expression for the instrument that maximizes the slope of the test under a fixed alternative $\overline{H}_a$.

The most powerful instrument. To construct the MPI, we use the following decomposition of m under $\overline{H}_a$:

$$m(W_i, \theta_0) = \underbrace{m(W_i, \theta_a)}_{\text{true residual under } \overline{H}_a} + \underbrace{m(W_i, \theta_0) - m(W_i, \theta_a)}_{\Delta^m_{0,a}(W_i)},$$

with $\Delta^m_{0,a}(W_i)$ being the correction term due to misspecification under the alternative $\overline{H}_a$. Now let us provide an expression for the most powerful instrument.

Proposition 3.2 (Most Powerful Instrument)
Under a fixed alternative $\overline{H}_a : (\beta, F) = (\beta_a, F_a) \neq (\beta_0, F_0)$, we have the following:

$$h^*_a(Z_i) \equiv \left(\mathbb{E}\left[m(W_i, \theta_0) m(W_i, \theta_0)^T | Z_i\right]^{-1} \mathbb{E}[\Delta^m_{0,a}(W_i) | Z_i]\right)^T \in \underset{h \in \mathcal{H}_0}{\operatorname{argmax}} \;\; c(h, \theta_0).$$

The proof is given in Appendix B.2.1. Intuitively, the MPI fully captures the exogenous variation contained in the correction term implied by the misspecification. Under $\overline{H}_a$, observe that the MPI equals $\tilde{h}(Z_i) \equiv \left(\mathbb{E}\left[m(W_i, \theta_0) m(W_i, \theta_0)^T | Z_i\right]^{-1} \mathbb{E}[m(W_i, \theta_0) | Z_i]\right)^T$. However, $\tilde{h} \notin \mathcal{H}_0$ because $\tilde{h}$ is equal to 0 under H_0. Furthermore, the expression in Proposition 3.2 is far more convenient to work with, as it leverages knowledge about the nature of the misspecification. A key property of the MPI h^*_a is its ability to achieve consistency under a fixed alternative $\overline{H}_a : (\beta, F) = (\beta_a, F_a) \neq (\beta_0, F_0)$, provided that m identifies θ. This is formalized in the following result.

Proposition 3.3 (Consistency of the test with the MPI) *Denote h_a^* the MPI associated with $\overline{H}_a$. Under the assumption that m identifies θ and Assumption 1, we have:*

$$\overline{H}_a : (\beta, F) = (\beta_a, F_a) \neq (\beta_0, F_0) \implies \forall q \in \mathbb{R}^+, \quad \mathbb{P}(S_N(h_a^*, \theta_0) > q) \underset{N \to +\infty}{\longrightarrow} 1.$$

The proof of this result is given in Appendix B.2.1. In the models presented previously, the most powerful instrument can be expressed as follows.

Example 1 continued.

$$h_{1,a}^*(Z_i)^T = \left(\mathbb{E}\left[m_1(W_i, \theta_0) m_1(W_i, \theta_0)^T | X_i\right]\right)^{-1} \left[\int G_\varepsilon(b - X_{1i}'\beta_0 - X_{2i}'v) dF_0(v) - \int G_\varepsilon(b - X_{1i}'\beta_a - X_{2i}'v) dF_a(v)\right]$$

Example 2 continued.

$$h_{2,a}^*(Z_i)^T = \mathbb{E}\left[m_2(W_i, \theta_0) m_2(W_i, \theta_0)^T | X_i\right]^{-1} \left[\rho(X_i; \theta_0) - \rho(X_i, \theta_a)\right]$$

Example 3 continued.

$$h_{3,a}^*(Z_t)^T = \mathbb{E}\left[m_3(W_t, \theta_0) m_3(W_t, \theta_0)^T | Z_t\right]^{-1} \mathbb{E}\left[X_{1t}(\beta_a - \beta_0) + \rho^{-1}(W_t, F_0) - \rho^{-1}(W_t, F_a) | Z_t\right]$$

We conclude this section with two important observations. First, the expressions for the MPI in the examples listed above can already be useful if the researcher has a specific alternative in mind and wishes to test against this alternative. Second, the concept of MPI extends well beyond the scope of this paper and may find applications in a variety of contexts, such as the estimation of structural instrumental variable models.

3.2 Composite hypothesis

We now turn our attention to the composite hypothesis, which is of the form $H_0 : F \in \mathcal{F}_0$ where $\mathcal{F}_0$ denotes a parametric family in which each distribution is uniquely characterized by a parameter λ (i.e. $\mathcal{F}_0 = \{F_0(\cdot|\lambda) : \lambda \in \Lambda_0\}$). The composite hypothesis is in general the most relevant one for practitioners as it corresponds to the scenario in which the researcher seeks to test the specification used to estimate the model. Our test for the composite hypothesis proceeds in two steps. We first estimate the model under the hypothesis H_0 for the distribution of RCs and then we test the validity of a set of unconditional moments evaluated at the estimated parameter. In contrast to the simple hypothesis test examined earlier, we must now estimate the parameter $\phi_0 = (\beta_0^T, \lambda_0^T)^T$ in the first stage, which generates parameter uncertainty. To address this challenge, we introduce a novel approach that involves transforming the instruments, thereby ensuring that the tested empirical moments are asymptotically insensitive to the estimation of ϕ_0. This approach is applicable across a broad spectrum of estimators, including the estimators obtained via maximum likelihood and GMM. Furthermore, we provide an explicit expression for the MPI within the space of instruments that are orthogonal to vanishing perturbations of $m(W_i, \phi)$ around ϕ_0.

3.2.1 First-stage estimator

In the first stage, the researcher estimates the parameter $\phi_0 = (\beta_0^T, \lambda_0^T)^T$ under the parametric assumption $F \in \mathcal{F}_0$. Here ϕ_0 either corresponds to the true parameter if we are under $H_0 : F \in \mathcal{F}_0$ or a pseudo-true value if we are under the alternative $\overline{H}_a : F = F_a \notin \mathcal{F}_0$. We assume the following properties for the first stage estimator.

Assumption 2 *First-stage estimator*

- $\phi_0 \in$ *interior*(Φ) *with* Φ *a compact subset of* $\mathbb{R}^{dim(\phi)}$
- *Under* $H_0 : F \in \mathcal{F}_0$*, the first stage estimator* $\hat{\phi}$ *is* $\sqrt{N}-$ *consistent for the true parameter* ϕ_0*:*

$$\hat{\phi} \underset{N \to +\infty}{\overset{P}{\longrightarrow}} \phi_0 \qquad \textit{and} \qquad \sqrt{N}(\hat{\phi} - \phi_0) = O_p(1).$$

- *Under* $\overline{H}_a : F = F_a \notin \mathcal{F}_0$ *the first stage estimator* $\hat{\phi}$ *is consistent for a pseudo-true value* ϕ_0

$$\hat{\phi} \underset{N \to +\infty}{\overset{P}{\longrightarrow}} \phi_0 \qquad \textit{with} \qquad \phi_0 = \underset{\phi \in \Phi}{\operatorname{argmax}}\ \mathcal{Q}(\phi)$$

  with $\mathcal{Q}(\cdot)$ *a population objective function.*

Important examples of $\mathcal{Q}$ include the population GMM objective function or the population log-likelihood. The usual primitive conditions for $\sqrt{N}$-consistency under H_0 and consistency under $\overline{H}_a$ can be found in Newey and McFadden (1994). Under the alternative, the pseudo-true value holds different interpretations depending on the objective function $\mathcal{Q}$. In the context of a maximum likelihood estimator, White (1982) shows that the pseudo true value minimizes the Kullback-Leibler distance between the specified parametric family of likelihoods and the true likelihood. In the context of the GMM estimator, Hall and Inoue (2003) show that the pseudo-true value minimizes a weighted sum of population moments.[10]

3.2.2 A class of instruments robust to parameter uncertainty

To test $H_0 : F \in \mathcal{F}_0$, we want to evaluate whether the conditional moment restrictions in (2.1) evaluated at $\theta_0 = (\beta_0, F_0(\cdot|\lambda_0))$ hold. We employ the same strategy as in the simple hypothesis case and transform the p conditional moment restrictions into L unconditional moment restrictions using the family of instruments $(h_l)_{l=1,\ldots,L}$.[11] The main difference with the simple hypothesis case stems from the fact that to derive our test statistic, we cannot directly plug in the (pseudo)-true value $\phi_0 = (\beta_0^T, \lambda_0^T)^T$. Instead, we must rely on the first stage estimator $\hat{\phi} = (\hat{\beta}, \hat{\lambda})$, which is random because of sampling error. Thus, we must take into account the randomness generated by the estimation of the (pseudo)-true value to determine the correct asymptotic distribution of our test statistic under the null hypothesis. This problem is commonly referred to as parameter uncertainty (or a nuisance parameter problem).

To tackle this issue, we apply a strategy commonly referred to as orthogonalization, which involves transforming the unconditional moment conditions ex-ante to ensure that they become asymptotically insensitive to the first-stage estimation of θ_0. For example, Bontemps and Meddahi (2005) apply this approach in the context of specification tests, while Chernozhukov et al. (2018) use it for valid inference in two-step estimation procedures.[12] To illustrate the intuition behind orthogonalization, let us consider a basic situation where the researcher wants to test $H_0 : \mathbb{E}\left[g(W_i, \phi_0)\right] = 0$ but only observes a $\sqrt{N}$-consistent estimator of ϕ_0. Under classical regularity conditions,

[10] In particular, they show that in the over-identified case, the value of the pseudo-true value depends on the weights assigned to each moment condition.

[11] As previously, $h(Z_i) \in \mathcal{M}_{L,p}(\mathbb{R})$ represents the matrix that stacks horizontally the L vectors $h_l(Z_i)$.

[12] The classical approach to deal with parameter uncertainty consists of correcting the asymptotic variance-covariance matrix either analytically or using resampling methods such as Bootstrap. The analytical way of modifying the variance can be difficult to implement as it requires to estimate the covariance between the moment used for estimation and the one used for estimation.

which we will elucidate later in our context, for any function $g(W_i, \phi)$,

$$\frac{1}{\sqrt{N}}\sum_{i=1}^{N} g(W_i, \hat{\phi}) = \frac{1}{\sqrt{N}}\sum_{i=1}^{N} g(W_i, \phi_0) + \mathbb{E}\left[\frac{\partial g(W_i, \phi_0)}{\partial \phi^T}\right]\sqrt{N}(\hat{\phi} - \phi_0) + o_p(1). \quad (3.5)$$

Thus, if g is such that $\mathbb{E}\left[\frac{\partial g(W_i,\phi_0)}{\partial \phi^T}\right] = 0$ $(*)$, then the empirical moment has the same asymptotic distribution under H_0, regardless of whether we substitute ϕ_0 or $\hat{\phi}$. Several methods have been proposed in the literature to ensure the fulfillment of $(*)$. In this paper, we propose a novel method to orthogonalize the moments, which can be applied broadly and features the advantage of not requiring any prior knowledge of the statistical properties of the first-stage estimator, provided that it is $\sqrt{N}$-consistent. Our approach consists of orthogonalizing the instruments instead of the unconditional moments directly. Specifically, we begin by delineating a subset of $\mathcal{H}_0$, denoted as $\mathcal{H}_0^{\perp}$, in which the instruments satisfy a Neyman orthogonality property analog to $(*)$. In a second stage, we provide a systematic way of transforming the initial instruments to ensure that they satisfy this Neyman orthogonality property.

Definition 2 *Let $\mathcal{H}_0^{\perp} \subset \mathcal{H}_0$ the subset of instruments that satisfy the following restrictions:*

1. *$\exists C'$, $\forall l$, $\mathbb{E}[\|h_l(Z_i)\|_2^4] < C'$,*
2. *The moment $\mathbb{E}\left[h(Z_i)m(W_i, \phi_0)\right]$ is Neyman orthogonal with respect to $\phi = (\beta, \lambda)$:*

$$\mathbb{E}\left[h(Z_i)\frac{\partial m(W_i, \phi_0)}{\partial \phi^T}\right] = 0_{L\times dim(\phi)}.$$

Again, we index $\mathcal{H}_0^{\perp}$ by 0 to underscore that its composition depends on H_0 and the pseudo-true value ϕ_0. As emphasized in equation (3.5), the Neyman orthogonality condition enables us to eliminate the first-order term in the Taylor expansion of our test statistic around ϕ_0. Consequently, our moment becomes asymptotically insensitive to vanishing fluctuations of $\hat{\phi}$ around ϕ_0. The condition (1) in Definition 2 ensures, among other things, that the uniform law of large numbers (ULLN) can be applied to derive a consistent estimator of Ω_0 even when ϕ_0 is estimated in the first stage. This condition would also be required even in the absence of any orthogonalization.

Furthermore, we impose that m satisfies the following regularity conditions to derive the asymptotic properties of our test.

Assumption 3 (Regularity conditions on $m(W_i, \phi)$) *The following conditions hold with probability one.*

1. $\phi \mapsto m(W_i, \phi)$ *is continuously differentiable for* $\phi \in \Phi$.
2. $\exists M(\cdot)$ *and* $\overline{M} \in \mathbb{R}_+$ *such that* $j \in [|1, p|]$, $\left|[m(W_i, \phi)]_j\right| \leq M(W_i)$ *with* $\mathbb{E}[M(W_i)^4] < \overline{M}$ *for* $\phi \in \Phi$.
3. $\exists \tilde{M}(\cdot)$ *and* $\overline{\tilde{M}} \in \mathbb{R}_+$ *such that* $\forall (k, j) \in [|1, p|] \times [|1, dim(\phi)|]$, $\left|\left[\frac{\partial m(W_i, \phi)}{\partial \phi^T}\right]_{k,j}\right| \leq \tilde{M}(W_i)$ *with* $\mathbb{E}[\tilde{M}(W_i)^2] < \overline{\tilde{M}}$ *for* $\phi \in \Phi$.

Assumption 3.2 is aimed at ensuring that $m(W_i, \phi)$ satisfies a uniform law of large numbers. Assumptions 3.1 and 3.3 ensure that: $\frac{\partial m(W_i, \phi_0)}{\partial \phi^T}$ (i) is well-defined, and (ii) satisfies a uniform law of large numbers over Φ. These assumptions are standard regularity conditions commonly used in parametric models. For instance, if m were used to estimate ϕ_0, the consistency of the estimator $\hat{\phi}$ and of its variance would require the fulfillment of all the aforementioned assumptions. Moreover, in most models with RCs, these conditions are likely to hold, as the part of m that depends on ϕ is usually a smooth and bounded probability function.[13] The following proposition emphasizes that when $h \in \mathcal{H}_0^\perp$ and under the classical regularity conditions outlined in Assumptions 2 and 3, the asymptotic properties of the test statistic studied in this paper remain unchanged, regardless of whether we substitute ϕ_0 or $\hat{\phi}$.

Proposition 3.4 *Assume that* $h \in \mathcal{H}_0^\perp$, $\hat{\phi}$ *satisfies Assumption 2, and* m *satisfies Assumption 3, then the following properties hold:*

- *Under* $H_0 : F \in \mathcal{F}_0, \qquad S_N(h, \hat{\phi}) \xrightarrow[N \to +\infty]{d} \chi^2_L,$
- *Under* $\overline{H}'_a :\ \mathbb{E}\left[h(Z_i) m(W_i, \phi_0)\right] \neq 0, \quad \forall q > 0, \quad \mathbb{P}(S_N(h, \hat{\phi}) > q) \underset{N \to +\infty}{\longrightarrow} 1,$

where a consistent estimator of Ω_0 *is simply the empirical variance-covariance matrix obtained by replacing* ϕ_0 *with* $\hat{\phi}$.

[13] The smoothness of m usually depends on the distributional assumption on the additive error shock. For instance, if the shock is distributed as a Gumbel, then m is infinitely differentiable.

We prove this proposition in Appendix B.2.2. There are two remaining issues to address: (i) how to ensure that $h \in \mathcal{H}_0^{\perp}$, and (ii) how to select $h \in \mathcal{H}_0^{\perp}$ to maximize the power of our test. These questions are examined in the following two subsections.

3.2.3 A systematic method to orthogonalize the instruments

In this section, we provide a systematic way of orthogonalizing an instrument. Let $h \in \mathcal{H}_0$, we define $h^{\perp}(Z_i)$ the orthogonalized counterpart of $h(Z_i)$ as follows:

$$h_{\Lambda}^{\perp}(Z_i) \equiv \left[h(Z_i)\Lambda(Z_i) - \mathbb{E}\left[BLP\left(h(Z_i)\Lambda(Z_i)\middle|\Lambda(Z_i)\frac{\partial m(W_i,\phi_0)}{\partial \phi^T}\right)\middle|Z_i\right]\right]\Lambda(Z_i). \quad (3.6)$$

where $\Lambda(Z_i)$ is a symmetric weighting matrix set by the researcher and $BLP(\cdot|\cdot)$ denotes the best linear predictor, which we define as follows.[14] For any random matrices $(A_i, B_i) \in \mathcal{M}_{k,p}(\mathbb{R}) \times \mathcal{M}_{p,m}(\mathbb{R})$ with $(k,m) \in \mathbb{N}^{*2}$, $BLP(A_i|B_i) = \mathbb{E}(A_iB_i)\mathbb{E}(B_i^TB_i)^{-1}B_i^T$. The expression of the orthogonalized instrument can be simplified if we set $\Lambda(Z_i) = I_p$. The general form exhibited above is however useful in deriving the MPI in the class of orthogonalized instruments. Next, we outline some standard regularity conditions that guarantee the existence of $h_{\Lambda}^{\perp}$.

Assumption 4 (Regularity conditions for the existence of $h_{\Lambda}^{\perp}$)

- $\exists C''$ *such that* $\mathbb{E}\left[\|h(Z_i)\Lambda(Z_i)\|_F^2\right] + \mathbb{E}\left[\left\|\Lambda(Z_i)\frac{\partial m(W_i,\phi_0)}{\partial \phi^T}^2\right\|_F^2\right] < C''$
- $\mathbb{E}\left[\frac{\partial m(W_i,\phi_0)}{\partial \phi}\Lambda(Z_i)^2\frac{\partial m(W_i,\phi_0)}{\partial \phi^T}\right]$ *is full rank.*

These conditions ensure that the best linear predictor of $h(Z_i)\Lambda(Z_i)$ on $\Lambda(Z_i)\frac{\partial m(W_i,\phi_0)}{\partial \phi^T}$ exists. The next proposition shows that under Assumptions 3 and 4, the orthogonalized instrument $h_{\Lambda}^{\perp}(Z_i)$ as defined above exists and satisfies the Neyman orthogonality property.

Proposition 3.5 *(Neyman orthogonality)*

Let $h \in \mathcal{H}_0$. Under Assumption 3 and 4, $h_{\Lambda}^{\perp}$ exists and the moment $\mathbb{E}\left[h_{\Lambda}^{\perp}(Z_i)m(W_i,\phi_0)\right]$ is Neyman orthogonal with respect to $\phi = (\beta, \lambda)$

[14] To avoid confusion with the BLP demand model, we use italic letters for the best linear predictor and consistently refer to the model as the BLP demand model.

Intuitively, $h_\Lambda^\perp(Z_i)$ can be interpreted as a linear projection of the initial instruments on the orthogonal complement to the linear subspace generated by $\frac{\partial m(W_i,\phi_0)}{\partial \phi^T}$. The proof of this Proposition is in Appendix B.2.2 and leverages the properties of the best linear predictor. In practice, the best linear predictor is not known but estimated and its derivation involves a first-stage estimator of ϕ_0. A potential concern is that the estimation of the best linear predictor could impact the asymptotic distribution of our test statistic. However, we show in Proposition 1.2 (Appendix A.3) that, under mild regularity conditions on m and Λ, the two test statistics have identical asymptotic distributions. The conditions we require on m are slightly more restrictive than in Assumption 3, in particular, we require m to be twice-differentiable in ϕ.

3.2.4 The Most Powerful instrument in $\mathcal{H}_0^\perp$

We now turn to the derivation of the MPI in the space of instruments $\mathcal{H}_0^\perp$ that are orthogonal to $\frac{\partial m(W_i,\phi_0)}{\partial \phi^T}$. The next proposition provides an expression for the MPI that maximizes the asymptotic slope of our test within the class of orthogonal instruments $\mathcal{H}_0^\perp$.

Proposition 3.6 (Most powerful instrument in $\mathcal{H}_0^\perp$)
Under a fixed alternative $\overline{H}_a : (\beta, F) = (\beta_a, F_a)$ where $F_a \notin \mathcal{F}_0$, we have the following:

$$h_a^\star(Z_i) \equiv \left(\mathbb{E}[\Delta_{0,a}^m(W_i)|Z_i]^T\right)_{\Sigma^{-1/2}}^\perp \in \underset{h\in\mathcal{H}_0^\perp}{\operatorname{argmax}} \;\; c(h,\theta_0)$$

with $\Sigma(Z_i) = \mathbb{E}\left[m(W_i,\phi_0)m(W_i,\phi_0)^T|Z_i\right]$ provided that:

1. *$\exists C'$ such that $\mathbb{E}[\|h_a^\star(Z_i)\|_F^4] < C'$*
2. *$\Omega_0 \equiv \mathbb{E}[h_a^\star(Z_i)m(W_i,\theta_0)m(W_i,\theta_0)^T h_a^\star(Z_i)^T]$ is full rank.*

The proposition above indicates that under the regularity conditions (1) and (2), the MPI in $\mathcal{H}_0^\perp$ corresponds to the orthogonalized counterpart of $\mathbb{E}[\Delta_{0,a}^m(W_i)|Z_i]$ with the usual inverse of the conditional variance term used as the weighting matrix. Consequently, the MPI within $\mathcal{H}_0^\perp$ can be interpreted as the orthogonalized version of the MPI within $\mathcal{H}_0$.

3.2.5 An alternative approach for the orthogonalization of the instruments

Our approach to orthogonalizing the instruments relies on computing $\mathbb{E}\left[\frac{\partial m(W_i,\phi_0)}{\partial \phi^T}\middle| Z_i\right]$. Depending on the example, this step may present challenges. In Examples 1 and 2, this computation is straightforward since $\frac{\partial m(W_i,\phi_0)}{\partial \phi^T}$ is a function of Z_i. However, this simplification does not apply in Example 3. To circumvent the problem of estimating $\mathbb{E}\left[\frac{\partial m(W_i,\phi_0)}{\partial \phi^T}\middle| Z_i\right]$, we present an alternative instrument orthogonalization method using a double linear projection, which can be implemented when $L > \dim(\phi_0)$.

3.2.6 Strengths and limitations of the orthogonalization strategy

We now briefly examine the merits and drawbacks of our orthogonalization strategy. On the positive side, our method for orthogonalization is straightforward to implement and operates independently of any knowledge of the first-stage estimator beyond its $\sqrt{N}$-consistency. Furthermore, as emphasized in Chernozhukov et al. (2018), it has the potential to improve the finite-sample properties of our test. However, the downside of focusing on $\mathcal{H}_0^{\perp}$ rather than $\mathcal{H}_0$, is the loss of the discriminative power included in the linear subspace generated by $\frac{\partial m(W_i,\phi_0)}{\partial \phi^T}$.[15] By contrast, the classical approach of correcting for the variance retains this identifying power but may inflate the variance of the test statistic, thereby reducing power in finite samples. Overall, determining the superior approach in terms of power remains ambiguous, whereas the orthogonalization strategy presents a clear advantage in terms of ease of implementation. Our simulation experiments suggest that the orthogonalization strategy outperforms the classical approach in terms of power while maintaining good size control.

3.3 Connection with the optimal instruments

In this section, we assume that $F \in \mathcal{F}_0$ and we aim to estimate the true parameter $\phi_0 = (\beta_0^T, \lambda_0^T)^T$, identified by the non-linear conditional moment restriction in (2.1), using GMM. The optimal instruments from Chamberlain (1987) enhance efficiency of the GMM estimator by minimizing its asymptotic variance-covariance. We have the

[15] In particular, if $\Delta_{0,a}^m$ lies within this subspace, distinguishing between H_0 and $\overline{H}_a$ becomes infeasible.

following expression for the optimal instruments:[16]

$$h_E^*(Z_i) = \left(\mathbb{E}\left[m(W_i,\phi_0)m(W_i,\phi_0)^T | Z_i \right]^{-1} \mathbb{E}\left[\frac{\partial m(W_i,\phi_0)}{\partial \phi^T} \middle| Z_i \right] \right)^T .$$

We now show that the MPI devoted to testing the specification of the model at the true parameter against any fixed local alternative can be expressed as a linear combination of the optimal instruments. By definition, we can write the MPI to test $\overline{H}_0 : \phi = \phi_0$ against a fixed alternative $\overline{H}_a : \phi = \phi_a$ as follows:

$$h_a^*(Z_i) = \left(\mathbb{E}\left[m(W_i,\phi_0)m(W_i,\phi_0) | Z_i \right]^{-1} \mathbb{E}[\Delta_{0,a}^m(W_i)|Z_i] \right)^T$$

Under the parametric assumption $F \in \mathcal{F}_0$, one can write a Taylor expansion of $m(W_i,\cdot)$ around ϕ_0 as follows

$$\Delta_{0,a}^m(W_i) = \frac{\partial m(W_i,\phi_0)}{\partial \phi^T}(\phi_a - \phi_0) + o(||\phi_a - \phi_0||_2)\,.$$

Thus, we can see that when ϕ_a is in a neighborhood of ϕ_0, the MPI, $h_a^*(Z_i)$, against this fixed alternative is a linear combination of the optimal instruments $h_E^*(Z_i)$:

$$h_a^*(Z_i)^T \approx \underbrace{\mathbb{E}\left[m(W_i,\phi_0)m(W_i,\phi_0)^T | Z_i \right]^{-1} \mathbb{E}\left[\frac{\partial m(W_i,\phi_0)}{\partial \phi^T} \middle| Z_i \right]}_{h_E^*(Z_i)^T} (\phi_a - \phi_0).$$

It follows that the classical optimal instruments can be interpreted as an approximation of the MPI devoted to testing $H_0 : \phi = \phi_0$ against any fixed local alternative.[17] We leave the implications of this relationship and the potential role of the MPI for estimation for future research.

4 A feasible approximation of the MPI

The MPI is the first-best instrument, in the context of a moment-based test, to reject $\overline{H}_0 : (\beta, F) = (\beta_0, F_0)$ or $H_0 : F \in \mathcal{F}_0$ under a fixed alternative $\overline{H}_a : (\beta, F) = (\beta_a, F_a)$.

[16]For the sake of completeness, we show this result in Appendix B.2.3.

[17]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.

Yet, its derivation requires the knowledge of $\overline{H}_a$, which is unknown in practice. In this section we provide an approximation of the most powerful instrument when the econometrician wants to remain agnostic about the fixed alternative, which is the most common scenario in empirical work.

Simplifying assumptions. For exposition, we omit the conditional variance term as it does not significantly modify the subsequent analysis. We adopt the homoscedastic MPI (i.e., $h_a^*(Z_i) = \mathbb{E}[\Delta_{0,a}^m(W_i)|Z_i]^T$) as our baseline MPI.[18] The literature offers various non-parametric methods to estimate the conditional variance term (eg: Ai and Chen (2003)), which could be applied in our setup. However, it is widely acknowledged that estimating the conditional variance, a component also involved in the derivation of the optimal instrument, poses practical challenges. Hence, researchers typically ignore this term or impose an ad-hoc structure on the form that it can take.[19]

4.1 The interval instruments

For ease of implementation, we focus on an approximation of the MPI using a finite set of instruments. We will show that this finite-dimensional set of instruments is straightforward to compute and generates an asymptotic slope that approaches the one implied by the MPI, even if it does not generally reach it. In Appendix A.5.3, we provide a preliminary analysis of the case where the number of interval instruments, denoted by $L \equiv L_N$, increases with the sample size.

In the majority of models featuring random coefficients, the correction term resulting from misspecification takes the form of a known function integrated over the unknown distribution of random coefficients (i.e., the fixed alternative). In this case, we say that the correction term is discretizable. Consequently, we can exploit this specific form to approximate this integral by a sum where solely the weights vary with the fixed alternative. Our strategy then consists of using the terms in the sum as our set of instruments.

[18] This last expression corresponds to the exact formulation of the MPI under homoscedasticity of the error term.

[19] For instance, most approximations of the optimal instruments in the BLP demand model simply ignore the conditional variance term (eg: Reynaert and Verboven (2014)).

Definition 3 *We say that $\Delta^m_{0,a}$ is discretizable if there exists an explicit function ψ_0 such that:*

$$\Delta^m_{0,a}(W_i) = \int_{\mathcal{V}} \psi_0(W_i, v, \beta_a) dF_a(v). \tag{4.7}$$

For instance, in Examples 1 and 2 the correction term is discretizable.

Example 1 continued. In the mixed linear model, we can write the correction term as follows:

$$\Delta^{m_1}_{0,a}(W_i) = \int_{\mathcal{V}} \underbrace{\int G_\varepsilon(b - X'_{1i}\beta_0 - X'_{2i}u)dF_0(u) - G_\varepsilon(b - X'_{1i}\beta_a - X'_{2i}v)}_{\psi_0(W_i,v,\beta_a)} dF_a(v).$$

Example 2 continued. In the mixed logit model, the correction term is written as:

$$\Delta^{m_2}_{0,a}(W_i) = \int_{\mathcal{V}} \underbrace{\rho(X_i;\theta_0) - \frac{\exp\{X_1\beta_a + X_2 v\}}{1 + \sum_{k=1}^J \exp\{X'_{1j}\beta_a + X'_{2j}v\}}}_{\psi_0(W_i,v,\beta_a)} dF_a(v).$$

In most models with RCs, it can be shown that β_a is identified independently of F_a and can be consistently estimated without any assumptions about F_a. For instance, in Example 1, β_a can be estimated using a straightforward linear regression. In Example 2, Fox et al. (2012) provides a constructive proof for the identification of β_a. Therefore, in what follows, we treat β_a as known (while in practice, we simply replace β_a by its empirical counterpart). Additionally, it is important to note that the choice of β_a does not impact the validity of the instruments, yet it does influence the power they generate. When $\Delta^m_{0,a}$ is discretizable, we can use a Riemann sum approximation of the integral to approximate the MPI as follows:

$$\begin{aligned} h^*_a(Z_i) = \mathbb{E}[\Delta^m_{0,a}(W_i)|Z_i] &= \mathbb{E}\left[\int_{\mathcal{V}} \psi_0(W_i, v, \beta_a)dF_a(v) \middle| Z_i\right] \approx \sum_{l=1}^L \omega_l(F_a)\mathbb{E}\left[\psi_0(W_i, v_l, \beta_a)|Z_i\right] \\ &\equiv \sum_{l=1}^L \omega_l(F_a)\pi_l(Z_i) \end{aligned}$$

with $\{v_l\}_{l=1,\ldots,L}$ L points chosen in the domain of definition of F_a, and $\omega_l(F_a)$ the corresponding weights.[20] The main benefit from this decomposition is that only the

[20] In the usual Riemann sum, each weight $\omega_l(F_a)$ corresponds to the density evaluated at point v_l

weights depend on the alternative.

The interval instruments. From what precedes, the MPI can be approximated as follows: $h_a^*(Z_i) \approx \sum_{l=1}^{L} \omega_l(F_a)\pi_l(Z_i)$. As we do not know the weights $\omega_l(F_a)$, we propose to take the functions $\{\pi_1(Z_i), ..., \pi_L(Z_i)\}$ as our testing instruments and we follow the procedures outlined in Section 3.1 for the simple hypothesis $\overline{H}_0 : (F, \beta) = (F_0, \beta_0)$ or in Section 3.2 for the composite hypothesis $H_0 : F \in \mathcal{F}_0$. For exposition purposes, we omit the dependence of these instruments in θ_0 and β_a. In the remainder of this paper, we use the terms interval instruments and feasible MPI interchangeably to refer to $\Pi(Z_i) \equiv (\pi_1(Z_i), ..., \pi_L(Z_i))^T$.

To derive the interval instruments, the researcher must select a collection of points within the domain of definition of F_a. It is worth noting that even if a point lies outside the support of F_a, the instrument remains valid. This approach is reminiscent of the way Fox et al. (2011) constructs its flexible estimator of the distribution of RCs, which involves discretizing the unknown distribution over a grid of fixed types and estimating the associated weights. While determining the optimal set of points falls beyond the scope of this paper, we discuss the effects of the location and the number of points on the properties of our test in Appendix A.5.1. Furthermore, it is straightforward to conduct robustness checks by testing different sets of points. Finally, in Appendix A.5.4, we present an alternative method to construct the interval instruments in settings where the correction term is not discretizable—for example, in the BLP demand model (Example 3).

4.2 A theoretical basis for the interval instruments

The objective is now to explain why the interval instruments work well for testing the distribution of random coefficients. Intuitively, if $m(W_i, \theta_0)$ is strongly correlated with a weighted sum of terms, it should also be strongly correlated with at least one of the terms in the sum. Beyond this simple intuition, we now introduce an alternative interpretation for the slope of the test, which in turn allows us to better grasp the strength of the interval instruments to test the distribution of RCs.

Up to this point, we have used the asymptotic slope of the test as a power criterion.

(i.e. $f_a(v_l)$) times the width of the interval around v_l.

The next proposition shows that, under homoscedasticity, the slope of the test is also a measure of the predictive power of the instruments on the residuals.

Proposition 4.1 (Alternative intepretation of the asymptotic slope)

Let $h \in \mathcal{H}_0$, and assume homoscedasticity (i.e., under $\overline{H}_0$, we assume $\mathbb{E}[m(W_i, \theta_0)m(W_i, \theta_0)^T \mid Z_i] = \sigma^2 I_p$). Then:

$$c(h, \theta_0) = \sigma^{-2}\ \mathbb{E}\big[||BLP(m(W_i, \theta_0)^T | h(Z_i)^T)||_2^2\big].$$

The preceding proposition shows that the slope associated with a given instrument $h(Z_i)$ is equal to the expected squared norm of the linear projection of the "residuals" $m(W_i, \theta_0)$ on the space spanned by the set of instruments $h(Z_i)^T$. The proof of this result is in Appendix B.3. In Appendix A.5.2, we demonstrate that a similar interpretation of the slope holds even without assuming homoscedasticity. This result also sheds a new light on the expression we obtained for the MPI. Under homoscedasticity, the MPI associated with the alternative $\overline{H}_a$ satisfies:

$$\mathbb{E}[\Delta_{0,a}^m(W_i) \mid Z_i] = \mathbb{E}[m(W_i, \theta_0) \mid Z_i] \quad \text{a.s. under } \overline{H}_a.$$

By definition, $\mathbb{E}[m(W_i, \theta_0) \mid Z_i]$ minimizes the mean squared prediction error between $m(W_i, \theta_0)$ and all square-integrable and measurable functions of Z_i. It is then straighforward to show that under $\overline{H}_a$, $h_a^*(Z_i)$ achieves the maximum predictive power among the functions of Z_i. Specifically, we have:

$$\forall h \in \mathcal{H}_0, \quad \mathbb{E}\left[\|m(W_i, \theta_0) - BLP(m(W_i, \theta_0)^T | h(Z_i)^T)\|_2^2\right] \geq \mathbb{E}\left[\|m(W_i, \theta_0) - \mathbb{E}[m(W_i, \theta_0) | Z_i]\|_2^2\right].$$

Under $\overline{H}_a$, the previous inequality implies:

$$\forall h \in \mathcal{H}_0, \quad \sigma^2 c(h, \theta_0) = \mathbb{E}\left[\|BLP(m(W_i, \theta_0)^T | h(Z_i)^T)\|_2^2\right] \leq \mathbb{E}\left[\|\mathbb{E}[m(W_i, \theta_0) | Z_i]\|_2^2\right] = \sigma^2 c(h_a^*, \theta_0).$$

With this new interpretation of the slope, it becomes quite clear why the interval instruments defined earlier can serve as a good approximation of the MPI. The intuition is given in the following lines:

$$\begin{aligned}
E[\Delta^m_{0,a}(W_i) \mid Z_i] \approx \Pi(Z_i)^T \omega(F_a) &\implies \text{under } \overline{H}_a,\ \mathbb{E}[m(W_i,\theta_0)|Z_i] \approx BLP(m(W_i,\theta_0)^T|\Pi(Z_i)^T)^T \\
&\implies \text{under } \overline{H}_a,\ \mathbb{E}\left[\|\mathbb{E}[m(W_i,\theta_0)|Z_i]\|_2^2\right] \approx \mathbb{E}\left[\|BLP(m(W_i,\theta_0)^T|\Pi(Z_i)^T)\|_2^2\right] \\
&\implies \text{under } \overline{H}_a,\ c(h_a^*,\theta_0) \approx c(\Pi,\theta_0).
\end{aligned}$$

In words, the Riemann sum approximation of the integral allows us to linearize the MPI. As a result, under $\overline{H}_a$, the predictive power of the MPI associated with $\overline{H}_a$ and that of $\Pi(Z_i)$ on the residuals are close to each other. The greater the accuracy of the Riemann approximation, the closer the slopes generated by the interval instruments and the MPI become. A more accurate approximation can generally be achieved via a finer discretization and thus an increase in the number of interval instruments. However, as we discuss in Appendix A.5.1, increasing the number of instruments can lead to finite-sample issues such as large size distortions due to the poor estimation of Ω_0 or an inflation of the critical value due to an increase in the degrees of freedom of the asymptotic chi-square distribution.

5 Model selection

Up to this point, our emphasis has been on constructing a powerful test designed to validate or reject a given specification $\mathcal{F}_0$ for the distribution of RCs. In this section, we shift our focus to model selection and show how the researcher can use the tools developed previously to discriminate between different specifications for the distribution of RCs, even in the case where all the specifications are rejected.

We follow the approach proposed in Vuong (1989) and extended in Rivers and Vuong (2002), which tests for differences in a user-specified lack-of-fit criterion between two potentially non-nested models. A key advantage of this method is that allows both competing models to be misspecified. It has been widely used in empirical work, including model selection in vertical relationships between manufacturers and retailers (Bonnet and Dubois (2010), Duarte et al. (2024)) and in models of firms' price-setting behaviors (Backus et al. (2021)).[21]

As a lack-of-fit criterion, we propose to use the slope of the test obtained with

[21]In Appendix A.6, we discuss alternative approaches to model selection and explain why we favor the method proposed in Rivers and Vuong (2002).

the interval instruments defined in Section 4. We motive this choice shortly after. The testing procedure is as follows. Suppose we want to discriminate between two specifications of the RCs $\mathcal{F}_1$ and $\mathcal{F}_2$.[22] We define for $k = 1, 2$, the asymptotic slope $C_k \equiv c(h_k, \phi_k)$ obtained with interval instruments h_k and at ϕ_k the (pseudo)-true value under specification $\mathcal{F}_k$.[23] The null hypothesis in a Vuong-type test asserts that both models fit the data equally well according to the chosen criterion, while each alternative implies that one model outperforms the other.

$$H_0^{RV} : C_1 = C_2, \qquad H_1^{RV} : C_1 < C_2 \qquad \text{and} \qquad H_2^{RV} : C_2 < C_1.$$

The Rivers and Vuong test statistic is expressed as follows:

$$T_N^{RV} = \frac{\sqrt{N}(\hat{C}_1 - \hat{C}_2)}{\hat{\sigma}_{RV}}$$

with $\hat{C}_k$ are empirical counterparts of the slope and $\hat{\sigma}_{RV}$ is the asymptotic variance of $\sqrt{N}(\hat{C}_1 - \hat{C}_2)$ under H_0^{RV}. Namely, $\sigma_{RV}^2 = R\Omega R^T$ with:

$$R = \begin{pmatrix} \mathbb{E}[h_1(Z_i)m(W_i, \phi_1)]^T \Omega_{1,1}^{-1} & -\mathbb{E}[h_2(Z_i)m(W_i, \phi_2)]^T \Omega_{2,2}^{-1} \end{pmatrix} \quad \textit{and} \quad \Omega = \begin{pmatrix} \Omega_{1,1} & \Omega_{1,2} \\ \Omega_{2,1} & \Omega_{2,2} \end{pmatrix},$$

with $\Omega_{k,j} = \mathbb{E}[h_k(Z_i)m(W_i, \phi_k)m(W_i, \phi_j)^T h_j(Z_i)^T]$ and ϕ_k the (pseudo)-true value under specification $\mathcal{F}_k$ (as defined in Assumption 2).

The decision rule for a level-α test is as follows: reject H_0^{RV} if $|T_N^{RV}|$ exceeds the $1 - \frac{\alpha}{2}$ quantile of the standard normal distribution $\mathcal{N}(0, 1)$, in favor of the appropriate alternative depending on the sign of T_N^{RV}. The next proposition describes the asymptotic behavior of the test statistic.

Proposition 5.1 *Assume that for $k \in \{1, 2\}$, $h_k \in \mathcal{H}_k^{\perp}$, $\hat{\phi}_k$ satisfies Assumption 2, m satisfies Assumption 3, and $\sigma_{RV} > 0$, then the following properties hold:*

- *Under $H_0^{RV} : C_1 = C_2$* $\qquad T_N^{RV} \xrightarrow[N \to +\infty]{d} \mathcal{N}(0, 1)$,

[22]For simplicity, we assume the competing models differ solely in the way the distribution of random coefficients is specified.

[23]It's worth noting that under misspecification, the pseudo-true value depends on the criterion function used to estimate the model in the first stage.

- *Under* $H_1^{RV} : C_1 < C_2 \qquad T_N^{RV} \xrightarrow[N\to+\infty]{d} -\infty$,
- *Under* $H_2^{RV} : C_1 > C_2 \qquad T_N^{RV} \xrightarrow[N\to+\infty]{d} +\infty$,

where a consistent estimator of σ_0^{RV} *is simply the empirical counterpart obtained by replacing* ϕ_1 *and* ϕ_2 *with* $\hat{\phi}_1$ *and* $\hat{\phi}_2$.

The proof of this proposition is in Appendix B.4. Several remarks are in order. First, under the regularity assumptions considered in this paper and the non-degeneracy of σ_{RV} (i.e. $\sigma_{RV} > 0$), which is also stipulated in Rivers and Vuong (2002), the asymptotic behavior of T_N^{RV} is identitical to the asymptotic behavior of the general test statistic in Rivers and Vuong (2002). In particular, our test for $H_0^{RV} : C_1 = C_2$ is asymptotically valid. Second, by taking $h_k \in \mathcal{H}_k^{\perp}$, we make the test statistic asymptotically insensitive to the first-stage estimators $\hat{\phi}_1$ and $\hat{\phi}_2$, which significantly simplifies the derivation of the asymptotic variance in comparison with the classical procedure in Rivers and Vuong (2002).

We now briefly discuss the assumption of the non-degeneracy of σ_{RV}. This assumption rules out two problematic scenarios. The first scenario occurs when all the tested moments are satisfied or almost satisfied as emphasized in Hall and Pelletier (2011) and Duarte et al. (2024) (as then, $R \approx 0$). This situation can happen if the tested moments are uninformative about the distribution of RCs. Assuming that F is non-parametrically identified, we can discard this first possibility, as the instruments are chosen to make the tested moments highly informative about the distribution of RCs. The second scenario occurs when the tested moments are identical for both specifications (as then, $\Omega_{j,k} = \tilde{\Omega}$ for $(j,k) \in \{1,2\}^2$ and $R = (\breve{R}, -\breve{R})$). This happens when the distributions are identitical, which may occur if the competing specifications are nested or overlap. This issue is also present in the original test in Vuong (1989). To address this issue, we propose pre-testing whether the distributions obtained from the different specifications are distinguishable. In Appendix A.7, we provide a sequential procedure in the case of nested specifications. If $\sigma_{RV} > 0$ is not satisfied, severe inferential problems may occur, including large size distortions. Alternative methods have been proposed in the literature to address the potential degeneracy of σ_{RV}, which could also be applied in our setting.[24]

[24] Among others, Shi (2015) proposes a corrected test statistic with a simulated critical to control

We conclude this section by discussing the choice of the asymptotic slope associated with the interval instruments as the lack-of-fit criterion. As emphasized in Section 4.2, under homoscedasticity, the slope associated with the interval instruments approximates that obtained using the MPI, which equals $\mathbb{E}[\|\mathbb{E}[m(W_i, \phi_k) \mid Z_i]\|_2^2]$. This quantity provides an objective, interpretable, and invariant measure of the extent to which the conditional moment restriction $\mathbb{E}[m(W_i, \phi_k) \mid Z_i] = 0$ is violated. This criterion is less sensitive to the researcher's choices. The derived slope depends only on the selection of points within the support of F_a, which can be identical across specifications, and indirectly on the model estimation choices in the first stage (affecting the pseudo-true value). In contrast, alternative criteria, such as the GMM objective function, depend directly on the moments used for model estimation and the weighting matrix, which can vary across specifications.

6 Monte Carlo experiments

To assess the finite-sample performance of our specification test and model selection procedure, we conduct a series of simulation experiments. For brevity, we report only a selected subset of results pertaining to the specification test in the mixed logit model (Example 2). In Appendix C.1, we present additional simulation results in the mixed logit case, which corroborate the findings presented below. In Appendix C.2, we present simulation results for the specification test applied to the BLP demand model. These results indicate that our test exhibits good size control and substantially outperforms the traditional alternatives in terms of power. In Appendix C.3, we present simulation evidence on the performance of our model selection procedure in the context of the mixed logit model (Example 2). The findings indicate that our procedure reliably selects the correct specification for the RC, favoring either the true specification or the null hypothesis.

for the size when σ_{RV} is "close" to zero for the log-likelihood version of the test developed in Vuong (1989); Duarte et al. (2024) provides an approach to diagnostic the problem of weak instruments in this setting.

6.1 Finite sample performance of our test in the mixed logit model

We now present a sample of the simulation results for the canonical mixed logit model. To evaluate the empirical size and power of our test, we consider several testing instruments: the infeasible most powerful instrument (MPI), used as a benchmark; a feasible approximation of the MPI, constructed following the procedure described in Section 4.1; and a polynomial transformation of the exogenous variables based on Hermite polynomials with decaying tails.[25] For the construction of the infeasible MPI and the feasible approximations of the MPI, we ignore the conditional variance term. We provide more details on the construction of each set of instruments in Appendix C.6. Moreover, we consider two procedures to adjust the test for parameter uncertainty: the orthogonalization procedure we developed in Section 3.2.2 and a classical variance correction approach that we explain in Appendix C.5. Estimation is done via simulated maximum likelihood. More detail on the maximization procedure and the asymptotic properties of the estimator is provided in the Appendix C.4.1.

The simulation design is standard. The market includes $J = 12$ products. The utility associated with each product j is given by:

$$u_{ij} = -0.5 + x_{aij} + 2x_{bij} - x_{cij} + x_{dij}v_i + \varepsilon_{ijt} \qquad \varepsilon_{ij} \sim EV1.$$

The exogenous products attributes x_a, x_b, x_c and x_d follow a joint normal distribution, which is given in Appendix C.1. Consumers differ in their preferences for x_d, but not for x_a, x_b, x_c. The RC v_i follows various distributions depending on the simulation exercise. We observe the choices made by N agents in this market. The sample size N varies across 500, 1000 and 5000 agents. The hypothesis tested throughout the simulations is $H_0 : F \in \mathcal{F}_0$, where $\mathcal{F}_0$ is the family of normal distributions. We set the nominal size to 5%.

Empirical size. The random coefficient v_i is normally distributed and we derive the empirical rejection rates. Below in Table 1, we report the empirical sizes of the test with

[25] The exponential decay is incorporated to enhance the finite sample size control of the test by reducing the influence of moments involving high-order polynomial, whose behavior is often poorly approximated by the Central Limit Theorem in finite samples.

the different sets of instruments described previously and for different distributions of the RC such that $v_i \sim F \in \mathcal{F}_0$. We do not report the MPI as under H_0, all the valid instruments have a slope that is equal to 0 and can be considered as MPIs.

The results in Table 1 indicate that our test provides a good size control across the sets of instruments and for both of the procedures used to control for parameter uncertainty. Moreover, we observe that when we increase the sample size, the empirical size for all the configurations converges towards the nominal size, which is a good indication of the asymptotic validity of our test. Furthermore, we observe that for most configurations, the orthogonalization procedure yields an empirical size that closely matches the nominal size, even more so than the classical correction, where the empirical rejection rate is more conservative. These results corroborate that the orthogonalization procedure makes the test statistic robust to parameter uncertainty.

Power against Gaussian mixture alternatives. The random coefficients is now distributed according to the Gaussian mixtures described below.[26] The rest of the simulation setup is unchanged. The empirical power corresponds to the empirical rejection rate of H_0.

$$
\begin{aligned}
&v = Dv_1 + (1-D)v_2, \quad \mathbb{P}(D=1) = p, \quad p \in \{0.1; 0.2; 0.3; 0.4; 0.5\} \\
&v_1 \sim \mathcal{N}\left(-\sqrt{\frac{3p}{1-p}} + 2, 1\right) \quad v_2 \sim \mathcal{N}\left(\sqrt{\frac{3(1-p)}{p}} + 2, 1\right).
\end{aligned}
$$

We report the results in Table 2. We observe that, under the orthogonalization procedure, as predicted by our analysis, the MPI generates the most powerful test, followed by the sets FMPI (6) and FMPI (10). We notice that FMPI (6) performs slightly better than FMPI (10), suggesting that a larger number of interval instruments does not necessarily result in a more powerful test. The instruments that perform the worst are, by far, the polynomial instruments. When we examine the results with the standard variance correction, we find that, for a same set of instruments, the test is slightly less powerful compared to the test with the orthogonalization procedure. However, the ranking of the instrument sets remains unchanged. It is also noTable that with a small sample size ($N = 500$), the infeasible MPI exhibits limited power under the standard variance correction procedure, although this trend reverses as N

[26] We plot the alternative distributions in Figure **??** in Appendix C.2

Table 1: Empirical size ($\alpha = 5\%$, 1000 replications)

Number of individuals	N=500					
2-step adjustment	orthogonalization			std. variance correction		
Instruments	FMPI (6)	FMPI (10)	Pol	FMPI (6)	FMPI (10)	Pol
$v_i \sim \mathcal{N}(-1, 0.5^2)$	0.048	0.037	0.105	0.005	0.017	0.068
$v_i \sim \mathcal{N}(0, 0.75^2)$	0.038	0.039	0.08	0.008	0.014	0.054
$v_i \sim \mathcal{N}(1, 1^2)$	0.05	0.051	0.08	0.005	0.019	0.042
$v_i \sim \mathcal{N}(2, 2^2)$	0.035	0.038	0.078	0.006	0.015	0.053
$v_i \sim \mathcal{N}(3, 3^2)$	0.053	0.046	0.043	0.016	0.038	0.022
Number of individuals	N=1000					
2-step adjustment	orthogonalization			std. variance correction		
Instruments	FMPI (6)	FMPI (10)	Pol	FMPI (6)	FMPI (10)	Pol
$v_i \sim \mathcal{N}(-1, 0.5^2)$	0.041	0.046	0.084	0.009	0.013	0.056
$v_i \sim \mathcal{N}(0, 0.75^2)$	0.035	0.042	0.086	0.005	0.008	0.054
$v_i \sim \mathcal{N}(1, 1^2)$	0.061	0.06	0.075	0.006	0.021	0.045
$v_i \sim \mathcal{N}(2, 2^2)$	0.056	0.055	0.07	0.015	0.018	0.045
$v_i \sim \mathcal{N}(3, 3^2)$	0.041	0.058	0.05	0.009	0.039	0.031
Number of individuals	N=5000					
2-step adjustment	orthogonalization			std. variance correction		
Instruments	FMPI (6)	FMPI (10)	Pol	FMPI (6)	FMPI (10)	Pol
$v_i \sim \mathcal{N}(-1, 0.5^2)$	0.051	0.051	0.085	0.01	0.019	0.059
$v_i \sim \mathcal{N}(0, 0.75^2)$	0.052	0.043	0.075	0.01	0.011	0.045
$v_i \sim \mathcal{N}(1, 1^2)$	0.052	0.052	0.059	0.011	0.021	0.037
$v_i \sim \mathcal{N}(2, 2^2)$	0.048	0.035	0.058	0.012	0.01	0.035
$v_i \sim \mathcal{N}(3, 3^2)$	0.055	0.055	0.046	0.014	0.018	0.029

Note: $FMPI(6)$ corresponds to the feasible MPI with 6 points chosen in the support of F, same for $FMPI(10)$ with 10 instruments and Pol corresponds to the polynomial transformation of the exogenous variables.

increases. One possible explanation is that the MPI does not maximize the slope of the test when using the standard classical variance correction. In Appendix C.1, we also consider the case where the true distribution of v_i is a Gamma distribution and we obtain similar findings.

Table 2: Empirical power against Gaussian mixture ($\alpha = 5\%$, 1000 replications)

Number of individuals	N=500							
2-step adjustment	orthogonalization				std. variance correction			
Instruments	MPI	FMPI (6)	FMPI (10)	Pol	MPI	FMPI (6)	FMPI (10)	Pol
Mixture 1	0.729	0.455	0.391	0.095	0.261	0.195	0.235	0.056
Mixture 2	0.699	0.428	0.355	0.1	0.255	0.16	0.192	0.056
Mixture 3	0.449	0.232	0.188	0.095	0.05	0.065	0.069	0.052
Mixture 4	0.283	0.14	0.11	0.091	0.006	0.032	0.03	0.055
Mixture 5	0.213	0.116	0.085	0.098	0.003	0.023	0.025	0.054
Number of individuals	N=1000							
2-step adjustment	orthogonalization				std. variance correction			
Instruments	MPI	FMPI (6)	FMPI (10)	Pol	MPI	FMPI (6)	FMPI (10)	Pol
Mixture 1	0.945	0.762	0.734	0.108	0.712	0.454	0.576	0.061
Mixture 2	0.936	0.8	0.727	0.115	0.582	0.442	0.526	0.064
Mixture 3	0.75	0.483	0.38	0.103	0.238	0.178	0.21	0.056
Mixture 4	0.465	0.259	0.218	0.088	0.074	0.086	0.088	0.057
Mixture 5	0.349	0.188	0.156	0.093	0.036	0.078	0.072	0.056
Number of individuals	N=5000							
2-step adjustment	orthogonalization				std. variance correction			
Instruments	MPI	FMPI (6)	FMPI (10)	Pol	MPI	FMPI (6)	FMPI (10)	Pol
Mixture 1	1	0.997	0.997	0.165	0.997	0.994	0.998	0.037
Mixture 2	1	0.999	1	0.157	0.999	0.992	0.998	0.047
Mixture 3	0.997	0.991	0.982	0.115	0.986	0.921	0.945	0.049
Mixture 4	0.963	0.862	0.799	0.085	0.849	0.632	0.641	0.041
Mixture 5	0.92	0.765	0.707	0.076	0.671	0.594	0.54	0.041

Note: MPI corresponds to the infeasible MPI, $FMPI(6)$ corresponds to the feasible MPI with 6 points chosen in the support of F, same for $FMPI(10)$ with 10 instruments and Pol corresponds to the polynomial transformation of the exogenous variables.

7 Empirical Application

In this empirical application, we estimate a demand system for automobiles in Germany over the period 2012 to 2018. We follow the classical BLP demand model initiated by Berry (1994) and Berry et al. (1995), which is the standard demand model in markets with differentiated products. Berry and Haile (2021) and Miravete et al. (2023) underline the importance of the distribution of price sensitivity in determining key quantities of interestsuch as demand elasticity and curvature. In particular, Miravete et al. (2023) shows that the choice of the specification for the distribution of the RC associated with price can impose strong restrictions on the demand curvature and thus limits the range of the implied pass-through rate, which is key to assess the impact of tariffs or a cost shock on consumer welfare. Yet, the literature provides no guidance on the choice of the specification, which can vary substantially across applications.[27]

To address this gap, we estimate the model under different specifications for the distribution of the RC associated with price. We show that different specifications lead to substantially different results for the quantities of interest. In line with Miravete et al. (2023), our results indicate that the degenerate and normal specifications yield lower demand curvatures and pass-through rates. Additionally, we find these specifications also result in lower price elasticities. Next, we apply our test and selection procedure to identify the best-supported specification. We find that all specifications are rejected except the triweight kernel.

Data. We draw on several data sources. First, we use state-level new car registration data for Germany from 2012 to 2018, publicly available from the German Federal Motor Transport Authority. Next, we use data on car characteristics including horsepower, engine type, size, weight, fuel cost,...[28] We also observe state-year-level demographic variables, such as average household income and yearly average gas prices. We define a market as a state-year pair, a product as a brand, model, engine type, engine size, and body combination and the market size as the number of households in the market.[29] To instrument for price, we use real exchange rates and the distance between Germany and the country of assembly as cost shifters, along with traditional BLP instruments. Descriptive statistics and instrument details are in the Appendix D.[30]

[27] Researchers commonly use three distinct approaches to model heterogeneity in price sensitivity. First, they may normalize prices by the market-level average income and assume homogeneous price sensitivity across consumers (Berry et al. (1995); Grigolon et al. (2018); Beresteanu and Li (2011)). Second, they can multiply price to a random coefficients with known distributions, such as empirical income distributions (Miravete et al. (2018)), or interact it with demographic characteristics (Nevo (2001)). Third, they can use a parametric specification for the distribution of the RC like normal or log-normal Barahona et al. (2023); Conlon and Rao (2023).

[28] Data on car characteristics and prices were scraped from General German Automobile Club. The merged dataset was kindly provided by Kevin Remmy (`https://kevinremmy.com/research/`).

[29] In aggregating the products from the HSN/TSN level, we use the characteristics of the car with the highest sales. We focus on combustion engines and drop the sales of electric vehicles, which is on average 1% of the total car sales across markets. Among the remaining products, we keep the top 35% selling products that cover approximately 90% of the total sales.

[30] Real effective exchange rates retrieved from `https://databank.worldbank.org/reports.aspx?source=2&series=PX.REX.REER&country=OED`

Empirical specification. We follow the generic BLP demand model introduced in Example 3. We specify the indirect utility of consumer i, purchasing product j in market t as follows:

$$u_{ijt} = x'_{1jt}\beta + \xi^*_{jt} + \frac{p_{jt}}{inc_t}\alpha_i + \varepsilon_{ijt}.$$

The variables in x_{1jt} include a constant, horsepower, engine type (diesel or gasoline), fuel cost, size, height, a dummy for foreign brands, and an interaction between the time trend and the diesel dummy to account for the Dieselgate scandal. The product characteristics in x_{1jt} are assumed to exhibit no preference heterogeneity. The demand shock on product j is decomposed as follows: $\xi^*_{jt} = Brand_j + Class_j + State_t + trend_t + \xi_{jt}$, where $Brand_j$ is a brand fixed effect, $Class_j$ is a category fixed effect, $State_t$ captures time invariant state-specific demand shocks and $trend_t$ is a time trend to capture variation in the overall demand for cars. To keep our model parsimonious, we assume that only the price is associated with a RC. To account for the income effect, we normalize price by the average household income in a given market. We estimate the model assuming different specifications for the distribution of the price RC: (i) degenerate (i.e. logit), (ii) normal, (iii) gaussian mixture (iv) triweight (v) log-normal distribution. The specifications are summarized in Table 3.

Table 3: Parametric specifications for the price RC in the empirical application.

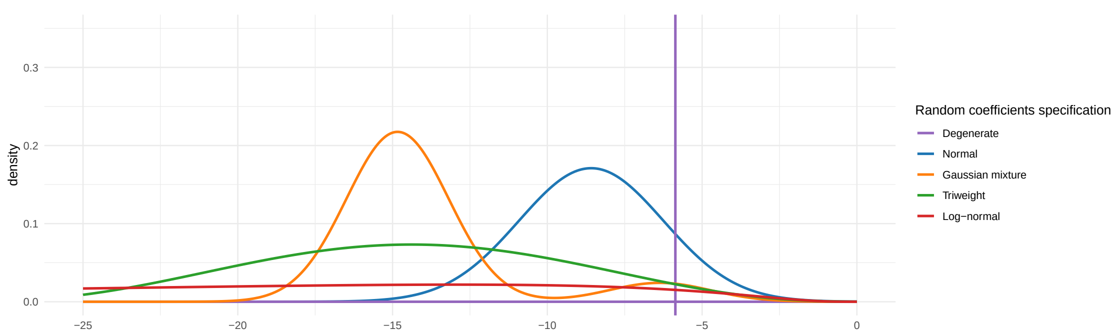
Distribution	Density	Parameters
Degenerate	$\delta(x-\mu)$	μ
Normal	$\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$	μ, σ
Mixture of Normals	$p\phi(x;\mu_1,\sigma) + (1-p)\phi(x;\mu_2,\sigma)$	μ_1, μ_2, σ, p
Triweight	$\frac{35}{32\sigma}\left(1-\left(\frac{x-\mu}{\sigma}\right)^2\right)^3 \mathbf{1}_{\lvert x-\mu\rvert\leq\sigma}$	μ, σ
$(-1)\times$Log-normal	$\frac{1}{x\sigma\sqrt{2\pi}}e^{-\frac{(\log x-\mu)^2}{2\sigma^2}}$	μ, σ

Results. We estimate the model via nested fixed-point GMM (see details in Appendix C.4.2). Table 4 reports the estimates of the parameters of the price RC for each specification, and Figure 1 plots the estimated distributions. The estimated distributions of price sensitivity vary significantly across specifications. The median consumer is substantially more price-sensitive under the triweight or Gaussian mixture specification than under the normal or degenerate one. Moreover, all non-degenerate specifications show significant heterogeneity in price sensitivity. The Gaussian mixture suggests the presence of at least two distinct modes in price sensitivity.[31] The estimates for the other car characteristics are presented in Table 14 in Appendix D.

Table 4: Estimated distributions of the price RC under different specifications

	Specification				
Parameter	Degenerate	Normal	Gaussian mixture	Triweight	Log-normal
Location 1	−5.86	−8.58	−14.84	−14.39	3.41
	(0.38)	(0.46)	(0.84)	(0.86)	(0.04)
Location 2			−6.37		
			(0.39)		
Mixture weight 1			0.90		
			(−)		
Dispersion		2.33	1.65	14.94	0.92
		(0.14)	(0.10)	(3.02)	(0.03)

Figure 1: Estimated distributions of the price RC under the different specifications



[31] For the Gaussian mixture, we apply a grid search for the mixing probability, and for each point in this grid, we estimate the model taking the probability as fixed. As a result, we cannot provide standard errors for the mixing probability. The Procedure is described in Appendix C.4.1.

For each specification, we use our estimation results to derive price elasticities, marginal costs, markups, demand curvatures and pass-through rates. To recover marginal costs and markups, we assume that multi-product firms pricing under Bertrand-Nash equilibrium. For the pass-through, we assume that the marginal cost of all the products increases by 1% and we recompute the equilibrium prices.[32] We report the median of the counterfactual quantities for the year 2018 in Table 5.[33]

Table 5: Median quantities of interest under different specifications for the RC

	Specification				
quantity of interest	Degenerate	Normal	Gaussian mixture	Triweight	Log-normal
Price Elasticity	-3.19	-3.3	-4.51	-3.65	-3.34
Marginal Cost (% of Price)	0.68	0.69	0.72	0.71	0.68
Markup	0.46	0.45	0.38	0.4	0.48
Demand Curvature	1	1.12	1.17	1.17	1.18
Pass-Through Rate	1	1.13	1.32	1.3	1.31

We can make several observations. First, the degenerate, normal, and log-normal specifications yield significantly higher price elasticities than the other two specifications. The Gaussian mixture specification results in the lowest price elasticity, which can be attributed to the presence of a group of highly price-sensitive consumers. These differences directly translate into lower markups and higher marginal costs under the Gaussian mixture and triweight specifications. Finally, in line with the findings of Miravete et al. (2023), we observe that the logit specification imposes a demand curvature and a pass-through rate equal to 1 for almost all products. The normal specification exhibits an overshifted median pass-through rate due to a demand curvature slightly greater than one. Lastly, the three remaining specifications produce higher demand curvatures and pass-through rates than the normal specification. These findings show that the choice of specification has important implications for policy analysis.

Specification test. To determine which specification to choose, we first apply the specification test developed in Section 4, which we denote Interval Test, for each spec-

[32]In Appendix D, we provide details on the calculation of the different counterfactual quantities.
[33]In Appendix D, we also report the empirical distirbution of the quantities of interest in Figure 4.

ification.[34] We report the results in Table 15. We also report the associated Sargan-Hansen J statistics.[35] The degrees of freedom vary in the Interval Test as colinear instruments are removed. Our test rejects the null hypothesis that the model is correctly specified for all the specifications except the triweight. In Appendix D, we also report the results of our model selection procedure (Table 15). The results are in line with the conclusions from the specification test. The triweight specification is consistently selected across comparisons. It is followed by the Gaussian mixture and log-normal distributions, for which we cannot reject equality of the slopes at the 5% level. The normal and degenerate distributions are the most strongly rejected specifications.

Table 6: Specification tests

	Interval test			J-test		
Specification	Statistic	Critical value	DF	Statistic	Critical value	DF
Degenerate	61.485	18.307	10	95.024	21.026	12
Normal	52.522	16.919	9	110.546	27.587	17
Gaussian mixture	22.293	14.067	7	105.888	26.296	16
Triweight	7.434	18.307	10	108.642	26.296	16
Log-normal	32.535	15.507	8	107.979	27.587	17

8 Conclusion

In this paper, we develop a simple yet powerful moment-based specification test for the distribution of random coefficients. We derive an expression for the instrument that maximizes the power of our test when the distribution of random coefficients is misspecified. This most powerful instrument is derived for both the simple hypothesis and the composite hypothesis cases. Next, we provide a feasible approximation of the MPI that can be derived without any knowledge of the alternative. Furthermore, we show that our test can be applied to a wide class of models featuring random coefficients and can be used to discriminate between different specifications. To evaluate

[34] In Appendix D, we provide greater detail on how the interval instruments are constructed.

[35] To facilitate the comparison, we keep the same set of estimation instruments across the different specifications except for the degenerate specification where no extra instruments are used to improve the identification of the distribution of RCs.

the performance of our test in finite sample, we conduct extensive sets of simulation experiments. Finally, we apply the test and model selection procedure developed in this paper to identify the specification of price sensitivity that is best supported by the data in our demand estimation for the German automobile market.

References

Ai, C. and X. Chen (2003): "Efficient estimation of models with conditional moment restrictions containing unknown functions," *Econometrica*, 71, 1795–1843.

Angrist, J. D. and G. W. Imbens (1995): "Two-stage least squares estimation of average causal effects in models with variable treatment intensity," *Journal of the American statistical Association*, 90, 431–442.

Arcidiacono, P. and R. A. Miller (2011): "Conditional choice probability estimation of dynamic discrete choice models with unobserved heterogeneity," *Econometrica*, 79, 1823–1867.

Backus, M., C. Conlon, and M. Sinkinson (2021): "Common ownership and competition in the ready-to-eat cereal industry," Tech. rep., National Bureau of Economic Research.

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

Barahona, N., C. Otero, and S. Otero (2023): "Equilibrium effects of food labeling policies," *Econometrica*, 91, 839–868.

Beresteanu, A. and S. Li (2011): "Gasoline Prices, Government Support, and the Demand for Hybrid Vehicles in the United States*," *International Economic Review*, 52, 161–182.

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 (2021): "Foundations of demand estimation," in *Handbook of industrial organization*, Elsevier, vol. 4, 1–62.

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

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

Bonnet, C. and P. Dubois (2010): "Inference on vertical contracts between manufacturers and retailers allowing for nonlinear pricing and resale price maintenance," *The RAND Journal of Economics*, 41, 139–164.

Bontemps, C. and N. Meddahi (2005): "Testing normality: a GMM approach," *Journal of Econometrics*, 124, 149–186.

Boyd, J. H. and R. E. Mellman (1980): "The effect of fuel economy standards on the US automotive market: an hedonic demand analysis," *Transportation Research Part A: General*, 14, 367–378.

Bunting, J. (2022): "Continuous permanent unobserved heterogeneity in dynamic discrete choice models," *arXiv preprint arXiv:2202.03960*.

Cardell, N. S. and F. C. Dunbar (1980): "Measuring the societal impacts of automobile downsizing," *Transportation Research Part A: General*, 14, 423–434.

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

Chernozhukov, V., D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins (2018): "Double/debiased machine learning for treatment and structural parameters," *The Econometrics Journal*, 21, C1–C68.

Compiani, G. (2022): "Market counterfactuals and the specification of multiproduct demand: A nonparametric approach," *Quantitative Economics*, 13, 545–591.

Conlon, C. and N. L. Rao (2023): "The cost of curbing externalities with market power: Alcohol regulations and tax alternatives," Tech. rep., National Bureau of Economic Research Cambridge, MA, USA.

Cox, D. R. (1961): "Tests of separate families of hypotheses," in *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability*, vol. 1, 96.

Davidson, R. and J. G. MacKinnon (1981): "Several tests for model specification in the presence of alternative hypotheses," *Econometrica*, 49, 781–793.

Duarte, M., L. Magnolfi, M. Sølvsten, and C. Sullivan (2024): "Testing firm conduct," *Quantitative Economics*, 15, 571–606.

Fosgerau, M. and M. Bierlaire (2007): "A practical test for the choice of mixing distribution in discrete choice models," *Transportation Research Part B: Methodological*, 41, 784–794.

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.

Fox, J. T., K. I. Kim, S. P. Ryan, and P. Bajari (2011): “A simple estimator for the distribution of random coefficients,” *Quantitative Economics*, 2, 381–418.

Fox, J. T., K. i. Kim, and C. Yang (2016): “A simple nonparametric approach to estimating the distribution of random coefficients in structural models,” *Journal of Econometrics*, 195, 236–254.

Gao, Z. and M. H. Pesaran (2023): “Identification and estimation of categorical random coefficient models,” *Empirical Economics*, 64, 2543–2588.

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

Godfrey, L. (1996): “Misspecification tests and their uses in econometrics,” *Journal of Statistical Planning and Inference*, 49, 241–260.

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

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.

Hall, A. R. and A. Inoue (2003): “The large sample behaviour of the generalized method of moments estimator in misspecified models,” *Journal of Econometrics*, 114, 361–394.

Hall, A. R. and D. Pelletier (2011): “Nonnested testing in models estimated via generalized method of moments,” *Econometric Theory*, 27, 443–456.

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

Hausman, J. A. (1978): “Specification Tests in Econometrics,” *Econometrica*, 46, 1251.

Heckman, J. J. and E. Vytlacil (2005): “Structural equations, treatment effects, and econometric policy evaluation 1,” *Econometrica*, 73, 669–738.

Hotz, V. J. and R. A. Miller (1993): “Conditional choice probabilities and the estimation of dynamic models,” *The Review of Economic Studies*, 60, 497–529.

Huber, P. J. and E. M. Ronchetti (2009): *Robust Statistics*, John Wiley & Sons, Inc.

Lehmann, E. L., J. P. Romano, and G. Casella (1986): *Testing statistical hypotheses*, vol. 3, Springer.

McCulloch, C. E. and S. R. Searle (2000): *Generalized, Linear, and Mixed Models*, Wiley.

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

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 (2023): *Elasticity and curvature of discrete choice demand models*, Centre for Economic Policy Research London, UK.

Nevo, A. (2001): "Measuring market power in the ready-to-eat cereal industry," *Econometrica*, 69, 307–342.

Newey, W. K. (1985): "Maximum Likelihood Specification Testing and Conditional Moment Tests," *Econometrica*, 53, 1047.

Newey, W. K. and D. McFadden (1994): "Large sample estimation and hypothesis testing," *Handbook of econometrics*, 4, 2111–2245.

Neyman, J. (1959): "Optimal asymptotic tests of composite hypotheses," *Probability and statsitics*, 213–234.

Neyman, J. and E. S. Pearson (1933): "IX. On the problem of the most efficient tests of statistical hypotheses," *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 231, 289–337.

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.

Rivers, D. and Q. Vuong (2002): "Model selection tests for nonlinear dynamic models," *The Econometrics Journal*, 5, 1–39.

Rust, J. (1987): "Optimal replacement of GMC bus engines: An empirical model of Harold Zurcher," *Econometrica: Journal of the Econometric Society*, 999–1033.

Sargan, J. D. (1958): "The estimation of economic relationships using instrumental variables," *Econometrica: Journal of the econometric society*, 393–415.

Shi, X. (2015): "A nondegenerate Vuong test," *Quantitative Economics*, 6, 85–121.

Smith, R. J. (1992): "Non-nested tests for competing models estimated by generalized method of moments," *Econometrica: Journal of the Econometric Society*, 973–980.

Train, K. E. (2001): *Discrete Choice Methods with Simulation*, Cambridge University Press.

Vuong, Q. and W. Wang (1993): "Selecting estimated models using chi-square statistics," *Annales d'Economie et de Statistique*, 143–164.

Vuong, Q. H. (1989): "Likelihood ratio tests for model selection and non-nested hypotheses," *Econometrica*, 57, 307–333.

Vytlacil, E. (2002): "Independence, monotonicity, and latent index models: An equivalence result," *Econometrica*, 70, 331–341.

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.

Zheng, J. X. (1996): "A consistent test of functional form via nonparametric estimation techniques," *Journal of Econometrics*, 75, 263–289.

Appendix

A Additional results

A.1 Identification

A.1.1 Non-parametric identification of the distribution of RCs in the BLP demand model

The BLP demand model (Example 3) is the empirical framework to which we devote the most attention in this paper. Thus, it is natural to provide a set of sufficient conditions under which the moment function m_3 non-parametrically identifies $\theta = (\beta, F)$. We show our result by exploiting a key finding in Wang (2022), which shows that if the demand functions are identified, then the distribution of RCs is identified. Before we state the assumptions under which m_3 identifies $\theta = (\beta, F)$ in our model, we provide a brief summary of the literature on the identification of the random coefficients in aggregate demand models (Fox and Gandhi (2011); Fox et al. (2012); Dunker et al. (2022); Wang (2022); Berry and Haile (2014); Allen and Rehbeck (2020)).

In their seminal paper, Berry and Haile (2014) show 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 characteristics has a coefficient that is not random and that is equal to 1. Let us underline that in the classical BLP model, the structure implied by the logit shock guarantees the invertibility of the demand functions.

Fox et al. (2012) provide 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) investigate 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 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}, and 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) study 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 more 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 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 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.[36] His proof exploits the real analytic property of the demand functions.[37]

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

[37] In particular, the real analytic property yields that the local identification of ρ on $\mathcal{D} \subset \mathbb{R}^J$ implies

Here, 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). Let us formally state the assumptions that we impose to recover the point identification of $\theta = (F, \beta)$.

Assumption A
(i) Strict exogeneity: $\forall j \;\; \mathbb{E}[\xi_{jt}|Z_t] = 0$ *a.s.;*
(ii) Completeness: for any measurable function g *such that* $\mathbb{E}[|g(Y_t, X_t)|] < \infty$*, if* $\mathbb{E}[g(Y_t, X_t)|Z_t] = 0$ *a.s., then* $g(Y_t, X_t) = 0$ *a.s.*;
(iii) The distribution of the data $(Y_t, X_{2t}, X_{1t}, Z_t)$ *is fully observed by the econometrician and market shares* Y_t *are generated by the demand model defined in Example 3 by equations (2.2) 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* $\mathbb{P}(X_t^T X_t$ *is positive definite*$) > 0$;
(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..*

Assumption A(i) is the classical conditional moment restriction on the unobserved demand shocks. Assumption A(ii) is a completeness assumption that states that the instruments are strongly relevant for (Y_t, X_t). This assumption is typical of semiparametric or nonparametric IV models and is 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.

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}.

Proposition 1.1 *Under Assumption A, the distribution of random coefficients F and the homogeneous preference parameters β are non-parametrically identified:*

$$(\tilde{F}, \tilde{\beta}) = (F, \beta) \iff \qquad \mathbb{E}[m_3(W_t, \tilde{\theta})|Z_t] = \mathbb{E}\left[\rho^{-1}(Y_t, X_{2t}, \tilde{F}) - X_{1t}\tilde{\beta}\middle|Z_t\right] = 0_J \quad a.s..$$

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. We conclude this section by adding that the strict exogeneity Assumption A(i) can be replaced by $\mathbb{E}[\sum_{j=1}^{J} \xi_{jt}|Z_t] = 0$ *a.s.*.[38] In words, we only require the demand shock ξ_{jt} to be mean independent of the instrumental variables Z_t across products, but we do not require this to hold for each product j taken separately. 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.

A.2 Bahadur's non-local approach

Bahadur's non-local approach to assess the power of a test. We now present the main ingredients to perform Bahadur's comparison approach. From Section 3.1.1, we have:

$$\text{Under } \overline{H}_0: \qquad S_N \equiv S_N(h, \theta_0) \xrightarrow[N\to+\infty]{P} S \qquad \text{with } \; S = \chi^2_L.$$

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:

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

The main idea in Bahadur's approach entails deriving the level of the test if one

[38] This is a result we showed in a previous version of the paper that circulated under the title "Testing and Relaxing Distributional Assumptions on Random Coefficients in Demand Models"

takes the value of the test statistic as the critical value (this is also known as the p-value). Namely:

$$\alpha_N = \Lambda(S_N).$$

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

Definition 4 (Asymptotic slope of the test)

(i) $K_N = -\frac{2}{N}\log(\Lambda(S_N))$ *is the approximate slope of the test,*

(ii) Under $\overline{H}_a$*: plim* $K_N = c^a(\theta_0)$ *is the asymptotic slope of the test,*

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

Under the alternative $\overline{H}_a : (\beta, F) = (\beta_a, F_a)$, consider two sequences of tests based on S_N^1 and S_N^2 with asymptotic slopes $c_1^a(\theta_0)$ and $c_2^a(\theta_0)$ respectively. The test based on S_N^1 is asymptotically preferred to the test based on S_N^2 in Bahadur's sense if and only if $c_1^a(\theta_0) > c_2^a(\theta_0)$. To derive the asymptotic slopes of our test, we apply an important result in Geweke (1981), which states that if under H_0: $S_N \underset{N\to+\infty}{\xrightarrow{d}} \chi_q^2$ (with any $q \in \mathbb{N}^*$), then $\frac{1}{N}S_N \underset{N\to+\infty}{\xrightarrow{P}} c^a(\theta_0)$ (when the limit exists). In our test, the limiting distribution is chi-squared. Thus, the asymptotic slope of our test with instrument $h(Z_i)$ can be written as follows:

$$c(h, \theta_0) = \text{plim } \frac{1}{N} S_N(h, \theta_0) = \mathbb{E}\left[h(Z_i)m(W_i, \theta_0)\right]^T \Omega_0^{-1} \mathbb{E}\left[h(Z_i)m(W_i, \theta_0)\right].$$

Bahadur's Non-Local Approach vs. Local Alternatives. The are several ways to compare the power of competing tests (see Gourieroux and Monfort (1995) for a comprehensive review). The main alternative to Bahadur's non-local approach is to discriminate between two tests based on their power against local alternatives. In a parametric framework, the researcher considers a sequence of local alternatives θ_N that converges to θ_0 at a given rate (usually $\frac{1}{\sqrt{N}}$). The econometrician can then compare two competing tests using their asymptotic power (or more precisely, the limits of the power functions when the sample size goes to $+\infty$) against this sequence of local alternatives.

We argue that Bahadur's non-local approach is more appropriate for the testing problem addressed in this paper. First, 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 local criterion when the test concerns non-parametric objects such as distributions. Moreover, we study the properties of our test under a fixed alternative $\overline{H}_a : (\beta, F) = (\beta_a, F_a)$ that is not necessarily local. Finally, the literature has highlighted several 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 $\overline{H}_0$ may perform poorly away from $\overline{H}_0$.

A.3 Asymptotic equivalence between $h^{\perp}$ and its empirical counterpart

In practice, the best linear predictor is not known but estimated and its derivation also involves a first-stage estimator of ϕ_0. This raises concerns regarding how this estimation might affect the asymptotic distribution of our test statistic. However, we show that under explicit mild regularity conditions described below, both test statistics share the same asymptotic distribution. These conditions impose slightly greater demands on the smoothness of the function m compared to those in Assumption 3; specifically, we require m to be twice-differentiable with respect to ϕ. Moreover, we require stronger boundedness conditions on the derivatives of m. Additionally, in Assumption 6, we require that Λ satisfies a finite moment condition.

Assumption 5 (Regularity conditions on m) *The following conditions hold with probability one.*

1. *$m(W_i, \phi)$ is twice-continuously differentiable in ϕ for all $\phi \in \Phi$.*
2. *$\exists M(\cdot)$ and $\overline{M} \in \mathbb{R}_+$ such that $j \in [|1, p|]$, $\left|[m(W_i, \phi)]_j\right| \leq M(W_i)$ with $\mathbb{E}[M(W_i)^4] < \overline{M}$ for all $\phi \in \Phi$.*
3. *$\exists \tilde{M}(\cdot)$ and $\overline{\tilde{M}} \in \mathbb{R}_+$ such that $\forall (k, j) \in [|1, p|] \times [|1, dim(\Phi)|]$, $\left|\left[\frac{\partial m(W_i,\phi)}{\partial \phi^T}\right]_{k,j}\right| \leq \tilde{M}(W_i)$ where $\mathbb{E}[\tilde{M}(W_i)^4] < \overline{\tilde{M}}$ for $\phi \in \Phi$.*

4. $\exists \tilde{\tilde{M}}(\cdot)$ *and* $\overline{\tilde{\tilde{M}}} \in \mathbb{R}_+$ *such that* $\forall j \in [|1, dim(\Phi)|]$, $\forall (k,r) \in [|1,p|] \times [|1, dim(\Phi)|]$, $\left| \left[\frac{\partial^2 m(W_i,\phi)}{\partial \phi_j \partial \phi^T} \right]_{k,r} \right| \leq \tilde{\tilde{M}}(W_i)$ *where* $\mathbb{E}[\tilde{\tilde{M}}(W_i)^4] < \overline{\tilde{\tilde{M}}}$ *for all* $\phi \in \Phi$.

5. $\forall j \in [|1, dim(\Phi)|]$, $\frac{\partial}{\partial \phi_j} \mathbb{E}\left[\frac{\partial m(W_i,\phi)}{\partial \phi^T} \middle| Z_i \right] = \mathbb{E}\left[\frac{\partial^2 m(W_i,\phi)}{\partial \phi_j \partial \phi^T} \middle| Z_i \right]$ *for all* $\phi \in \Phi$.

6. $\exists C'''$ *such that* $\mathbb{E}\left[\left\| \mathbb{E}\left[\frac{\partial m(W_i,\phi_0)}{\partial \phi} \middle| Z_i \right] \Lambda(Z_i)^2 m(W_i, \phi_0) \right\|_F^2 \right] < C'''$.

Assumption 6 (Regularity conditions on Λ) *The weighting matrix* Λ *satisfies the following restriction:* $\exists \bar{\Lambda}$, $\mathbb{E}[\|\Lambda(Z_i)\|_F^4] < \bar{\Lambda}$.

Proposition 1.2 *Assume that* $h \in \mathcal{H}_0$ *and satisfies Assumption 2 (1). Moreover, Assumption 4 is satisfied,* $\hat{\phi}$ *satisfies Assumption 2,* m *satisfies Assumption 5 and* Λ *satisfies 6, then the following property holds.*

- *Under* $H_0: \ F \in \mathcal{F}_0$,

$$\frac{1}{\sqrt{N}} \sum_{i=1}^{N} \hat{h}_\Lambda^\perp(Z_i) m(W_i, \hat{\phi}) = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} h_\Lambda^\perp(Z_i) m(W_i, \phi_0) + o_p(1).$$

- *Under* $\overline{H}'_a : (F, \beta) = (F_a, \beta_a)$ *with* $F_a \notin \mathcal{F}_0$ *and* $\mathbb{E}\left[h_\Lambda^\perp(Z_i) m(W_i, \phi_0) \right] \neq 0$,

$$\frac{1}{N} \sum_{i=1}^{N} \hat{h}_\Lambda^\perp(Z_i) m(W_i, \hat{\phi}) \xrightarrow[N \to +\infty]{P} \mathbb{E}\left[h_\Lambda^\perp(Z_i) m(W_i, \phi_0) \right].$$

with $h_\Lambda^\perp$ and $\hat{h}_\Lambda^\perp$ respectively the population and the sample orthogonalized counterparts of h as in (3.6).

The previous proposition indicates that the orthogonalization procedure can be carried out in practice without affecting the asymptotic properties of the test statistic. We prove this result in Appendix B.2.2.

A.4 An alternative approach for the orthogonalization of the instruments

In this section, we present an alternative way of orthogonalizing the set of instruments that, contrary to the main approach, does not require deriving or estimating $\mathbb{E}\left[\frac{\partial m(W_i,\phi_0)}{\partial \phi^{\perp}}|Z_i\right]$, which can be difficult in some models such as the BLP demand model. Let $h \in \mathcal{H}_0$ and let us further assume that $L > dim(\phi)$. We define $h^{\dagger}$ as the orthogonalized version of h, obtained using the alternative procedure described below.

$$
\begin{aligned}
h^{\dagger}(Z_i) \equiv h(Z_i) - BLP\left(h(Z_i)\middle|BLP\left(\frac{\partial m(W_i,\phi_0)}{\partial \phi}\middle|h(Z_i)^T\right)^T\right) \\
\equiv BLP^{\perp}\left(h(Z_i)\middle|BLP\left(\frac{\partial m(W_i,\phi_0)}{\partial \phi}\middle|h(Z_i)^T\right)^T\right),
\end{aligned}
$$

where $BLP(\cdot|\cdot)$ denotes the best linear predictor, which we define as follows. For any random matrices $(A_i, B_i) \in \mathcal{M}_{k,p}(\mathbb{R}) \times \mathcal{M}_{p,m}(\mathbb{R})$ with $(k,m) \in \mathbb{N}^{*2}$, $BLP(A_i|B_i) = \mathbb{E}(A_iB_i)\mathbb{E}(B_i^TB_i)^{-1}B_i^T$. This approach consists of a double linear projection. First, we project $\frac{\partial m(W_i,\phi_0)}{\partial \phi}$ on the space spanned by $h(Z_i)$ and then we project $h(Z_i)$ onto the orthogonal space to $BLP\left(\frac{\partial m(W_i,\phi_0)}{\partial \phi}\middle|h(Z_i)^T\right)^T$. Therefore, $h^{\dagger}(Z_i)$ is a linear transformation of the instruments $h(Z_i)$, which ensures that they are orthogonal to $\frac{\partial m(W_i,\phi_0)}{\partial \phi^T}$. Next, let us state some classical regularity conditions that ensure the existence of $h^{\dagger}$.

Assumption 7 (Regularity conditions for the existence of $h^{\dagger}$)

- $\exists C^{\dagger}$ *such that* $\mathbb{E}\left[\|h(Z_i)\|_F^4\right] + \mathbb{E}\left[\|\frac{\partial m(W_i,\phi_0)}{\partial \phi^T}\|_F^2\right] < C^{\dagger}$.
- $\mathbb{E}\left[h(Z_i)h(Z_i)^T\right]$ *is full rank.*
- $\mathbb{E}\left[\frac{\partial m(W_i,\phi_0)}{\partial \phi}h(Z_i)^T\right]$ *is full rank.*

The next proposition shows that under Assumptions 3 and 7, the orthogonalized instrument $h^{\dagger}$ as defined above exists and satisfies the Neyman orthogonality property.

Proposition 1.3 *(Neyman orthogonality for $h^\dagger$)*

Let $h \in \mathcal{H}_0$. Under Assumption 3 and 7, $h^\dagger$ exists and the moment $\mathbb{E}\left[h^\dagger_\Lambda(Z_i)m(W_i,\phi_0)\right]$ is Neyman orthogonal with respect to $\phi = (\beta, \lambda)$. Moreover, $h^\dagger(Z_i) = (I_L - \Gamma_h)h(Z_i)$ with Γ_h a projection in $\mathbb{R}^L$ and $rank(\Gamma_h) = dim(\phi)$.

We prove this proposition in Appendix B.2.2. It follows from the previous proposition that $h^\dagger(Z_i)$ is a projection of $h(Z_i)$ onto a linear subspace of $\mathcal{L}(h(Z_i))$ that is of rank $L - dim(\phi)$. Hence, we can show that $\Omega_0 = \mathbb{E}\left[(h^\dagger(Z_i)m(W_i,\phi_0))(h^\dagger(Z_i)m(W_i,\phi_0))^T\right]$ can be written as follows:

$$\Omega_0 = (I_L - \Gamma_h)\mathbb{E}\left[(h(Z_i)m(W_i,\phi_0))(h(Z_i)m(W_i,\phi_0))^T\right](I_L - \Gamma_h^T) \equiv (I_L - \Gamma_h)\tilde{\Omega}_0(I_L - \Gamma_h^T).$$

As $h \in \mathcal{H}_0$, it follows that Ω_0 is of rank $L - dim(\phi)$. To account for the fact that Ω_0 is no longer invertible, we must slightly change the definition of our test statistic by taking the pseudo-inverse instead of the inverse. Namely, we consider the following test statistic:

$$\tilde{S}_N(h,\phi) = N\left(\frac{1}{N}\sum_{i=1}^N h(Z_i)m(W_i,\phi)\right)^T \hat{\Omega}_0^+ \left(\frac{1}{N}\sum_{i=1}^N h(Z_i)m(W_i,\phi)\right),$$

where $(\cdot)^+$ denotes the Moore-Penrose inverse with $\hat{\Omega}_0$ a consistent estimator of Ω_0. The next proposition describes the asymptotic behavior of our modified test statistic when we use $\hat{h}^\dagger$ as an instrument (the empirical counterpart of $h^\dagger$). In particular, under the usual regularity assumptions, our test is asymptotically pivotal regardless of whether we use ϕ_0 or $\hat{\phi}$ to compute $m(W_i,\phi)$.

Before we state the main asymptotic result, we need a final regularity condition on the rank of the empirical estimators of $\tilde{\Omega}_0$ and Γ_h. These conditions ensure that we can apply a result in Andrews (1987), which allows us to recover a chi-square asymptotic distribution under $\overline{H}_0$ even if we consider a generalized inverse for the weighting matrix.

Assumption 8 (Rank condition)

$$\mathbb{P}\left(\{rank(\hat{\tilde{\Omega}}_0) = L\} \cap \{rank(\hat{\Gamma}_h) = dim(\phi)\}\right) \underset{N\to+\infty}{\longrightarrow} 1.$$

Proposition 1.4 *Assume that $h \in \mathcal{H}_0$, $\hat{\phi}$ satisfies Assumption 2, and m satisfies Assumption 3, Assumption 7 and 8 are satisfied, then the following properties hold:*

- *Under $H_0 : F \in \mathcal{F}_0$,* $\qquad \tilde{S}_N(\hat{h}^{\dagger}, \hat{\phi}) \xrightarrow[N\to+\infty]{d} \chi^2_{L-dim(\phi)}$,
- *Under $\overline{H}'_a : (I_L - \Gamma_h)\mathbb{E}\left[h(Z_i)m(W_i, \phi_0)\right] \neq 0$,*

$$\frac{1}{N}\sum_{i=1}^{N} \hat{h}^{\dagger}(Z_i)m(W_i, \hat{\phi}) \xrightarrow[N\to+\infty]{P} \mathbb{E}\left[h^{\dagger}(Z_i)m(W_i, \phi_0)\right].$$

where a consistent estimator of Ω_0 is simply the empirical variance-covariance matrix obtained by replacing ϕ_0 with $\hat{\phi}$ and $\hat{h}^{\dagger} = (I_L - \hat{\Gamma}_h)h$, with $\hat{\Gamma}_h$ the empirical counterpart of Γ_h.

We prove this proposition in Appendix B.2.2. Several remarks are in order. First, in comparison with the asymptotic results in the general case delineated in Proposition 3.4, the test statistic converges to a chi-square distribution with $L - dim(\phi)$ degrees of freedom. This reduction in degrees of freedom, akin to the Sargan-Hansen overidentification test, arises due to the projection. Second, as $h^{\dagger}(Z_i) = (I_L - \Gamma_h)h(Z_i)$ is a linear transformation of $h(Z_i)$, it is sufficient to provide conditions under which $\hat{\Gamma}_h \xrightarrow[N\to+\infty]{P} \Gamma_h$ to replace $h^{\dagger}$ with its empirical counterpart. In particular, these conditions are less stringent compared to those outlined in Assumption 5 necessary for replacing $h^{\perp}_{\Lambda}$ to be replaced by $\hat{h}^{\perp}_{\Lambda}$. Lastly, for practical reasons, we give a closed-form expression for Γ_h. One can easily show that:

$$\Gamma_h = \mathbb{E}\left[h(Z_i)\frac{\partial m(W_i, \phi_0)}{\partial \phi^T}\right]\left(\mathbb{E}\left[\frac{\partial m(W_i, \phi_0)}{\partial \phi}h(Z_i)^T\right]\mathbb{E}\left[h(Z_i)h(Z_i)^T\right]^{-1}\mathbb{E}\left[h(Z_i)\frac{\partial m(W_i, \phi_0)}{\partial \phi^T}\right]\right)^{-1} \mathbb{E}\left[\frac{\partial m(W_i, \phi_0)}{\partial \phi}h(Z_i)^T\right]\mathbb{E}\left[h(Z_i)h(Z_i)^T\right]^{-1}.$$

A.5 A feasible approximation of the MPI

A.5.1 Practical considerations for the derivation of the interval instruments.

In practice, to derive the interval instruments, the researcher must guess the support of F_a and choose a collection of points within this support. We now discuss how these choices affect the properties of our test.

Guess on the support of F_a. The support of F_a is in general unknown and the researcher must make an educated guess. Often, the test is preceded by a first-stage estimator, and the estimated parametric distribution can assist in refining this guess. This is a practice we follow in the simulations. Furthermore, since our test is straightforward to implement, conducting robustness checks by testing different sets of points is easily achievable.

Location of the points. The choice of the points does not affect the validity of our test but will affect its power. To see this, we consider the simple mixed logit case (Example 2) and we further assume homoscedasticity, $\beta_0 = \beta_a$ and that F_0 and F_a have the same finite support $\{v_1, ..., v_L\}$. In this simple example, the correction term is written:

$$\mathbb{E}[\Delta^m_{0,a}(W_i)|Z_i] = \sum_{l=1}^{L}(\omega_l(F_0) - \omega_l(F_a)) \underbrace{\frac{\exp\{X_1\beta_a + X_2 v_l\}}{1 + \sum_{k=1}^{J}\exp\{X'_{1j}\beta_a + X'_{2j}v_l\}}}_{\pi_l(Z_i)}$$

In this case, $BLP(\Delta^m_{0,a}(W_i)|\Pi(Z_i)) = \mathbb{E}[\Delta^m_{0,a}(W_i)|Z_i]$ and the slope of the test is equal to:

$$\begin{aligned} c(\Pi, \theta_0) = c(h^*_a, \theta_0) &= \sum_j \mathbb{E}\left(\sum_{l=1}^{L}(\omega_l(F_0) - \omega_l(F_a))\pi_{lj}(Z_i)\right)^2 \\ &= \sum_j \sum_{l,l'}(\omega_l(F_0) - \omega_l(F_a))(\omega_{l'}(F_0) - \omega_{l'}(F_a))\mathbb{E}[\pi_{lj}(Z_i)\pi_{l'j}(Z_i)] \end{aligned}$$

One can immediately see that the contribution of instrument $\pi_l(Z_i)$ to the slope of the test is equal to:

$$(\omega_l(F_0) - \omega_l(F_a))\sum_j \sum_{l'}(\omega_{l'}(F_0) - \omega_{l'}(F_a))\mathbb{E}[\pi_{lj}(Z_i)\pi_{l'j}(Z_i)]$$

In particular, the contribution of instrument l is proportional to $\omega_l(F_0) - \omega_l(F_a)$ which is the difference in probability masses at point v_l between F_0 and F_a. One may observe that if this difference is zero, then the contribution of this instrument is zero. If we extrapolate this result to continuous distributions, a point v_l yields a powerful instrument $\pi_l(Z_i)$ if the integrated difference in distributions in the interval around this point is large. One important insight we can draw from this exercise is that points in the domain of definition of F_a where f_0 and f_a are close, won't generate powerful instruments.

Number of points. The choice of the number of instruments L is subject to the following trade-off. On the one hand, increasing the number of instruments L allows the researcher to increase the slope of the test. Indeed, as we saw in Section A.5.3 when the discretization of the integral gets finer, the Riemann approximation of the integral gets closer to the true integral, and thus, the slope gets closer to the slope reached by the MPI. On the other hand, increasing the number of instruments can pose important problems in finite sample. First, it increases finite-sample bias and/or variance. Second, it is well known in the literature that a large number of moments can hinder the estimation of the variance-covariance matrix, potentially leading to substantial size distortions in finite samples. Several finite-sample corrections, such as the one proposed by Windmeijer (2005), have been introduced in the literature and could be successfully exploited in our context. Moreover, a fine discretization can create some colinearity problems between the instruments. The importance of this issue can be assessed empirically. Finally, an increase in the number of instruments entails an increase in the degrees of freedom in the asymptotic chi-square distribution and an inflation of the critical value, which can be detrimental in terms of power. We defer the determination of the optimal number of instruments to future research.

A.5.2 The slope of the test as a measure of the predictive power of the instruments on the residuals: the general case

Here, we state a result analog to Proposition 4.1, which provides an alternative interpretation of the asymptotic slope in the general case (without assuming homoskedasiticity). Namely, the next proposition shows that the slope associated with $h(Z_i)$ is equal to the expected squared norm of the projection of the standardized "residuals" $\Sigma(Z_i)^{-1/2}m(W_i,\theta_0)$ on the space spanned by the set of instruments $\Sigma(Z_i)^{1/2}h(Z_i)$ with $\Sigma(Z_i) = \mathbb{E}\left[m(W_i,\theta_0)m(W_i,\theta_0)^T|Z_i\right]$

Proposition 1.5 (Alternative intepretation of the asymptotic slope in the general case)

Let $h \in \mathcal{H}_0$, then: $c(h,\theta_0) = \mathbb{E}\big[||BLP((\Sigma(Z_i)^{-1/2}m(W_i,\theta_0))^T|\Sigma(Z_i)^{1/2}h(Z_i)^T)||_2^2\big]$

with $\Sigma(Z_i) = \mathbb{E}\left[m(W_i,\theta_0)m(W_i,\theta_0)^T|Z_i\right]$

The proof of this proposition is in Appendix B.3.

A.5.3 Towards a Sieve approximation of the MPI

In this section, we provide a preliminary analysis of the case where the number of interval instruments, denoted by $L \equiv L_N$, increases with the sample size. Since this topic is peripheral to our paper, we provide intuitive explanations and limit our discussion to the homoscedastic case without presenting formal results. Letting L diverge with the sample size offers two main advantages relative to taking a fixed number of instruments. First, it can restore the consistency of the test, which is achieved by taking the infeasible MPI as the instrument and under the assumption that m identifies nonparametrically θ. Second, allowing L to diverge enables the slope of the test obtained with the feasible MPI to reach the optimal slope generated by the MPI and possibly increase the power of the test. Nevertheless, as discussed in Appendix A.5.1, the impact on finite-sample power is ambiguous, as increasing L simultaneously increases the slope and inflates the critical value.

When we let the number of interval instruments grow as the sample size increases, the linear approximation of the MPI using interval instruments can be interpreted as a Sieve estimation of the MPI. Following the notations of Chen (2007), under $\overline{H}_a$, the homoscedastic MPI $h_a^*(Z_i)$ equals $\mathbb{E}[m(W_i,\theta_0)|Z_i]$, which is the unique minimizer of the following population criterion in the space of square-integrable functions of Z_i, denoted by $\mathcal{G}$:

$$\mathcal{Q}(g) = \mathbb{E}\left[\|m(W_i,\theta_0) - g(Z_i)\|_2^2\right]$$

Since it is computationally infeasible to solve for h_a^*, the Sieve method replaces $\mathcal{G}$ with a Sieve space where the minimization is much easier and whose complexity increases with the sample size. In accordance with our Riemann sum approximation of the integral, it is natural to consider the following linear and finite dimensional sieve space $\mathcal{G}_N$:

$$\mathcal{G}_N = \left\{ g : \mathcal{Z} \to \mathbb{R}^p, g(z) = \sum_{k=1}^{L_N} \gamma_k \pi_k(z)\ ,\ \gamma_1, ..., \gamma_{L_N} \in \mathbb{R} \right\},$$

with $dim(\mathcal{G}_N) = L_N$ increases slowly as $N \to +\infty$. The series estimator $\hat{g}$ has the following closed-form expression:

$$\hat{h}(z) = \Pi(z)^T \hat{\Gamma}_{L_N} \qquad \text{with } \hat{\Gamma}_{L_N} = \left(\sum_{i=1}^{N} \Pi_{L_N}(Z_i) \Pi_{L_N}(Z_i)^T \right)^{-} \sum_{i=1}^{N} \Pi_{L_N}(Z_i) m(W_i, \theta_0).$$

With $()^{-}$ the Moore–Penrose generalized inverse. Under high-level regularity conditions, such as those in Chen (2007), one can show that for any $\varepsilon > 0$,

$$\mathbb{P}\left(\left| \mathcal{Q}(\hat{h}) - \mathcal{Q}(h^*) \right| > \varepsilon \right) \underset{N \to +\infty}{\longrightarrow} 0. \tag{A.8}$$

The convergence result above implies that the MPI and the interval instruments (with increasing L) have the same slope. Using analog arguments as in Fox et al. (2016) to show the consistency of their non-parametric estimator for the distribution of RCs, it is possible to derive sufficient conditions for our Riemann sum approximation to satisfy the assumptions in Chen (2007). Typically, these conditions include that F_a is absolutely continuous and has a bounded support, as then the Riemann sum approximation can get arbitrarily close to the true integral. However, analogous results can be derived for less restrictive classes of distributions by employing alternative methods to approximate the integral.

Regarding consistency, under any fixed alternative, the convergence result in (A.8) implies that

$$\frac{1}{N} S_N(\Pi_{L_N}, \theta_0) \underset{N \to +\infty}{\overset{P}{\longrightarrow}} c(h^*, \theta_0) > 0,$$

which is positive if m non-parametrically identifies θ, proving consistency.

It remains unclear whether the asymptotic distribution of our test statistic under H_0, with an increasing number of moments and a potentially singular weighting matrix, remains chi-square distributed. To our knowledge, this precise setting has not been studied in the literature and could be investigated in a separate paper.

A.5.4 A feasible approximation of the MPI when the correction term is not discretizable

In the main text, we focus on the case where the correction term is discretizable. However, in some cases such as the BLP demand model (Example 3), the correction term is not discretizable. For these cases, we propose to linearize the correction term locally around F_0. To do so, we define the concept of locally discretizable:

Definition 5 *We say that $\Delta^m_{0,a}$ is locally discretizable if for any θ_0, there exists an explicit function ψ_0 such that:*

$$\Delta^m_{0,a}(W_i) = \int_{\mathcal{V}} \psi_0(W_i, v, \beta_a) dF_a(v) + o(\|F_0 - F_a\|_1) \tag{A.9}$$

In words, the correction is term is locally discretizable if the first term in the Taylor expansion around F_0 is itself discretizable. In the case of the BLP demand model, we can show that the correction term is locally discretizable by exploiting a local expansion of $\Delta^{m_3}_{0,a}(W_i)$ "around F_0", which is recovered by exploiting the properties of the inverse demand function, which is both $\mathcal{C}^\infty$ and bijective in Y_t.

Proposition 1.6
A first order expansion of $\Delta^{m_3}_{0,a}(W_t)$ around F_0 writes:

$$\Delta^{m_3}_{0,a}(W_t) = X_{1t}(\beta_0 - \beta_a) + \left(\frac{\partial \rho(\delta^0_t, X_{2t}, F_0)}{\partial \delta}\right)^{-1} \int_{\mathcal{V}} \left[\frac{\exp\{\delta^0_t + X_{2t} v\}}{1 + \sum_{k=1}^J \exp\{\delta^0_{kt} + X'_{2kt} v\}} - \rho(\delta^0_t, X_{2t}, F_0)\right] dF_a(v) + \mathcal{R}_0,$$

with $\delta^0_t = \rho^{-1}(Y_t, X_{2t}, F_0)$ and $\mathcal{R}_0 = o\big(\int_{\mathcal{V}} |dF_a(v) - dF_0(v)|\big)$.

The proof is in Appendix B.3.2. We first observe that for any distribution F_0, we can construct artificial market shares Y^0_t such that $\rho^{-1}(Y_t, X_{2t}, F_a) = \rho^{-1}(Y^0_t, X_{2t}, F_0)$. Then, we recover the final result by taking a Taylor expansion of $\rho^{-1}(Y^0_t, X_{2t}, F_0)$ around Y_t and showing that the remainder is bounded.[39] 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.

A.6 Overview of Model Selection Methods

Several approaches exist in the literature for selecting between two non-nested models. One set involves maximizing goodness-of-fit criteria such as the R^2, the adjusted R^2, or the Akaike information criterion, but these have noTable theoretical limitations. In particular, they are not broadly applicable; the R^2 increases mechanically with the complexity of the model; the Akaike information criterion and the adjusted R^2 are not

[39]The expansion is taken around Y_t because Y^0_t depends on F_a and is thus unknown to the researcher.

consistent in the sense that they do not select the correct model with a probability that converges to 1.

Another set involves testing the models against each other, with well-known procedures including those proposed in Cox (1961), Davidson and MacKinnon (1981), and Vuong (1989). In this paper, we adopt the framework initially proposed in Vuong (1989) and subsequently expanded by Rivers and Vuong (2002). Vuong (1989) introduces a test for comparing the fit of two non-nested models estimated via quasi-maximum likelihood, using the Kullback-Leibler distance to measure the lack of fit. Rivers and Vuong (2002) extend this approach to a wider range of criterion functions and models, encompassing those defined by moment conditions.[40] Crucially, Vuong (1989) and Rivers and Vuong (2002) allow for both of the competing models to be misspecified, rejecting the null hypothesis if either model diverges further from the true data-generating process based on a lack-of-fit criterion chosen by the researcher. In contrast, a major drawback of the two other testing approaches is that they require either the model under the null or the alternative to be the true model. When both models are misspecified, they are hardly interpretable and may lead to some inconsistencies when H_1 and H_2 are successively chosen to be the null hypothesis.

A.7 Model selection: a sequential procedure when the models are nested

Sequential procedure. As we emphasized in Section 5, if two different specifications on the distribution of RCs $\mathcal{F}_1$ and $\mathcal{F}_2$ yield the same distributions F_{1,λ_1} and F_{2,λ_2}, the variance σ_{RV} becomes degenerate (again, for simplicity, we assume the competing models differ solely in the way the distribution of random coefficients is specified).

This issue can arise when the competing specifications are nested or overlap, meaning some parameters in both models yield the same distributions. To address this problem, we adopt a sequential procedure similar to the one proposed in Vuong (1989) in the case of overlapping specifications. This procedure is applicable regardless of the estimation method used. Let us note that when estimation is performed via maximum likelihood, and under classical regularity conditions, an LR test for nested models can

[40] Let us observe that Vuong and Wang (1993) already extend Vuong (1989) to a Pearson chi-squared test statistic.

be conducted following the procedure proposed in Vuong (1989). However, one key regularity condition that limits the applicability of this approach is that the pseudo-true value must lie in the interior of the parameter space. This excludes, for example, the case where we test a family of degenerate RCs against a family of Gaussian RCs. Under $H_0 : \sigma = 0$, the asymptotic distribution of the test statistic becomes non-standard, making the usual LR test inapplicable (see Andrews (1999) on the derivation of the asymptotic distribution of an extremum estimator when the true parameter lies on the boundary of the parameter space).

Now we describe the sequential procedure in the case of nested specifications. Without loss of generality, we assume that $\mathcal{F}_2 \subset \mathcal{F}_1$. We propose a sequential procedure which works as follows:

1. First, we test whether the distributions estimated under each specifications are the same. This test can be conducted via a classical parametric Wald test.[41]

2. If H_0^{pre} is not rejected, we can conclude that F_{1,λ_1} and F_{2,λ_2} cannot be distinguished given the data, and a fortiori, we cannot reject that the slopes are equal. If H_0^{pre} is rejected, then we can test for the equality of the slopes $H_0^{RV} : C_1 = C_2$ using the procedure outlined in section 5.

Control of the size. Let α_{pre} the significance level used to perform the pre-test and α_{RV} the significance level for the Rivers and Vuong test. As in Vuong (1989), we can show that the significance level of the sequential test for equality of the slopes presented above is asymptotically bounded by the maximum of α_{pre} and α_{RV}. Indeed, $H_0 = H_0^{pre} \cup (H_0 - H_0^{pre})$ (in particular, this is implied by the fact that $H_0^{pre} \subset H_0$). It follows that:

$$\begin{aligned}
\Pr[\text{reject } H_0 \mid H_0] &= \Pr[\text{reject } H_0^{pre} \cap \text{reject } H_0^{RV} \mid H_0] \\
&\leq \max\{\Pr(\text{reject } H_0^{pre} \cap \text{reject } H_0^{RV} \mid H_0^{pre}), \Pr(\text{reject } H_0^{pre} \cap \text{reject } H_0^{RV} \mid H_0 - H_0^{pre})\} \\
&\leq \max\{\Pr(\text{reject } H_0^{pre} \mid H_0^{pre}), \Pr(\text{reject } H_0^{RV} \mid H_0 - H_0^{pre})\}.
\end{aligned}$$

Under the usual regularity conditions, $\Pr(\text{reject } H_0^{pre} \mid H_0^{pre}) \underset{N\to+\infty}{\longrightarrow} \alpha_{pre}$. Furthermore, under Proposition 5.1, $\Pr(\text{reject } H_0^{RV} \mid H_0 - H_0^{pre}) \underset{N\to+\infty}{\longrightarrow} \alpha_{RV}$. Thus,

[41] The Wald test applies more broadly than the LR test or the Score test

the significance level of the sequential testing procedure is asymptotically at most $\max\{\alpha_{pre}, \alpha_{RV}\}$.

Pre-test in the Gaussian mixture case. In our simulation experiments and our empirical application, we encounter the case where the two nested specifications are Gaussian mixtures with repectively L and $L+1$ components. Here, we briefly introduce a pre-test to test whether the two distributions are the same. Let us denote $\mathcal{F}_L$ and $\mathcal{F}_{L+1}$, the families of mixtures with respectively L and $L+1$ components. Let λ_L and λ_{L+1} the pseudo-true values obtained under each specification. We want to test: H_0^{pre} : $F_{\lambda_L} = F_{\lambda_{L+1}}$. This hypothesis can be difficult to test in the general case. To address this, we propose to replace H_0^{pre} by a more conservative hypothesis $H_0^{pre'} : F_{\lambda_{L+1}} \in \mathcal{F}_L$. Let us observe that these hypothesis are equivalent in the case of maximum likelihood or if the tested moment conditions are the same in the context of GMM.

In the general setting, since $H_0^{pre'} \not\subseteq H_0$, substituing $H_0^{pre'}$ for H_0^{pre} results in a more conservative null hypothesis $H_0' : H_0 \cup H_0^{pre'}$. The sequential procedures remains unchanged; we simply replace H_0 by H_0' and H_0^{pre} by $H_0^{pre'}$. One can easily show that the size control for H_0' is preserved by using the same line of reasoning as for H_0. Next, we provide a test for $H_0^{pre'}$. In both, the simulations and the empirical application, we assume that for $\forall k \in \{1, ..., L+1\}$, p_k belongs to a fixed grid that excludes 0. It follows that :

$$H_0^{pre'} : F_{\lambda_{L+1}} \in \mathcal{F}_L \iff \left\{ \exists (k,j) \in \{1, ..., L+1\}^2, k \neq j, \ (\mu_k, \sigma_k) = (\mu_j, \sigma_j) \right\}$$

Therefore, $H_0^{pre'}$ is a union of null hypothesis, each of which can be tested using a separate Wald test. To maintain proper size control, we apply a Bonferroni correction.

B Proof of Propositions

B.1 Identification

Proof of Proposition 1.1.

We want to show that under Assumptions A, the following implication holds:

$$(\tilde{F},\tilde{\beta})=(F,\beta) \iff \mathbb{E}[m_3(W_i,\tilde{\theta})|Z_t]=0 \ a.s.$$
$$\iff \mathbb{E}\left[\rho_j^{-1}(Y_t,X_{2t},\tilde{F})-X_{1t}\tilde{\beta}\middle|Z_t\right]=0 \ a.s..$$

The direct implication is straightforward to show:

$$(\tilde{F},\tilde{\beta})=(F,\beta) \implies \forall j, \quad \mathbb{E}\left[\rho_j^{-1}(Y_t,X_{2t},\tilde{F})-X_{1t}\tilde{\beta}\middle|Z_t\right]=\mathbb{E}[\xi_{jt}|Z_t]=0 \ a.s.$$

The reverse implication is much more intricate to prove and we will exploit other results in the literature. We want to show:

$$(\tilde{F},\tilde{\beta})\neq(F,\beta) \implies \mathbb{E}\left[\rho^{-1}(Y_t,X_{2t},\tilde{F})-X_{1t}\tilde{\beta}\middle|Z_t\right]=0 \text{ a.s. does not hold.}$$

Case 1: First, let us assume that $\tilde{F}=F$ and $\tilde{\beta}\neq\beta$, then we have:

$$\rho^{-1}(Y_t,X_{2t},\tilde{F})-X_{1t}\tilde{\beta}=\underbrace{\rho^{-1}(Y_t,X_{2t},F)-X_{1t}\beta}_{\xi_t}+X_{1t}(\beta-\tilde{\beta})$$

By assumption, we have: $\mathbb{P}(X_{1t}^TX_{1t} \ \ pd)>0$.

Therefore, we have $\forall\ \gamma\in\mathbb{R}^{dim(\beta)}\backslash\{0\}$,

$$\mathbb{P}(X_{1t}\gamma\neq0)=P(\|\hat{X}_{1t}\gamma\|^2>0)=P(\gamma^TX_{1t}^TX_{1t}\gamma>0)>\Pr(X_{1t}^TX_{1t} \ \ pd)>0$$

Thus, $X_{1t}(\beta-\tilde{\beta})=0 \ a.s.$ does not hold. To conclude:

$\mathbb{E}[\rho^{-1}(Y_t,X_{2t},F)-X_{1t}\tilde{\beta}|Z_t]=\underbrace{\mathbb{E}[\xi_t|Z_t]}_{=0}+\mathbb{E}[X_{1t}(\beta-\tilde{\beta})|Z_t]=\ 0\ a.s$ does not hold from the completeness

Case 2: Now let us assume that $\tilde{F}\neq F$ and we want to show that $\forall\tilde{\beta}\in\mathbb{R}^{\dim(\beta)}$, $\exists j'$

such that:
$$\mathbb{E}\left[\rho_{j'}^{-1}(Y_t, X_{2t}, \tilde{F}) - X_{1j't}^T\tilde{\beta}\middle|Z_t\right] = 0 \ a.s. \text{ does not hold.}$$

First, let us observe that $\forall j'$,

$$\mathbb{E}\big[\rho_{j'}^{-1}(Y_t, X_{2t}, \tilde{F}) - X_{1j't}^T\tilde{\beta}\big|Z_t\big] = \underbrace{\mathbb{E}\big[\xi_{j't}\big|Z_t\big]}_{=0} + \mathbb{E}\big[\rho_{j'}^{-1}(Y_t, X_{2t}, \tilde{F}) - \rho_{j'}^{-1}(Y_t, X_{2t}, F) - X_{1j't}^T(\tilde{\beta} - \beta)\big|Z_t\big].$$

As a consequence, we need to show that $\exists j'$ such that $\mathbb{E}\Big[\rho_{j'}^{-1}(Y_t, X_{2t}, \tilde{F}) - \rho_{j'}^{-1}(Y_t, X_{2t}, F) - X'_{1j't}(\tilde{\beta} - \beta)\Big|Z_t\Big] = 0$ a.s. does not hold. From the completeness condition, a sufficient condition is: $\exists j'$ such that $\rho_{j'}^{-1}(Y_t, X_{2t}, \tilde{F}) - \rho_{j'}^{-1}(Y_t, X_{2t}, F) - X_{1j't}^T(\tilde{\beta} - \beta) = 0 \ a.s.$ does not hold.

- **Step 1:** First let us show that $\forall \gamma \in \mathbb{R}^{dim(\beta)}$, $\tilde{F} \neq F \implies \rho(\delta_t, X_{2t}, F) - \rho(\delta_t + X_{1t}\gamma, X_{2t}, F) = 0$ 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 $\mu_i = X_{1t}\Gamma + X_{2t}v_i = X_t\mathbf{v}$ 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 $G_{\mu|X_t}$ the distribution of $\mu_i|X_t$ under $F^\dagger = (c = 0, F)$ and $G_{\tilde{\mu}|X_t}$ the distribution of $\mu_i|X_t$ under $\tilde{F}^\dagger = (c = \gamma, \tilde{F})$. The following result is shown in Wang (2022): for any $\bar{X}_t \in Supp(X_t)$,

  $$\rho(\delta_t, G_{\mu|\bar{x}_t}) - \rho(\delta_t, G_{\tilde{\mu}|\bar{x}_t}) = 0 \text{ on open set } \mathcal{D} \subset \mathbb{R}^J \implies G_{\mu|\bar{X}_t} = G_{\tilde{\mu}|\bar{X}_t}.$$

  Thanks to the real analytic property of the demand functions ρ, Wang (2022) does not require a full support assumption on δ_t.

  By assumption, there exists $\bar{X}_t \in Supp(X_t)$ such that $\bar{X}_t'\bar{X}_t$ is positive definite and $\delta_t = \bar{X}_{1t}\beta + \xi_t$ varies on an open set $\bar{\mathcal{D}}$ almost surely. Given the result in Wang (2022), in order to prove that $\rho(\delta_t, X_{2t}, F) - \rho(\delta_t + X_{1t}\gamma, X_{2t}, F) = 0 \ a.s.$ does not hold, we just need to prove that $\forall \gamma$, $\tilde{F} \neq F \implies G_{\tilde{\mu}|\bar{X}_t} \neq G_{\mu|\bar{X}_t}$. By definition (see assumption A (iv)), $\tilde{F} \neq F \implies \exists v^* \in \mathbb{R}^{K_2}$ $\tilde{F}(v^*) \neq F(v^*)$. Take $X^* = (0_{K_1}, \bar{X}_{2t}v^*)^T = \bar{X}_t(0_{K_1}, v^*)^T$:

$$G_{\mu|\bar{X}_t}(X^*) = P(X_t\mathbf{v}_i \leq X^*|X_t = \bar{X}_t) = P((X_t^TX_t)^{-1}X_t^TX_t\mathbf{v}_i \leq (X_t^TX_t)^{-1}X_t^T\bar{X}_t(0_{K_1}, v^*)^T|X_t = \bar{X}_t)$$
$$= (1_{K_1}, P(v_i \leq v^*|X_t = \bar{x}_t))^T = (1_{K_1}, F(v^*))^T$$

The last equality comes from the independence of v_i and X_t. Likewise, $G_{\tilde{\mu}|\bar{X}_t}(X^*) = (1\{\gamma > 0\}, \tilde{F}(v^*))^T$

Therefore, $\exists X^*$, $\forall \gamma \quad G_{\tilde{\mu}|\bar{X}_t}(X^*) \neq G_{\mu|\bar{X}_t}(X^*)$. Following the result in Wang (2022), we have that for all γ, $\rho(\delta_t, X_{2t}, F) - \rho(\delta_t + X_{1t}\gamma, X_{2t}, \tilde{F}) = 0$ $a.s.$ does not hold.

- **Step 2:** let $\gamma = (\tilde{\beta} - \beta)$. We want to show that $\rho(\delta_t, X_{2t}, F) - \rho(\delta_t + X_{1t}\gamma, X_{2t}, \tilde{F}) = 0$ $a.s.$ does not hold $\implies \exists j'$ $\rho_{j'}^{-1}(Y_t, X_{2t}, \tilde{F}) = \rho_j^{-1}(Y_t, X_{2t}, F) + \gamma^T X_{1j't}$ $a.s.$ does not hold.

  We proceed by contradiction. Assume that $\rho(\delta_t, X_{2t}, F) - \rho(\delta_t + X_{1t}\gamma, X_{2t}, \tilde{F}) = 0$ $a.s.$ does not hold and $\forall j'$ $\rho_{j'}^{-1}(Y_t, X_{2t}, \tilde{F}) = \rho_{j'}^{-1}(Y_t, X_{2t}, F) + \gamma^T X_{1j't} = 0$ $a.s.$. Then, we have: $\rho(\rho^{-1}(Y_t, X_{2t}, \tilde{F}), X_{2t}, \tilde{F}) = \rho(\rho^{-1}(Y_t, X_{2t}, F) + X_{1t}\gamma, X_{2t}, \tilde{F}) = \rho(\delta_t + X_{1t}\gamma, X_{2t}, \tilde{F}) = \rho(\delta_t, X_{2t}, F) = Y_t$ $a.s.$ does not hold. Therefore, we have a contradiction.

  This ends the proof

□

B.2 The infeasible most powerful instrument (MPI)

B.2.1 Simple hypothesis

Proof of Proposition 3.1.

The proof follows from classical arguments.

- Under $\overline{H}_0 : (\beta, F) = (\beta_0, F_0)$. By assumption, $\{W_i\}_{i=1,\ldots,N}$ are i.i.d. , $\mathbb{E}[\|h(Z_i)m(W_i, \theta_0)\|_2^2] < C$, the multivariate CLT applies:

$$\frac{1}{\sqrt{N}}\sum_{i=1}^{N} h(Z_i)m(W_i, \theta_0) \xrightarrow[N\to+\infty]{d} \mathcal{N}(0, \Omega_0),$$

with:

$$\Omega_0 = \mathbb{E}\left[\left(h(Z_i)m(W_i,\theta_0)\right)\left(h(Z_i)m(W_i,\theta_0)\right)^T\right].$$

Let us note that the assumption $\mathbb{E}[\|h(Z_i)m(W_i,\theta_0)\|^2] < C$ ensures that Ω_0 is finite. To see this, consider the following:

$$\mathbb{E}[\|\left(h(Z_i)m(W_i,\theta_0)\right)\left(h(Z_i)m(W_i,\theta_0)\right)^T\|_F] \leq \mathbb{E}[\|\left(h(Z_i)m(W_i,\theta_0)\right)\|_2^2] < C.$$

The inequality follows from the fact that $\|AB\|_F \leq \|A\|_F\|B\|_F$ for any two compatible matrices A and B. Thus, the law of large numbers applies: $\hat{\Omega}_0 \xrightarrow[N\to+\infty]{P} \Omega_0$. Additionally, by assumption, Ω_0 is full rank. Therefore, by the Continuous Mapping Theorem (CMT), we conclude that:

$$S_N(h,\theta_0) = N\left(\frac{1}{N}\sum_{i=1}^{N} h(Z_i)m(W_i,\theta_0)\right)^T \hat{\Omega}_0^{-1}\left(\frac{1}{N}\sum_{i=1}^{N} h(Z_i)m(W_i,\theta_0)\right) \xrightarrow[N\to+\infty]{d} \chi_L^2.$$

- Under H_a' : $\mathbb{E}\left[h(Z_i)m(W_i,\theta_0)\right] \neq 0$. The data are i.i.d., by the law of large numbers: $\frac{1}{N}\sum_{i=1}^{N} h(Z_i)m(W_i,\theta_0) \xrightarrow{\mathbb{P}} \mathbb{E}\left[h(Z_i)m(W_i,\theta_0)\right]$. It follows by the continuous mapping theorem:

$$\frac{S_N(h,\theta_0)}{N} \xrightarrow{\mathbb{P}} \underbrace{\mathbb{E}\left[h(Z_i)m(W_i,\theta_0)\right]^T \Omega_0^{-1}\mathbb{E}\left[h(Z_i)m(W_i,\theta_0)\right]}_{\kappa(h,\theta_0)}$$

Under H_a', $\kappa(h,\theta_0)$ is strictly positive because Ω_0 is positive definite. Thence,

$$\begin{aligned}
\forall q \in \mathbb{R}, \quad \underset{N\to\infty}{lim}\ \mathbb{P}(S_N(h,\theta_0) > q) &= \underset{N\to\infty}{lim}\ \mathbb{P}\left(\frac{S_N(h,\theta_0) - q}{N} > 0\right) \\
&= \mathbb{P}(\kappa(h,\theta_0) > 0) \\
&= 1,
\end{aligned}$$

where the second equality holds because convergence in probability implies convergence in distribution.

$\square$

Proof of Proposition 3.2.

To shorten notations, for any $\tilde{\theta}$, let $m(\tilde{\theta}) \equiv m(W_i, \tilde{\theta})$. Under $\overline{H}_a : (\beta, F) = (\beta_a, F_a)$, the asymptotic slope of the test is written as:

$$
\begin{aligned}
c(h, \theta_0) =& \mathbb{E}\left[h(Z_i)m(W_i, \theta_0)\right]^T \Omega_0^{-1} \mathbb{E}\left[h(Z_i)m(W_i, \theta_0)\right] \\
=& \mathbb{E}\left[h(Z_i)m(W_i, \theta_0)\right]^T \mathbb{E}\left[\big(h(Z_i)m(W_i, \theta_0)\big)\big(h(Z_i)m(W_i, \theta_0)\big)^T\right]^{-1} \mathbb{E}\left[h(Z_i)m(W_i, \theta_0)\right], \\
=& \mathbb{E}\left[h(Z_i)(m(\theta_0) - m(\theta_a))\right]^T \mathbb{E}\left[\big(h(Z_i)m(\theta_0)\big)\big(h(Z_i)m(\theta_0)\big)^T\right]^{-1} \mathbb{E}\left[h(Z_i)(m(\theta_0) - m(\theta_a))\right] \\
=& \mathbb{E}\left[h(Z_i)\Delta^m_{0,a}\right]^T \mathbb{E}\left[\big(h(Z_i)m(\theta_0)\big)\big(h(Z_i)m(\theta_0)\big)^T\right]^{-1} \mathbb{E}\left[h(Z_i)\Delta^m_{0,a}\right].
\end{aligned}
$$

The third line comes from the fact that under $\overline{H}_a : (\beta, F) = (\beta_a, F_a)$, $\mathbb{E}\left[h(Z_i)(m(\theta_0) - m(\theta_a)\right] = \mathbb{E}[h(Z_i)m(\theta_0)]$. Then the slope of the test taking $h^*_a(Z_i) = \big(\mathbb{E}(m(\theta_0)m(\theta_0)^T|Z_i)^{-1}\mathbb{E}(\Delta^m_{0,a}|Z_i)\big)^T$ is equal to:

$$
c(h^*_a, \theta_a) = \mathbb{E}\left(\mathbb{E}(\Delta^m_{0,a}|Z_i)^T \mathbb{E}(m(\theta_0)m(\theta_0)^T|Z_i)^{-1}\mathbb{E}(\Delta^m_{0,a}|Z_i)\right).
$$

To finish the proof, we must show that for any set of instruments h, we have: $c(h^*_a, \theta_a) \geq c(h, \theta_a)$.

Denote $\tilde{h}(Z_i) = \mathbb{E}(m(\theta_0)m(\theta_0)^T|Z_i)^{1/2}h(Z_i)^T$ and $\tilde{h}^*_a(Z_i) = \mathbb{E}(m(\theta_0)m(\theta_0)^T|Z_i)^{1/2}h^*(Z_i)^T$. With these new notations, we have:

$$
\begin{aligned}
c(h^*_a, \theta_0) - c(h, \theta_0) &= \mathbb{E}\left(\tilde{h}^*(Z_i)^T\tilde{h}^*_a(Z_i)\right) - \mathbb{E}\left(\tilde{h}^*(Z_i)^T\tilde{h}(Z_i)\right)\mathbb{E}\left(\tilde{h}(Z_i)^T\tilde{h}(Z_i)\right)^{-1}\mathbb{E}\left(\tilde{h}(Z_i)^T h^*_a(Z_i)\right) \\
&= G^T \begin{pmatrix} \mathbb{E}\left(\tilde{h}^*_a(Z_i)^T\tilde{h}^*_a(Z_i)\right) & \mathbb{E}\left(\tilde{h}^*_a(Z_i)^T\tilde{h}(Z_i)\right) \\ \mathbb{E}\left(\tilde{h}(Z_i)^T\tilde{h}^*_a(Z_i)\right) & \mathbb{E}\left(\tilde{h}(Z_i)^T\tilde{h}(Z_i)\right) \end{pmatrix} G \\
&= G^T\mathbb{E}\left(\tilde{H}\tilde{H}^T\right) G \geq 0,
\end{aligned}
$$

with $\tilde{H} = (\tilde{h}^*_a(Z_i), \tilde{h}(Z_i))^T$ and $G = \left(1, -\mathbb{E}\left(\tilde{h}^*_a(Z_i)^T\tilde{h}(Z_i)\right)\mathbb{E}\left(\tilde{h}(Z_i)^T\tilde{h}(Z_i)\right)^{-1}\right)^T$.

□

Proof of Proposition 3.3.

Assuming that m identifies θ, we have the following implication:

$$
\begin{aligned}
\overline{H}_a : (\beta, F) = (\beta_a, F_a) \neq (\beta_0, F_0) \implies \quad & \mathbb{E}[m(W_i,\theta_0)|Z_i] = 0 \;\text{ does not hold } a.s. \\
\implies \quad & \mathbb{E}\big[m(W_i,\theta_0)|Z_i\big]^T \mathbb{E}[m(W_i,\theta_0)m(W_i,\theta_0)^T|Z_i]^{-1}\mathbb{E}[m(W_i,\theta_0)|Z_i\big] > 0 \; a.s. \\
\implies \quad & \mathbb{E}\Big(\mathbb{E}\big[m(W_i,\theta_0)|Z_i\big]^T \mathbb{E}[m(W_i,\theta_0)m(W_i,\theta_0)^T|Z_i]^{-1}\mathbb{E}[m(W_i,\theta_0)|Z_i]\Big) > 0 \\
\implies \quad & \mathbb{E}\Big(\mathbb{E}\Big[\mathbb{E}\big[\Delta^m_{0,a}|Z_i\big]^T \mathbb{E}[m(W_i,\theta_0)m(W_i,\theta_0)^T|Z_i]^{-1} m(W_i,\theta_0)\Big| Z_i\Big]\Big) > 0 \\
\implies \quad & \mathbb{E}\Big(\underbrace{\mathbb{E}\big[\Delta^m_{0,a}|Z_i\big]^T \mathbb{E}[m(W_i,\theta_0)m(W_i,\theta_0)^T|Z_i]^{-1}}_{h^*_a(Z_i)} m(W_i,\theta_0)\Big) > 0 \\
\implies \quad & \overline{H}'_a : \; \mathbb{E}\big[h^*_a(Z_i)m(W_i,\theta_0)\big] \neq 0.
\end{aligned}
$$

The second line comes from the fact that $\mathbb{E}[m(W_i,\theta_0)m(W_i,\theta_0)^T|Z_i]$ is assumed positive definite. The fifth line comes from the law of iterated expectations. Under the same assumptions as 3.1, we have the following:

$$
\overline{H}'_a : \; \mathbb{E}\big[h^*_a(Z_i)m(W_i,\theta)\big] \neq 0 \implies \forall q \in \mathbb{R}^+, \;\; \mathbb{P}(S_N(h^*,\theta_0) > q) \underset{N\to+\infty}{\longrightarrow} 1.
$$

□

B.2.2 Composite hypothesis

Proof of Proposition 3.4.

We assume that $h \in \mathcal{H}_0^\perp$, $\hat{\phi}$ satisfies Assumption 2 and m satisfies Assumption 3.

(i) **First, we examine the case** $H_0 : F \in \mathcal{F}_0$. We consider a mean value expansion of $\frac{1}{\sqrt{N}}\sum_{i=1}^N h(Z_i)m(W_i,\hat{\phi})$ around ϕ_0. This expansion exists as $\phi \mapsto m(W_i,\hat{\phi})$ is continuously differentiable on Φ from Assumption 3 (1). Moreover, Assumption 2 (1) implies that $\phi_0 \in \text{interior}(\Phi)$ and thus, with probability that goes to one, $\hat{\phi} \in \Phi$.

$$
\frac{1}{\sqrt{N}}\sum_{i=1}^N h(Z_i)m(W_i,\hat{\phi}) = \underbrace{\frac{1}{\sqrt{N}}\sum_{i=1}^N h(Z_i)m(W_i,\phi_0)}_{A_N} + \underbrace{\frac{1}{\sqrt{N}}\sum_{i=1}^N h(Z_i)\frac{\partial m(W_i,\tilde{\phi})}{\partial \phi^T}(\hat{\phi}-\phi_0)}_{B_N}
$$

where $\tilde{\phi} \in conv(\phi_0, \hat{\phi})$.

– First, we show that $B_N = o_p(1)$. From Assumption 2, $\sqrt{N}(\hat{\phi} - \phi_0) = O_p(1)$. Next, we verify that the conditions to apply the Uniform Law of Large Numbers as stated in Theorem 18.2 in the Econometrics textbook Hansen (2022) are satisfied.

  1. $V_i \equiv \{W_i, Z_i\}$ are i.i.d.
  2. Φ is compact
  3. Bounded envelope. With probability one and $\forall \phi \in \Phi$, $g(V_i, \phi) \equiv h(Z_i)\frac{\partial m(W_i,\phi)}{\partial \phi^T}$ has the following envelope:

$$\|g(V_i, \phi)\|_F \leq G(V_i) \equiv \tilde{M}(W_i)\sqrt{dim(\phi)}\sqrt{\sum_{l=1}^{L}\|h_l(Z_i)\|_2^2}$$

     where $\|\cdot\|_F$ is the Frobenius norm. This can be established as follows:

$$\|g(V_i, \phi)\|_F = \sqrt{\sum_{l=1}^{L}\sum_{j=1}^{\dim(\phi)}\left|h_l(Z_i)^T\frac{\partial m(W_i, \phi)}{\partial \phi_j}\right|^2} \leq \sqrt{\sum_{l=1}^{L}\sum_{j=1}^{\dim(\phi)}\|h_l(Z_i)\|_2^2\tilde{M}(W_i)^2}$$
$$= \tilde{M}(W_i)\sqrt{dim(\phi)}\sqrt{\sum_{l=1}^{L}\|h_l(Z_i)\|_2^2}.$$

     The first inequality comes from Assumption 3 (3). The last inequality comes from the triangular inequality. Moreover, from Assumptions 3 (3) and condition (2) in the definition of $\mathcal{H}_0^{\perp}$, along with the application of the Cauchy-Schwarz inequality, we obtain:

$$\mathbb{E}\left[G(V_i)\right] \leq \sqrt{dim(\phi)}\mathbb{E}[\tilde{M}(W_i)^2]^{1/2}\mathbb{E}\left[\sum_{l=1}^{L}\|h_l(Z_i)\|_2^2\right]^{1/2} \leq \sqrt{dim(\phi)}\overline{\tilde{M}}^{1/2}L^{1/2}C'^{1/4}.$$

  4. From Assumption 3, $g(V_i, \phi)$ is continuous in ϕ with probability one.

  It follows that the Uniform Law of Large Numbers applies:

$$\sup_{\phi \in \Phi}\left\|\frac{1}{N}\sum_{i=1}^{N}g(V_i, \phi) - \mathbb{E}\left[g(V_i, \phi)\right]\right\|_F \xrightarrow[N \to +\infty]{P} 0. \tag{B.10}$$

Next, the triangular inequality entails:

$$\left\|\frac{1}{N}\sum_{i=1}^{N} h(Z_i)\frac{\partial m(W_i,\tilde{\phi})}{\partial \phi^T} - \mathbb{E}\left[h(Z_i)\frac{\partial m(W_i,\phi_0)}{\partial \phi^T}\right]\right\| \leq \sup_{\phi\in\Phi}\left\|\frac{1}{N}\sum_{i=1}^{N} g(V_i,\phi) - \mathbb{E}\left[g(V_i,\phi)\right]\right\|_F$$
$$+\left\|\mathbb{E}\left[g(V_i,\hat{\phi})\right] - \mathbb{E}\left[g(V_i,\phi_0)\right]\right\|_F.$$

The first term in the upper bound is $o_p(1)$ because of (B.10). The second term is $o_p(1)$ because $\hat{\phi} \xrightarrow[N\to+\infty]{P} \phi_0$ and $\phi \mapsto \mathbb{E}\left[g(V_i,\cdot)\right]$ is continuous on Φ. The continuity of $\phi \mapsto \mathbb{E}\left[g(V_i,\cdot)\right]$ follows directly from Lemma 2.4 in Newey and McFadden (1994)) which applies under the following conditions: (i) continuity of g ensured by Assumption 3 and (ii) g has a bounded envelope as shown previously.

Thus, $\frac{1}{N}\sum_{i=1}^{N} h(Z_i)\frac{\partial m(W_i,\tilde{\phi})}{\partial \phi^T} = \mathbb{E}\left[h(Z_i)\frac{\partial m(W_i,\phi_0)}{\partial \phi^T}\right] + o_p(1) = o_p(1)$ where the last equality comes from the fact that $h \in \mathcal{H}_0^{\perp}$. As a consequence,

$$B_N = \frac{1}{N}\sum_{i=1}^{N} h(Z_i)\frac{\partial m(W_i,\tilde{\phi})}{\partial \phi^T}\sqrt{N}(\hat{\phi}-\phi_0) = o_p(1)O_p(1) = o_p(1)$$

– Second, following the same classical steps as in the proof of Proposition 3.1, we can show that $A_N \xrightarrow[N\to+\infty]{d} \mathcal{N}(0,\Omega_0)$ and thus, $S_N(h,\hat{\phi}) \xrightarrow[N\to+\infty]{d} \chi^2_L$. This result holds provided that we have a consistent estimator of Ω_0. We show in (iii) that the empirical counterpart of Ω_0 obtained by replacing $\hat{\phi}$ with ϕ_0 is a consistent estimator.

(ii) **Second, we examine the case** H_a' : $\mathbb{E}\left[h(Z_i)m(W_i,\phi_0)\right] \neq 0$. Directly, from case (i), we have:

$$\frac{1}{N}\sum_{i=1}^{N} h(Z_i)m(W_i,\hat{\phi}) = \frac{1}{N}\sum_{i=1}^{N} h(Z_i)m(W_i,\phi_0) + o_p\left(\frac{1}{\sqrt{N}}\right).$$

The data are i.i.d. , $\mathbb{E}\left[\|h(Z_i)m(W_i,\phi_0)\|_2\right] < \sqrt{C}$ (as $h \in \mathcal{H}_0$), therefore, we can apply the law of large numbers:

$$\frac{1}{N}\sum_{i=1}^{N} h(Z_i)m(W_i,\hat{\phi}) \xrightarrow[N\to+\infty]{P} \mathbb{E}\left[h(Z_i)m(W_i,\phi_0)\right].$$

Finally, following the same classical steps as in the proof of Proposition 3.1 and provided that we have a consistent estimator of Ω_0, we can show the second statement in Proposition 3.4.

(iii) **Third, we show that** $\hat{\Omega} \xrightarrow[N\to+\infty]{P} \Omega_0$. First, let us show:

$$\frac{1}{N}\sum_{i=1}^{N}\left(h(Z_i)m(W_i,\hat{\phi})\right)\left(h(Z_i)m(W_i,\hat{\phi})\right)^T = \frac{1}{N}\sum_{i=1}^{N}\left(h(Z_i)m(W_i,\phi_0)\right)\left(h(Z_i)m(W_i,\phi_0)\right)^T + o_p(1)$$

Hence, we need to show that a ULLN holds for: $g^{\dagger\dagger}(V_i,\phi) \equiv \left(h(Z_i)m(W_i,\phi)\right)\left(h(Z_i)m(W_i,\phi)\right)^T$ and therefore, one can replace ϕ_0 by $\hat{\phi}$ without affecting the limit in probability.

Again, the conditions to apply the Uniform Law of Large Numbers as stated in Theorem 18.2 in Hansen (2022) are satisfied.

1. $\{V_i\} \equiv \{W_i, Z_i\}$ are i.i.d.
2. Φ is compact
3. Bounded envelope. With probability one, $g^{\dagger}(V_i,\phi) \equiv h(Z_i)m(W_i,\phi_0)$ has the following envelope: $\forall \phi \in \Phi$, $\|g^{\dagger}(V_i,\phi)\|_2 \leq G^{\dagger}(V_i) \equiv M(W_i)\sum_{l=1}^{L}\|h_l(Z_i)\|_2$.

$$\begin{aligned}
\|g^{\dagger}(V_i,\phi)\|_2 &= \sqrt{\sum_{l=1}^{L}|h_l(Z_i)^T m(W_i,\phi)|^2} \leq \sqrt{\sum_{l=1}^{L}\|h_l(Z_i)\|_2^2 M(W_i)^2} \\
&= M(W_i)\sqrt{\sum_{l=1}^{L}\|h_l(Z_i)\|_2^2}.
\end{aligned}$$

   The first inequality comes from Assumption 3 (2). The last inequality comes from the triangular inequality. It follows, using the fact that $\|AB\|_F \leq \|A\|_F\|B\|_F$ for any two compatible matrices A and B, that $\forall \phi \in \Phi$:

$$\|g^{\dagger\dagger}(V_i,\phi)\|_F \leq \|g^{\dagger}(V_i,\phi)\|_2^2 \leq G^{\dagger}(V_i)^2.$$

   Thus, $g^{\dagger\dagger}(V_i,\phi)$ is bounded by the envelope $G^{\dagger}(V_i)^2$. Moreover, from Assumptions 3 (2), condition (2) in the definition of $\mathcal{H}_0^{\perp}$, along with the application of the Cauchy-Schwarz inequality, we have: $\mathbb{E}\left[G^{\dagger}(V_i)^2\right] \leq \overline{M}^{1/2}LC'^{1/2}$.
4. From Assumption 3, $g^{\dagger\dagger}(V_i,\phi)$ is continuous in ϕ with probability one.

It follows that the Uniform Law of Large Numbers applies:

$$\sup_{\phi\in\Phi}\left|\left|\sum_{i=1}^{N} g^{\dagger\dagger}(V_i,\phi)-\mathbb{E}\left[g^{\dagger\dagger}(V_i,\phi)\right]\right|\right|_2 \xrightarrow[N\to+\infty]{P} 0.$$

From the previous statement, the triangular inequality, and the continuous mapping theorem, one can easily show B.2.2 by following the same steps as in (i). Finally, Ω_0 is finite by Assumption 1 (1). Consequently, the law of large numbers applies, leading to the desired result.

□

Proof of Proposition 3.5.

Let $h\in\mathcal{H}_0$. First, let us prove that under Assumption 4 and Assumption 3, $h_\Lambda^\perp$ exists. By definition (when it exists):

$$\begin{aligned}h_\Lambda^\perp(Z_i) &= \left[h(Z_i)\Lambda(Z_i)-\mathbb{E}\left[BLP\left(h(Z_i)\Lambda(Z_i)\middle|\Lambda(Z_i)\frac{\partial m(W_i,\phi_0)}{\partial\phi^T}\right)\middle|Z_i\right]\right]\Lambda(Z_i)\\ &= h(Z_i)\Lambda(Z_i)^2-\Gamma_h\mathbb{E}\left[\frac{\partial m(W_i,\phi_0)}{\partial\phi}\middle|Z_i\right]\Lambda(Z_i)^2\end{aligned}$$

with $\Gamma_h=\mathbb{E}\left[h(Z_i)\Lambda(Z_i)^2\frac{\partial m(W_i,\phi_0)}{\partial\phi^T}\right]\mathbb{E}\left[\frac{\partial m(W_i,\phi_0)}{\partial\phi}\Lambda(Z_i)^2\frac{\partial m(W_i,\phi_0)}{\partial\phi^T}\right]^{-1}$

From Assumption 3, $\frac{\partial m(W_i,\phi_0)}{\partial\phi^T}$ exists and has a finite first moment with probability one. Therefore, $\mathbb{E}\left[\frac{\partial m(W_i,\phi_0)}{\partial\phi^T}\middle|Z_i\right]$ is well-defined. Second, we want to show that Γ_h is well-defined.

- From Assumption 4, $\mathbb{E}\left[\frac{\partial m(W_i,\phi_0)}{\partial\phi}\Lambda(Z_i)^2\frac{\partial m(W_i,\phi_0)}{\partial\phi^T}\right]$ is finite:

$$\mathbb{E}\left[\left|\left|\frac{\partial m(W_i,\phi_0)}{\partial\phi}\Lambda(Z_i)^2\frac{\partial m(W_i,\phi_0)}{\partial\phi^T}\right|\right|_F\right]\le\mathbb{E}\left[\left|\left|\frac{\partial m(W_i,\phi_0)}{\partial\phi}\Lambda(Z_i)\right|\right|_F^2\right]<C''$$

- From Assumption 4, $\mathbb{E}\left[h(Z_i)\Lambda(Z_i)^2\frac{\partial m(W_i,\phi_0)}{\partial\phi^T}\right]$ is finite:

$$\begin{aligned}\mathbb{E}\left[\left\|h(Z_i)\Lambda(Z_i)^2\frac{\partial m(W_i,\phi_0)}{\partial\phi^T}\right\|_F\right] &\leq \mathbb{E}\left[\|h(Z_i)\Lambda(Z_i)\|_F\left\|\Lambda(Z_i)\frac{\partial m(W_i,\phi_0)}{\partial\phi^T}\right\|_F\right]\\ &\leq\sqrt{\mathbb{E}\left[\|h(Z_i)\Lambda(Z_i)\|_F^2\right]\mathbb{E}\left[\left\|\Lambda(Z_i)\frac{\partial m(W_i,\phi_0)}{\partial\phi^T}\right\|_F^2\right]}<C''\end{aligned}$$

- From Assumption 4, $\mathbb{E}\left[h(Z_i)\Lambda(Z_i)^2\frac{\partial m(W_i,\phi_0)}{\partial\phi^T}\right]$ is full rank.

The previous computations rely on the submultiplicativity of the Frobenius norm and the Cauchy-Schwarz inequality. It follows that Γ_h exists.

Second, we prove that the moment $\mathbb{E}\left[h^{\perp}(Z_i)m(W_i,\phi_0)\right]$ is Neyman orthogonal. The proof leverages properties of the best linear predictor and the conditional expectation operators.

$$\begin{aligned}\mathbb{E}\left[h^{\perp}(Z_i)\frac{\partial m(W_i,\phi_0)}{\partial\phi^T}\right] &= \mathbb{E}\left[h(Z_i)\Lambda(Z_i)^2\frac{\partial m(W_i,\phi_0)}{\partial\phi^T}\right]\\ &-\mathbb{E}\left[\mathbb{E}\left[BLP\left(h(Z_i)\Lambda(Z_i)\middle|\Lambda(Z_i)\frac{\partial m(W_i,\phi_0)}{\partial\phi^T}\right)\middle|Z_i\right]\Lambda(Z_i)\frac{\partial m(W_i,\phi_0)}{\partial\phi^T}\right]\\ &= \mathbb{E}\left[h(Z_i)\Lambda(Z_i)^2\frac{\partial m(W_i,\phi_0)}{\partial\phi^T}\right]\\ &-\mathbb{E}\left[BLP\left(h(Z_i)\Lambda(Z_i)\middle|\Lambda(Z_i)\frac{\partial m(W_i,\phi_0)}{\partial\phi^T}\right)\Lambda(Z_i)\frac{\partial m(W_i,\phi_0)}{\partial\phi^T}\right]\\ &= \mathbb{E}\left[h(Z_i)\Lambda(Z_i)^2\frac{\partial m(W_i,\phi_0)}{\partial\phi^T}\right]-\mathbb{E}\left[h(Z_i)\Lambda(Z_i)^2\frac{\partial m(W_i,\phi_0)}{\partial\phi^T}\right]=0.\end{aligned}$$

The last line comes from the fact that for any $(A_i,B_i)\in\mathcal{M}_{k,p}(\mathbb{R})\times\mathcal{M}_{p,m}(\mathbb{R})$ with $(k,m)\in\mathbb{N}^{*2}$, $\mathbb{E}\left[A_iB_i\right]=\mathbb{E}\left[BLP(A_i|B_i)B_i\right]$.

□

Proof of Proposition 1.2.

We assume that $h\in\mathcal{H}_0$ and satisfies Assumption 2 (1). Furthermore, we assume that Assumption 4 is satisfied, $\hat{\phi}$ satisfies Assumption 2, m satisfies Assumption 5, Λ satisfies Assumption 6. To shorten the notations, we will sometimes write $m_i(\phi)\equiv m(W_i,\phi)$.

First, we want to show that under H_0:

$$\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\hat{h}_{\Lambda}^{\perp}(Z_i)m(W_i,\hat{\phi})=\frac{1}{\sqrt{N}}\sum_{i=1}^{N}h_{\Lambda}^{\perp}(Z_i)m(W_i,\phi_0)+o_p(1)$$

with $h^{\perp}$ and $\hat{h}^{\perp}$ respectively the population and the sample orthogonalized counterparts of h (as in 3.6). By definition of $\hat{h}^{\perp}$ and $h^{\perp}$, we have:

$$\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\hat{h}_{\Lambda}^{\perp}(Z_i)m(W_i,\hat{\phi})=\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\left(h(Z_i)\Lambda(Z_i)^2-\hat{\Gamma}_h\mathbb{E}\left[\frac{\partial m(W_i,\hat{\phi})}{\partial\phi}\middle|Z_i\right]\Lambda(Z_i)^2\right)m(W_i,\hat{\phi})$$

with:

$$\hat{\Gamma}_h=\left(\frac{1}{N}\sum_{i=1}^{N}h(Z_i)\Lambda(Z_i)^2\frac{\partial m_i(\hat{\phi})}{\partial\phi^T}\right)\left(\frac{1}{N}\sum_{i=1}^{N}\frac{\partial m_i(\hat{\phi})}{\partial\phi}\Lambda(Z_i)^2\frac{\partial m_i(\hat{\phi})}{\partial\phi^T}\right)^{-1}$$

and its population counterpart:

$$\Gamma_h=\mathbb{E}\left[h(Z_i)\Lambda(Z_i)^2\frac{\partial m_i(\phi_0)}{\partial\phi^T}\right]\mathbb{E}\left[\frac{\partial m_i(\phi_0)}{\partial\phi}\Lambda(Z_i)^2\frac{\partial m_i(\phi_0)}{\partial\phi^T}\right]^{-1}.$$

From Proposition 1.2, we know that under Assumptions 5 and 4, Γ_h exists. We proceed in two steps.

(i) **First step:** we show that $\hat{\Gamma}_h \xrightarrow[N\to+\infty]{P} \Gamma_h$. To show this statement, we proceed in three steps.

- $\frac{1}{N}\sum_{i=1}^{N}h(Z_i)\Lambda(Z_i)^2\frac{\partial m_i(\hat{\phi})}{\partial\phi^T} \xrightarrow[N\to+\infty]{P} \mathbb{E}\left[h(Z_i)\Lambda(Z_i)^2\frac{\partial m_i(\phi_0)}{\partial\phi^T}\right]$ $(*)$
  To show this statement, we verify that the conditions to apply the Uniform Law of Large Numbers as stated in Theorem 18.2 in Hansen (2022) are satisfied.

  1. $\{V_i\}=\{W_i,Z_i\}$ are i.i.d.
  2. Φ is compact
  3. With probability one, $g(V_i,\phi)\equiv h(Z_i)\Lambda(Z_i)^2\frac{\partial m(W_i,\phi)}{\partial\phi^T}$ has the following envelope: $\|g(V_i,\phi)\|_F\le G(V_i)\equiv\sqrt{\dim(\phi)}\|h(Z_i)\Lambda(Z_i)\|_F\tilde{M}(W_i)\|\Lambda(Z_i)\|_F$ where $\|\cdot\|_F$ is the Frobenius norm. Indeed,

$$
\begin{aligned}
\|g(V_i,\phi)\|_F &\leq \|h(Z_i)\Lambda(Z_i)\|_F \left\|\Lambda(Z_i)\frac{\partial m(W_i,\phi)}{\partial \phi^T}\right\|_F \\
&= \|h(Z_i)\Lambda(Z_i)\|_F \sqrt{\sum_{j=1}^{p}\sum_{l=1}^{\dim(\phi)}\left|(\Lambda(Z_i)_j)^T\frac{\partial m(W_i,\phi)}{\partial \phi_l}\right|^2} \\
&\leq \|h(Z_i)\Lambda(Z_i)\|_F \sqrt{\dim(\phi)}\tilde{M}(W_i)\sqrt{\sum_{j=1}^{p}\|\Lambda(Z_i)_j\|_2^2} \\
&\leq \sqrt{\dim(\phi)}\|h(Z_i)\Lambda(Z_i)\|_F \tilde{M}(W_i)\|\Lambda(Z_i)\|_F.
\end{aligned}
$$

The first inequality stems from the under-multiplicativity of the Frobenius norm. Moreover, from Assumptions 4, 5, 6, along with the application of the Cauchy-Schwarz inequality, we obtain:

$$
\begin{aligned}
\mathbb{E}\left[G(V_i)\right] &\leq \sqrt{dim(\phi)}\mathbb{E}\left[\|h(Z_i)\Lambda(Z_i)\|_F^2\right]^{1/2}\mathbb{E}\left[\tilde{M}(W_i)^4\right]^{1/4}\mathbb{E}\left[\|\Lambda(Z_i)\|_F^4\right]^{1/4} \\
&< \sqrt{dim(\phi)}C''^{1/2}\overline{\tilde{M}}^{1/4}\overline{\Lambda}^{1/4}.
\end{aligned}
$$

4. From Assumption 5, $g(V_i,\phi)$ is continuous in ϕ with probability one.

It follows that the Uniform Law of Large Numbers applies:

$$
\sup_{\phi\in\Phi}\left\|\frac{1}{N}\sum_{i=1}^{N} g(V_i,\phi) - \mathbb{E}\left[g(V_i,\phi)\right]\right\|_F \xrightarrow[N\to+\infty]{P} 0. \tag{B.11}
$$

Next, it follows from the triangular inequality:

$$
\begin{aligned}
&\left\|\frac{1}{N}\sum_{i=1}^{N} h(Z_i)\Lambda(Z_i)^2\frac{\partial m_i(\hat{\phi})}{\partial\tilde{\phi}^T} - \mathbb{E}\left[h(Z_i)\Lambda(Z_i)^2\frac{\partial m_i(\phi_0)}{\partial\tilde{\phi}^T}\right]\right\| \leq \sup_{\phi\in\Phi}\left\|\frac{1}{N}\sum_{i=1}^{N} g(V_i,\phi) - \mathbb{E}\left[g(V_i,\phi)\right]\right\|_F \\
&+\left\|\mathbb{E}\left[g(V_i,\hat{\phi})\right] - \mathbb{E}\left[g(V_i,\phi_0)\right]\right\|_F.
\end{aligned}
$$

Where the first term in the upper bound is $o_p(1)$ because of (B.11) and the second term is $o_p(1)$ because $\hat{\phi} \xrightarrow[N\to+\infty]{P} \phi_0$ and $\phi \mapsto \mathbb{E}\left[g(V_i,\cdot)\right]$. The continuity of $\phi \mapsto \mathbb{E}\left[g(V_i,\cdot)\right]$ follows directly from Lemma 2.4 in Newey and McFadden (1994)) which applies under the following conditions: (i) continuity of g ensured by Assumption 5 and (ii) g has a bounded envelope as shown previously.

- $\frac{1}{N}\sum_{i=1}^N \frac{\partial m_i(\hat{\phi})}{\partial \phi}\Lambda(Z_i)^2 \frac{\partial m_i(\hat{\phi})}{\partial \phi^T} \xrightarrow[N\to+\infty]{P} \mathbb{E}\left[\frac{\partial m_i(\phi_0)}{\partial \phi}\Lambda(Z_i)^2 \frac{\partial m_i(\phi_0)}{\partial \phi^T}\right]$ $(**)$.

  As for $(*)$, the main difficulty is to show the uniform convergence of $g^\dagger(V_i,\phi) \equiv \frac{\partial m_i(\phi)}{\partial \phi}\Lambda(Z_i)^2 \frac{\partial m_i(\phi)}{\partial \phi^T}$. To do so, we use the same arguments as previously, we just need to find an envelope on $g^\dagger$. Specifically, we can show that $\|g^\dagger(V_i,\phi)\|_F \leq G^\dagger(V_i) \equiv p\ dim(\phi)\tilde{M}(W_i)^2 \|\Lambda(Z_i)\|_F^4$:

  $$\begin{aligned}\|g^\dagger(V_i,\phi)\|_F &\leq \left\|\frac{\partial m_i(\phi)}{\partial \phi}\Lambda(Z_i)\right\|_F^2 \leq \left\|\frac{\partial m_i(\phi)}{\partial \phi}\right\|_F^2 \|\Lambda(Z_i)\|_F^2 \\ &\leq p\ dim(\phi)\tilde{M}(W_i)^2 \|\Lambda(Z_i)\|_F^2 .\end{aligned}$$

  The first two inequalities stem from the under-multiplicativity of the Frobenius norm. The last inequality comes from Assumption 5 (3). Moreover, from Assumption 5 (3) and Asumption 6, along with the application of the Cauchy-Schwarz inequality, we have: $\mathbb{E}\left[G^\dagger(V_i)\right] \leq p\, dim(\phi)\tilde{\overline{M}}^{1/2}\sqrt{\overline{\Lambda}}$. Then, the uniform convergence of $g^\dagger$ and the statement $(**)$ follow from the same line of argument as for $(*)$.

- Finally, by the continuous mapping theorem and using the fact that $\mathbb{E}\left[\frac{\partial m_i(\phi_0)}{\partial \phi}\Lambda(Z_i)^2 \frac{\partial m_i(\phi_0)}{\partial \phi^T}\right]$ is invertible from Proposition 3.5, we obtain that $\hat{\Gamma}_h \xrightarrow[N\to+\infty]{P} \Gamma_h$.

(ii) **Second step:** we proceed to demonstrate the final result:

$$\frac{1}{\sqrt{N}}\sum_{i=1}^N \left(h(Z_i)\Lambda(Z_i)^2 - \hat{\Gamma}_h \mathbb{E}\left[\frac{\partial m(W_i,\hat{\phi})}{\partial \phi}\middle| Z_i\right]\Lambda(Z_i)^2\right) m(W_i,\hat{\phi}) = \frac{1}{\sqrt{N}}\sum_{i=1}^N h_\Lambda^\perp(Z_i) m(W_i,\phi_0) + o_p(1).$$

We consider a mean value expansion of $m(W_i,\hat{\phi})$ around ϕ_0. This expansion exists as $\phi \mapsto m(W_i,\hat{\phi})$ is continuously differentiable on Φ from Assumption 5 (1). Moreover, Assumption 2 (1) implies that $\phi_0 \in \text{interior}(\Phi)$ and thus, with probability that goes to one, $\hat{\phi} \in \Phi$.

Thus, we have the following decomposition:

$$\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\left(h(Z_i)\Lambda(Z_i)^2-\hat{\Gamma}_h\mathbb{E}\left[\frac{\partial m(W_i,\hat{\phi})}{\partial\phi}\middle|Z_i\right]\Lambda(Z_i)^2\right)m(W_i,\hat{\phi})=\frac{1}{\sqrt{N}}\sum_{i=1}^{N}h_\Lambda^{\perp}(Z_i)m(W_i,\phi_0)\quad(1)$$

$$+\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\left(h(Z_i)\Lambda(Z_i)^2-\hat{\Gamma}_h\mathbb{E}\left[\frac{\partial m(W_i,\hat{\phi})}{\partial\phi}\middle|Z_i\right]\Lambda(Z_i)^2\right)\frac{\partial m(W_i,\tilde{\phi})}{\partial\phi^T}(\hat{\phi}-\phi_0)\quad(2)$$

$$-\hat{\Gamma}_h\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\mathbb{E}\left[\frac{\partial m(W_i,\hat{\phi})}{\partial\phi}-\frac{\partial m(W_i,\phi_0)}{\partial\phi}\middle|Z_i\right]\Lambda(Z_i)^2m(W_i,\phi_0)\quad(3)$$

$$-(\hat{\Gamma}_h-\Gamma_h)\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\mathbb{E}\left[\frac{\partial m(W_i,\phi_0)}{\partial\phi}\middle|Z_i\right]\Lambda(Z_i)^2m(W_i,\phi_0)\quad(4)$$

where $\tilde{\phi}\in conv(\phi_0,\hat{\phi})$. We need to show that the second, third, and fourth terms converge to 0 in probability.

- For second term, from Assumption 2 on the first-stage estimator $\sqrt{N}(\hat{\phi}-\phi_0)=O_p(1)$ and from Proposition 3.5:

  $$\mathbb{E}\left[h_\Lambda^{\perp}(Z_i)\frac{\partial m(W_i,\phi_0)}{\partial\phi^T}\right]=0.$$

  Thus, it is sufficient to show:

  $$\frac{1}{N}\sum_{i=1}^{N}\left(h(Z_i)\Lambda(Z_i)^2-\hat{\Gamma}_h\mathbb{E}\left[\frac{\partial m(W_i,\hat{\phi})}{\partial\phi}\middle|Z_i\right]\Lambda(Z_i)^2\right)\frac{\partial m(W_i,\tilde{\phi})}{\partial\phi^T}\xrightarrow[N\to+\infty]{P}\mathbb{E}\left[h_\Lambda^{\perp}(Z_i)\frac{\partial m(W_i,\phi_0)}{\partial\phi^T}\right].\quad\text{(B.12)}$$

  Let us define: $g^{\dagger\dagger}(V_i,\hat{\phi},\tilde{\phi})\equiv\mathbb{E}\left[\frac{\partial m(W_i,\hat{\phi})}{\partial\phi}\middle|Z_i\right]\Lambda(Z_i)^2\frac{\partial m(W_i,\tilde{\phi})}{\partial\phi}$.

  And following almost the same steps that we used for $g^{\dagger}$ in (i) one can show that the following ULLN holds:

  $$\sup_{(\phi_a,\phi_b)\in\Phi^2}\left\|\frac{1}{N}\sum_{i=1}^{N}g^{\dagger\dagger}(V_i,\phi_a,\phi_b)-\mathbb{E}\left[g^{\dagger\dagger}(V_i,\phi_a,\phi_b)\right]\right\|\xrightarrow[N\to+\infty]{P}0$$

  Therefore, by the triangular inequality and the continuity of $g^{\dagger\dagger}$, we obtain $\frac{1}{N}\sum_{i=1}^{N}g^{\dagger\dagger}(V_i,\hat{\phi},\tilde{\phi})\xrightarrow[N\to+\infty]{P}\mathbb{E}\left[g^{\dagger\dagger}(V_i,\phi_0,\phi_0)\right]$. Finally, (i) yields $\hat{\Gamma}\xrightarrow[N\to+\infty]{P}\Gamma$ and combining these results with the LLN and the continuous mapping theorem, we get (B.12).

– Now we focus on the third term. From Assumption 5, we have that $m(W_i, \cdot)$ is twice continuously differentiable in ϕ on Φ with probability one and $\frac{\partial}{\partial\phi_j}\mathbb{E}\left[\frac{\partial m(W_i,\phi)}{\partial\phi^T}\middle|Z_i\right] = \mathbb{E}\left[\frac{\partial^2 m(W_i,\phi)}{\partial\phi_j\partial\phi^T}\middle|Z_i\right]$. Moreover, Assumption 2 (1) implies that $\phi_0 \in$ interior(Φ) and thus, with probability that goes to one, $\hat{\phi} \in \Phi$. Therefore, we can write a mean value expansion of $\mathbb{E}\left[\frac{\partial m(W_i,\hat{\phi})}{\partial\phi}\middle|Z_i\right]$ around ϕ_0, which gives us:

$$
\begin{aligned}
&\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\mathbb{E}\left[\frac{\partial m(W_i,\hat{\phi})}{\partial\phi} - \frac{\partial m(W_i,\phi_0)}{\partial\phi}\middle|Z_i\right]\Lambda(Z_i)^2 m(W_i,\phi_0)\\
&= \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\sum_{j=1}^{dim(\phi)}\mathbb{E}\left[\frac{\partial^2 m(W_i,\tilde{\phi})}{\partial\phi_j\partial\phi}\middle|Z_i\right](\hat{\phi}_j - \phi_{j0})\Lambda(Z_i)^2 m(W_i,\phi_0)\\
&= \sum_{j=1}^{dim(\phi)}\sqrt{N}(\hat{\phi}_j - \phi_{j0})\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left[\frac{\partial^2 m(W_i,\tilde{\phi})}{\partial\phi_j\partial\phi}\middle|Z_i\right]\Lambda(Z_i)^2 m(W_i,\phi_0)
\end{aligned}
$$

with $\tilde{\phi} \in conv(\hat{\phi}, \phi_0)$. From Assumption 2 on the first-stage estimator $\forall j \in [|1, dim(\phi)|]$, $\sqrt{N}(\hat{\phi} - \phi_0) = O_p(1)$. Moreover, under H_0: $\forall j$,

$$
\mathbb{E}\left[\mathbb{E}\left[\frac{\partial^2 m(W_i,\phi_0)}{\partial\phi_j\partial\phi}\middle|Z_i\right]\Lambda(Z_i)^2 m(W_i,\phi_0)\right] = 0.
$$

Therefore, it is sufficient to show that : $\forall j \in [|1, dim(\phi)|]$,

$$
\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left[\frac{\partial^2 m(W_i,\tilde{\phi})}{\partial\phi_j\partial\phi}\middle|Z_i\right]\Lambda(Z_i)^2 m(W_i,\phi_0) \xrightarrow[N\to+\infty]{P} \mathbb{E}\left[\mathbb{E}\left[\frac{\partial^2 m(W_i,\phi_0)}{\partial\phi_j\partial\phi}\middle|Z_i\right]\Lambda(Z_i)^2 m(W_i,\phi_0)\right]. \tag{B.13}
$$

As previously, let us define: $g_j^{\dagger\dagger\dagger}(V_i,\tilde{\phi}) \equiv \mathbb{E}\left[\frac{\partial^2 m(W_i,\tilde{\phi})}{\partial\phi_j\partial\phi}\middle|Z_i\right]\Lambda(Z_i)^2 m(W_i,\phi_0)$
Again, we want to show that a ULLN holds for $g_j^{\dagger\dagger\dagger}$.

$$
\sup_{\phi\in\Phi}\left|\left|\frac{1}{N}\sum_{i=1}^{N}g_j^{\dagger\dagger\dagger}(V_i,\phi) - \mathbb{E}\left[g_j^{\dagger\dagger\dagger}(V_i,\phi)\right]\right|\right|_2 \xrightarrow[N\to+\infty]{P} 0
$$

The usual regularity conditions hold and we have the following envelope for $g_j^{\dagger\dagger\dagger}$. With probability one, $\|g^{\dagger\dagger\dagger}(V_i,\phi)\|_2 \leq G^{\dagger\dagger\dagger}(V_i) \equiv \sqrt{dim(\phi)}\sqrt{p}\mathbb{E}[\tilde{\tilde{M}}(W_i)|Z_i]\|\Lambda(Z_i)\|_F\|\Lambda(Z_i)m(W_i,\phi_0)\|_F$

where $\|\cdot\|_F$ is the Frobenius norm.

$$
\begin{aligned}
\|g_j^{\dagger\dagger\dagger}(V_i,\phi)\|_2 &\leq \left\| \mathbb{E}\left[\frac{\partial^2 m(W_i,\tilde{\phi})}{\partial\phi_j\partial\phi}\middle| Z_i\right]\Lambda(Z_i)\right\|_F \|\Lambda(Z_i)m(W_i,\phi_0)\|_F \\
&\leq \left\| \mathbb{E}\left[\frac{\partial^2 m(W_i,\tilde{\phi})}{\partial\phi_j\partial\phi}\middle| Z_i\right]\right\|_F \|\Lambda(Z_i)\|_F\|\Lambda(Z_i)m(W_i,\phi_0)\|_F \\
&\leq \sqrt{dim(\phi)}\sqrt{p}\mathbb{E}[\tilde{\tilde{M}}(W_i)|Z_i]\|\Lambda(Z_i)\|_F\|\Lambda(Z_i)m(W_i,\phi_0)\|_F.
\end{aligned}
$$

The first two inequalities stems from the under-multiplicativity of the Frobenius norm. The last inequality comes from Assumption 5 (4). Moreover, from Assumptions 4, 5 (4) and Asumption 6, along with the application of the Cauchy-Schwarz and Jensen inequalities, we have:

$$
\begin{aligned}
\mathbb{E}[G^{\dagger\dagger\dagger}(V_i)] &\leq \sqrt{dim(\phi)}\sqrt{p}\mathbb{E}[\mathbb{E}[\tilde{\tilde{M}}(W_i)|Z_i]^4]^{1/4}\mathbb{E}[\|\Lambda(Z_i)\|_F^4]^{1/4}\mathbb{E}\left[\|\Lambda(Z_i)m(W_i,\phi_0)\|_F^2\right]^{1/2} \\
&\leq \sqrt{dim(\phi)}\sqrt{p}\mathbb{E}[\tilde{\tilde{M}}(W_i)^4]^{1/4}\mathbb{E}[\|\Lambda(Z_i)\|_F^4]^{1/4}\mathbb{E}\left[\|\Lambda(Z_i)m(W_i,\phi_0)\|_F^2\right]^{1/2} \\
&\leq \sqrt{dim(\phi)}\sqrt{p}\overline{\tilde{\tilde{M}}}^{1/4}\overline{\Lambda}^{1/4}\sqrt{\overline{M}}.
\end{aligned}
$$

Therefore, a ULLN holds for $g_j^{\dagger\dagger\dagger}(W_i,\tilde{\phi})$ and combined with the triangular inequality, we obtain ($B.13$). Finally from the continuous mapping theorem and $\hat{\Gamma}_h \xrightarrow[N\to+\infty]{P} \Gamma_h$, we get that (3) converges in probability to 0.

– For the fourth term, we have under H_0 : $\mathbb{E}\left[\left[\frac{\partial m(W_i,\phi_0)}{\partial\phi}\middle| Z_i\right]\Lambda(Z_i)^2 m(W_i,\phi_0)\right] = 0$. Moreover, Assumption 5 (6) implies $\exists C^{'''}$ such that:

$$
\mathbb{E}\left[\left\|\mathbb{E}\left[\frac{\partial m(W_i,\phi_0)}{\partial\phi}\middle| Z_i\right]\Lambda(Z_i)^2 m(W_i,\phi_0)\right\|_F^2\right] < C^{'''}.
$$

Thus, we can invoke the multivariate CLT, which implies: $\frac{1}{\sqrt{N}}\sum_{i=1}^N \mathbb{E}\left[\frac{\partial m(W_i,\phi_0)}{\partial\phi}\middle| Z_i\right]\Lambda(Z_i)^2 m(W_i,\phi_0) = O_p(1)$ and (i) gives $\hat{\Gamma}_h \xrightarrow[N\to+\infty]{P} \Gamma_h$. By combining these two results, we obtain that the fourth term converges to 0 in probability.

Second, we want to show that under $\overline{H}'_a$:

$$\frac{1}{N}\sum_{i=1}^{N}\hat{h}_\Lambda^\perp(Z_i)m(W_i,\hat{\phi}) = \mathbb{E}\left[h_\Lambda^\perp(Z_i)m(W_i,\phi_0)\right] + o_p(1).$$

By definition:

$$\frac{1}{N}\sum_{i=1}^{N}\hat{h}_\Lambda^\perp(Z_i)m(W_i,\hat{\phi}) = \frac{1}{N}\sum_{i=1}^{N}\left(h(Z_i)\Lambda(Z_i)^2 - \hat{\Gamma}_h\mathbb{E}\left[\frac{\partial m(W_i,\hat{\phi})}{\partial\phi}\middle|Z_i\right]\Lambda(Z_i)^2\right)m(W_i,\hat{\phi})$$
$$\equiv \frac{1}{N}\sum_{i=1}^{N}g^+(V_i,\hat{\phi}) - \hat{\Gamma}_h\frac{1}{N}\sum_{i=1}^{N}g^{++}(V_i,\hat{\phi}).$$

with $g^+(V_i,\phi) = h(Z_i)\Lambda(Z_i)^2m(W_i,\phi)$ and $g^{++}(V_i,\phi) = \mathbb{E}\left[\frac{\partial m(W_i,\phi)}{\partial\phi}\middle|Z_i\right]\Lambda(Z_i)^2m(W_i,\phi)$.

- From the first part of the proof: $\hat{\Gamma}_h \xrightarrow[N\to+\infty]{P} \Gamma_h$.
- We can prove that a ULLN holds for $g^+(V_i,\phi)$ on Φ. The usual conditions stated previously hold to apply Theorem 18.2 in Hansen (2022). The only argument that we show explicitly is the envelope on g^+. We have $\forall\phi\in\Phi$,

$$\|g^+(V_i,\phi)\|_2 = \|h(Z_i)\Lambda(Z_i)^2m(W_i,\phi)\|_2 \leq \|h(Z_i)\Lambda(Z_i)\|_F\|\Lambda(Z_i)m(W_i,\phi)\|_2$$
$$\leq \|h(Z_i)\Lambda(Z_i)\|_F\|\Lambda(Z_i)\|_F\|m(W_i,\phi)\|_2 = \|h(Z_i)\Lambda(Z_i)\|_F\|\Lambda(Z_i)\|_F\sqrt{p}M(W_i) \equiv G^+(V_i).$$

  The first two inequalities stem from the under-multiplicativity of the norms. The last inequality comes from Assumption 5 (1). Moreover, from Assumptions 4, 5 (1) and Asumption 6, along with the application of the Cauchy-Schwarz, we have:

$$\mathbb{E}[G^+(V_i)] \leq \sqrt{p}\sqrt{C''}\,\overline{M}^{1/4}\overline{\Lambda}^{1/4}.$$

  Following the same steps that we did several times previously, we have $\frac{1}{N}\sum_{i=1}^{N}g^+(V_i,\hat{\phi}) \xrightarrow[N\to+\infty]{P} \mathbb{E}\left[h(Z_i)\Lambda(Z_i)^2m(W_i,\phi_0)\right]$.
- With very analogous arguments that we omit for conciseness, we can easily show that $\frac{1}{N}\sum_{i=1}^{N}g^{++}(V_i,\hat{\phi}) \xrightarrow[N\to+\infty]{P} \mathbb{E}\left[\mathbb{E}\left[\frac{\partial m(W_i,\phi)}{\partial\phi}\middle|Z_i\right]\Lambda(Z_i)^2m(W_i,\phi)\right]$.

By combining the results above and applying the continuous mapping theorem, we obtain the desired result.

□

Proof of Proposition 3.6.

Let $h \in \mathcal{H}_0^{\perp}$. To shorten notations, for any $\tilde{\phi}$, let $m(\tilde{\phi}) \equiv m(W_i, \tilde{\phi})$. Following the proof of Proposition 3.2, under $\overline{H}_a : (\beta, F) = (\beta_a, F_a)$, the asymptotic slope of the test is equal to:

$$\begin{aligned} c(h, \phi_0) =& \mathbb{E}\left[h(Z_i)m(W_i, \phi_0)\right]^T \Omega_0^{-1} \mathbb{E}\left[h(Z_i)m(W_i, \phi_0)\right] \\ =& \mathbb{E}\left[h(Z_i)\Delta_{0,a}^m\right]^T \mathbb{E}\left[\big(h(Z_i)m(\phi_0)\big)\big(h(Z_i)m(\phi_0)\big)^T\right]^{-1} \mathbb{E}\left[h(Z_i)\Delta_{0,a}^m\right]. \end{aligned}$$

As $h \in \mathcal{H}_0^{\perp}$, for any $\Gamma \in \mathcal{M}_{dim(\phi)\times 1}(\mathbb{R})$, we can rewrite the slope as follows:

$$c(h, \phi_0) = \mathbb{E}\left[h(Z_i)\left(\Delta_{0,a}^m - \frac{\partial m(\phi_0)}{\partial \phi^T}\Gamma\right)\right]^T \mathbb{E}\left[\big(h(Z_i)m(\theta_0)\big)\big(h(Z_i)m(\theta_0)\big)^T\right]^{-1} \mathbb{E}\left[h(Z_i)\left(\Delta_{0,a}^m - \frac{\partial m(\phi_0)}{\partial \phi^T}\Gamma\right)\right].$$

Following the same steps as the ones displayed in the proof of Proposition 3.2, we can show that for any $\Gamma \in \mathcal{M}_{dim(\phi)\times 1}(\mathbb{R})$, the instrument h_Γ^* that maximizes the slope has the following expression:

$$h_\Gamma^*(Z_i) = \big(\mathbb{E}(m(\phi_0)m(\phi_0)^T|Z_i)^{-1}\mathbb{E}\big(\,\Delta_{0,a}^m - \frac{\partial m(\phi_0)}{\partial \phi^T}\Gamma|Z_i\,)\big)^T$$

It follows that for any Γ, h_Γ^* yields an asymptotic slope that is greater than the one generated by any instrument in $\mathcal{H}_0^{\perp}$. To finish the proof, we must find Γ such that $h_\Gamma^* \in \mathcal{H}_0^{\perp}$. We can show that the following choice of $\Gamma^\star$ makes the instrument Neyman orthogonal:

$$\Gamma^\star = \mathbb{E}\left[\frac{\partial m(\phi_0)}{\partial \phi}\Sigma(Z_i)^{-1}\frac{\partial m(\phi_0)}{\partial \phi^T}\right]^{-1}\left[\frac{\partial m(\phi_0)}{\partial \phi}\Sigma(Z_i)^{-1}\mathbb{E}[\Delta_{0,a}^m|Z_i]\right]$$

with $\Sigma(Z_i) = \mathbb{E}(m(\phi_0)m(\phi_0)^T|Z_i)$. One can easily check that the resulting MPI $h_{\Gamma^\star}^*$

corresponds to the following orthogonalized instrument:

$$h^*_{\Gamma^\star}(Z_i) = \left(\mathbb{E}[\Delta^m_{0,a}(W_i)|Z_i]^T\right)^{\perp}_{\Sigma^{-1/2}}.$$

Moreover, under the assumptions stated in Proposition 3.6, and Assumption 3 (2), along with the Cauchy-Schwarz inequality, the following inequalities hold:

- there exists a constant C' such that $\mathbb{E}[\|h^*_{\Gamma^\star}(Z_i)\|_F^4] < C'$,
- $\mathbb{E}[\|h^*_{\Gamma^\star}(Z_i)m(W_i,\phi_0)\|_2^2] \leq \sqrt{\mathbb{E}[\|h^*_{\Gamma^\star}(Z_i)\|_2^4 \mathbb{E}[\|m(W_i,\phi_0)\|_2^4]} \leq \sqrt{p^2 C' \overline{M}}$,
- $\Omega_0 = \mathbb{E}[h^*_{\Gamma^\star}(Z_i)m(W_i,\phi_0)\left(h^*_{\Gamma^\star}(Z_i)m(W_i,\phi_0)\right)^T]$ is full rank.

Thus, $h^*_{\Gamma^\star} \in \mathcal{H}_0^\perp$. □

Proof of Proposition 1.3.

Let $h \in \mathcal{H}_0$. Furthermore, we assume that Assumption 7 is satisfied, m satisfies Assumption 3. By definition (when $h^\dagger$ exists),

$$h^\dagger(Z_i) = h(Z_i) - BLP\left(h(Z_i)\middle| BLP\left(\frac{\partial m(W_i,\phi_0)}{\partial \phi}\middle| h(Z_i)^T\right)^T\right).$$

First, we develop:

$$\begin{aligned} BLP\left(\frac{\partial m(W_i,\phi_0)}{\partial \phi}\middle| h(Z_i)^T\right) &= \mathbb{E}\left[\frac{\partial m(W_i,\phi_0)}{\partial \phi} h(Z_i)^T\right]\mathbb{E}\left[h(Z_i)h(Z_i)^T\right]^{-1} h(Z_i) \\ &\equiv \tilde{\Gamma}_h h(Z_i) \end{aligned}$$

with $\tilde{\Gamma}_h = \mathbb{E}\left[\frac{\partial m(W_i,\phi_0)}{\partial \phi} h(Z_i)^T\right]\mathbb{E}\left[h(Z_i)h(Z_i)^T\right]^{-1}$. It follows:

$$\begin{aligned} BLP\left(h(Z_i)\middle| BLP\left(\frac{\partial m(W_i,\phi_0)}{\partial \phi}\middle| h(Z_i)^T\right)^T\right) &= \mathbb{E}\left[h(Z_i)h(Z_i)^T\tilde{\Gamma}_h^T\right]\mathbb{E}\left[\tilde{\Gamma}_h h(Z_i)h(Z_i)^T\tilde{\Gamma}_h^T\right]^{-1}\tilde{\Gamma}_h h(Z_i) \\ &= \mathbb{E}\left[h(Z_i)h(Z_i)^T\right]\tilde{\Gamma}_h^T\left(\tilde{\Gamma}_h\mathbb{E}\left[h(Z_i)h(Z_i)^T\right]\tilde{\Gamma}_h^T\right)^{-1}\tilde{\Gamma}_h h(Z_i) \\ &\equiv \Gamma_h h(Z_i). \end{aligned}$$

Now let us verify that Γ_h exists and is finite. From Assumption 3, $\frac{\partial m(W_i,\phi_0)}{\partial \phi}$ exists with probability one. Moreover, from Assumption 7, the following moments are bounded:

$$\begin{aligned}\mathbb{E}\left[\left\|\frac{\partial m(W_i,\phi_0)}{\partial\phi}h(Z_i)^T\right\|_F\right] &\leq \mathbb{E}\left[\left\|\frac{\partial m(W_i,\phi_0)}{\partial\phi}\right\|_F \|h(Z_i)\|_F\right] \\ &\leq \sqrt{\mathbb{E}\left[\left\|\frac{\partial m(W_i,\phi_0)}{\partial\phi}\right\|_F^2\right][\|h(Z_i)\|_F^2]} < C^{\dagger 3/4} \\ &\text{and } \mathbb{E}\left[\|h(Z_i)h(Z_i)^T\|_F\right] \leq \mathbb{E}\left[\|h(Z_i)\|_F^2\right] < \sqrt{C^\dagger}.\end{aligned}$$

where we use the fact that $\|AB\|_F \leq \|A\|_F\|B\|_F$ for two compatible matrices A and B and the Cauchy Schwarz inequality. Furthermore, from Assumption 7, $\mathbb{E}\left[h(Z_i)h(Z_i)^T\right]$ is full rank and thus invertible. Consequently, $\tilde{\Gamma}_h$ exists. Using standard results on matrix rank and Assumption 7 (2), which states that $rank\left(\mathbb{E}\left[\frac{\partial m(W_i,\phi_0)}{\partial\phi}h(Z_i)^T\right]\right) = dim(\phi)$, we obtain:

$$rank(\tilde{\Gamma}_h) = dim(\phi).$$

It follows, that $rank(\tilde{\Gamma}_h\mathbb{E}\left[h(Z_i)h(Z_i)^T\right]\tilde{\Gamma}_h{}^T) = dim(\phi)$ and thus, it is invertible and Γ_h is well-defined. Next, one can easily check that Γ_h is a projection in $\mathbb{R}^L$ as it satisfies $\Gamma_h^2 = \Gamma_h$. Thus, $Spectrum(\Gamma_h) \subset \{0,1\}$ and it follows that:

$$\begin{aligned}rank(\Gamma_h) = tr(\Gamma_h) &= tr(\mathbb{E}\left[h(Z_i)h(Z_i)^T\right]\tilde{\Gamma}_h^T\left(\tilde{\Gamma}_h\mathbb{E}\left[h(Z_i)h(Z_i)^T\right]\tilde{\Gamma}_h^T\right)^{-1}\tilde{\Gamma}_h) \\ &= tr\left(\tilde{\Gamma}_h\mathbb{E}\left[h(Z_i)h(Z_i)^T\right]\tilde{\Gamma}_h^T\left(\tilde{\Gamma}_h\mathbb{E}\left[h(Z_i)h(Z_i)^T\right]\tilde{\Gamma}_h^T\right)^{-1}\right) = tr(I_{dim(\phi)}) = dim(\phi)\end{aligned}$$

Finally, we show that $\mathbb{E}\left[h^\dagger(Z_i)m(W_i,\phi_0)\right]$ satisfies the Neyman orthogonality property. We leverage the properties of the best linear projection operator, which, in particular, entails that for any $(A_i,B_i) \in \mathcal{M}_{k,p}(\mathbb{R}) \times \mathcal{M}_{p,m}(\mathbb{R})$ with $(k,m) \in \mathbb{N}^{*2}$, $\mathbb{E}\left[A_iB_i\right] = \mathbb{E}\left[BLP(A_i|B_i)B_i\right] = \mathbb{E}\left[A_iBLP(B_i^T|A_i^T)^T\right]$ $(*)$

$$
\begin{aligned}
\mathbb{E}\left[h^{\dagger}(Z_i)\frac{\partial m(W_i,\phi_0)}{\partial \phi^T}\right] = \; & \mathbb{E}\left[h(Z_i)\frac{\partial m(W_i,\phi_0)}{\partial \phi^T}\right] - \mathbb{E}\left[BLP\left(h(Z_i)\middle|BLP\left(\frac{\partial m(W_i,\phi_0)}{\partial \phi}\middle|h(Z_i)^T\right)^T\right)\frac{\partial m(W_i,\phi_0)}{\partial \phi^T}\right] \\
= \; & \mathbb{E}\left[h(Z_i)BLP\left(\frac{\partial m(W_i,\phi_0)}{\partial \phi}\middle|h(Z_i)^T\right)^T\right] \\
& - \mathbb{E}\left[BLP\left(h(Z_i)\middle|BLP\left(\frac{\partial m(W_i,\phi_0)}{\partial \phi}\middle|h(Z_i)^T\right)^T\right)\frac{\partial m(W_i,\phi_0)}{\partial \phi^T}\right] \\
= \; & \mathbb{E}\left[h(Z_i)BLP\left(\frac{\partial m(W_i,\phi_0)}{\partial \phi}\middle|h(Z_i)^T\right)^T\right] - \mathbb{E}\left[h(Z_i)BLP\left(\frac{\partial m(W_i,\phi_0)}{\partial \phi}\middle|h(Z_i)\right)^T\right]\times \\
& \mathbb{E}\left[BLP\left(\frac{\partial m(W_i,\phi_0)}{\partial \phi}\middle|h(Z_i)^T\right)BLP\left(\frac{\partial m(W_i,\phi_0)}{\partial \phi}\middle|h(Z_i)^T\right)^T\right]^{-1}\times \\
& \mathbb{E}\left[BLP\left(\frac{\partial m(W_i,\phi_0)}{\partial \phi}\middle|h(Z_i)^T\right)\frac{\partial m(W_i,\phi_0)}{\partial \phi^T}\right] \\
= \; & \mathbb{E}\left[h(Z_i)BLP\left(\frac{\partial m(W_i,\phi_0)}{\partial \phi}\middle|h(Z_i)^T\right)^T\right] - \mathbb{E}\left[h(Z_i)BLP\left(\frac{\partial m(W_i,\phi_0)}{\partial \phi}\middle|h(Z_i)\right)^T\right]\times \\
& \mathbb{E}\left[BLP\left(\frac{\partial m(W_i,\phi_0)}{\partial \phi}\middle|h(Z_i)\right)BLP\left(\frac{\partial m(W_i,\phi_0)}{\partial \phi}\middle|h(Z_i)^T\right)^T\right]^{-1}\times \\
& \mathbb{E}\left[BLP\left(\frac{\partial m(W_i,\phi_0)}{\partial \phi}\middle|h(Z_i)^T\right)BLP\left(\frac{\partial m(W_i,\phi_0)}{\partial \phi}\middle|h(Z_i)^T\right)^T\right] \\
= \; & \mathbb{E}\left[h(Z_i)BLP\left(\frac{\partial m(W_i,\phi_0)}{\partial \phi}\middle|h(Z_i)^T\right)^T\right] - \mathbb{E}\left[h(Z_i)BLP\left(\frac{\partial m(W_i,\phi_0)}{\partial \phi}\middle|h(Z_i)\right)^T\right] = 0
\end{aligned}
$$

$\square$

Proof of Proposition 1.4.

Let $h \in \mathcal{H}_0$. Further, assume that Assumptions 7 and 8 are satisfied, $\hat{\phi}$ satisfies Assumption 2, m satisfies Assumption 3.

(i) As a preliminary result, let us show that $\hat{\Gamma}_h \xrightarrow[N\to+\infty]{P} \Gamma_h$. To show this statement, we proceed in three steps.

- $\frac{1}{N}\sum_{i=1}^{N} h(Z_i)\frac{\partial m(W_i,\hat{\phi})}{\partial \phi^T} \xrightarrow[N\to+\infty]{P} \mathbb{E}\left[h(Z_i)\frac{\partial m(W_i,\phi_0)}{\partial \phi^T}\right]$ $(*)$
  To show this statement, we can prove that a ULLN holds for $g(V_i,\phi) \equiv h(Z_i)\frac{\partial m(W_i,\phi_0)}{\partial \phi^T}$ on Φ. The usual conditions stated previously hold to apply Theorem 18.2 in Hansen (2022) (see for instance the proof of proposition 3.4). The only argument that we show explicitly is the envelope on g:

$\|g(V_i,\phi)\|_F \leq G(V_i) \equiv \tilde{M}(W_i)\sqrt{dim(\phi)}\sum_{l=1}^{L}\|h_l(Z_i)\|_2$ where $\|\cdot\|_F$ is the Frobenius norm. Indeed,

$$
\begin{aligned}
\|g(V_i,\phi)\|_F &= \sqrt{\sum_{l=1}^{L}\sum_{j=1}^{\dim(\phi)}\left|h_l(Z_i)^T\frac{\partial m(W_i,\phi)}{\partial\phi_j}\right|^2} \leq \sqrt{\sum_{l=1}^{L}\sum_{j=1}^{\dim(\phi)}\|h_l(Z_i)\|_2^2\tilde{M}(W_i)^2} \\
&= \tilde{M}(W_i)\sqrt{dim(\phi)}\sqrt{\sum_{l=1}^{L}\|h_l(Z_i)\|_2^2} \\
&= \tilde{M}(W_i)\sqrt{dim(\phi)}\|h(Z_i)\|_F.
\end{aligned}
$$

The first inequality comes from Assumption 3 (3). The last inequality comes from the triangular inequality. Moreover, from Assumption 7 (1) Assumption 3 (3) and the Cauchy–Schwarz inequality, we have: $\mathbb{E}\left[(g(V_i)\right] \leq \sqrt{\tilde{\tilde{M}}}\sqrt{dim(\phi)}L{C^{\dagger}}^{1/4}$. Therefore, we can apply a ULLN and obtain $(*)$ by using the same steps as in the proof of proposition 3.4 for instance.

- From the proof of proposition 1.3, we have $\mathbb{E}\left[\|h(Z_i)h(Z_i)^T\|_F\right]$ is bounded. Thus by the LLN, $\frac{1}{N}\sum_{i=1}^{N}h(Z_i)h(Z_i)^T \xrightarrow[N\to+\infty]{P} \mathbb{E}\left[h(Z_i)h(Z_i)^T\right]$ $(**)$.
- Combining $(*)$ and $(**)$ and applying the continuous mapping theorem, we have: $\hat{\Gamma}_h \xrightarrow[N\to+\infty]{P} \Gamma_h$.

(ii) **Second, we examine the case** $H_0 : F \in \mathcal{F}_0$. We consider a mean value expansion of $m(W_i,\hat{\phi})$ around ϕ_0. This expansion exists as $\phi \mapsto m(W_i,\hat{\phi})$ is continuously differentiable on Φ from Assumption 3 (1). Moreover, Assumption 2 (1) implies that $\phi_0 \in \text{interior}(\Phi)$ and thus, with probability that goes to one, $\hat{\phi} \in \Phi$.

$$
\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\hat{h}^{\dagger}(Z_i)m(W_i,\hat{\phi}) = \underbrace{\frac{1}{\sqrt{N}}(I_L-\hat{\Gamma}_h)\sum_{i=1}^{N}h(Z_i)m(W_i,\phi_0)}_{A_N} + \underbrace{\frac{1}{\sqrt{N}}(I_L-\hat{\Gamma}_h)\sum_{i=1}^{N}h(Z_i)\frac{\partial m(W_i,\tilde{\phi})}{\partial\phi^T}(\hat{\phi}-\phi_0)}_{B_N}
$$

where $\tilde{\phi} \in conv(\phi_0,\hat{\phi})$.

- Following the same steps as in the proof of Proposition 3.4 and using part

(i) $\hat{\Gamma}_h \xrightarrow[N\to+\infty]{P} \Gamma_h$, we can show that:

$$(I_L - \hat{\Gamma}_h)\frac{1}{N}\sum_{i=1}^{N} h(Z_i)\frac{\partial m(W_i,\tilde{\phi})}{\partial \phi^T} \xrightarrow[N\to+\infty]{P} (I_L - \Gamma_h)\mathbb{E}\left[h(Z_i)\frac{\partial m(W_i,\phi_0)}{\partial \phi^T}\right].$$

From Proposition 1.3, $(I_L - \Gamma_h)\mathbb{E}\left[h(Z_i)\frac{\partial m(W_i,\phi_0)}{\partial \phi^T}\right] = 0$. From Assumption 2, $\sqrt{N}(\hat{\phi} - \phi_0) = O_p(1)$. It follows from the continuous mapping theorem that $B_N = o_p(1)$.

– Second, by using (i) and applying Slutsky's lemma, we have that $A_N \xrightarrow[N\to+\infty]{d} (I_L - \Gamma_h)\mathcal{N}(0, \tilde{\Omega}_0)$. Furthermore, let us assume that we have a consistent estimator of Ω_0. We show in (iv) that the empirical counterpart of $\tilde{\Omega}_0$ obtained by replacing $\hat{\phi}$ with ϕ_0 is a consistent estimator of $\tilde{\Omega}_0$.

  Now the difference here with the classical case in proposition 3.1, is that $\Omega_0 \equiv (I_L - \Gamma_h)\tilde{\Omega}_0(I_L - \Gamma_h^T)$ is not invertible. Therefore, in the definition of $\tilde{S}_N$ we use the general inverse instead of the inverse of Ω_0. However, this is not without loss of generality as highlighted in Andrews (1987), the continuous mapping theorem does not apply because general inverses are not continuous. Thus, to derive the asymptotic distribution of our test statistic, we need additional regularity conditions stated in Andrews (1987). Namely we require assumption 8 to hold, $\mathbb{P}\left(\{rank(\hat{\tilde{\Omega}}_0) = L\} \cap \{rank(\hat{\Gamma}_h) = dim(\phi)\}\right) \xrightarrow[N\to+\infty]{} 1$. If these conditions are satisfied, we can apply theorem 1 in Andrews (1987) to recover the following result:

$$\tilde{S}_N(\hat{h}^\dagger, \hat{\phi}) \xrightarrow[N\to+\infty]{d} \chi^2_{L-dim(\phi)}.$$

(iii) **Third, we examine the case** $\overline{H}'_a :\ (I_L - \Gamma_h)\mathbb{E}\left[h(Z_i)m(W_i,\phi_0)\right] \neq 0$. Directly, from case (i), we have:

$$(I_L - \hat{\Gamma}_h)\frac{1}{N}\sum_{i=1}^{N} h(Z_i)m(W_i,\hat{\phi}) = (I_L - \Gamma_h)\frac{1}{N}\sum_{i=1}^{N} h(Z_i)m(W_i,\phi_0) + o_p\left(\frac{1}{\sqrt{N}}\right)$$

The data are i.i.d. , $\mathbb{E}\left[\|h(Z_i)m(W_i,\phi_0)\|_2\right] < \sqrt{C}$ (as $h \in \mathcal{H}_0$). Therefore, we can apply the law of large numbers:

$$\frac{1}{N}\sum_{i=1}^{N} h(Z_i)m(W_i,\hat{\phi}) \underset{N\to+\infty}{\xrightarrow{P}} \mathbb{E}\left[h(Z_i)m(W_i,\phi_0)\right]$$

Moreover, (i) implies $\hat{\Gamma}_h \underset{N\to+\infty}{\xrightarrow{P}} \Gamma_h$. Therefore, by the continuous mapping theorem:

$$(I_L-\hat{\Gamma}_h)\frac{1}{N}\sum_{i=1}^{N} h(Z_i)m(W_i,\hat{\phi}) \underset{N\to+\infty}{\xrightarrow{P}} (I_L-\Gamma_h)\mathbb{E}\left[h(Z_i)m(W_i,\phi_0)\right].$$

(iv) Following the same line of argument as in the proof of Proposition 3.4 (iii) , we can show that $\hat{\tilde{\Omega}}_0 \underset{N\to+\infty}{\xrightarrow{P}} \tilde{\Omega}_0$.

□

B.2.3 Connection with the optimal instruments

In this section, we assume that the model is well-specified ($F \in \mathcal{F}_0$) and that the parameter of interest ϕ_0 is identified by the non-linear conditional moment restriction in (2.1).[42] The researcher wants to estimate the true parameter $\phi_0 = (\beta_0^T, \lambda_0^T)^T$ by GMM and must choose the set of instruments $h_E(Z_i)$ (or equivalently, the unconditional moments) to include in the GMM objective function:

$$\hat{\phi} = \underset{\tilde{\phi}\in\Phi}{\operatorname{argmin}}\ N\left(\frac{1}{N}\sum_{i=1}^{N} h_E(Z_i)m(W_i,\tilde{\phi})\right)^T \hat{\Omega}_0^{-1}\left(\frac{1}{N}\sum_{i=1}^{N} h_E(Z_i)m(W_i,\tilde{\phi})\right).$$

with $\hat{\Omega}_0$ a consistent estimator of Ω_0 the asymptotic variance-covariance matrix of $\frac{1}{\sqrt{N}}\sum_i h_E(Z_i)m(W_i,\phi_0)$. The optimal instruments $h_E(Z_i)$ are chosen to minimize the asymptotic variance-covariance of the estimator $\hat{\phi}$. The resulting estimator is said to be efficient in the sense that its asymptotic variance cannot be reduced by using additional moment conditions. The derivation of the optimal instruments in this context has been

[42] The identification conditions in the parametric case are less stringent than the conditions for the non-parametric identification as defined in 1.1.

studied by Amemiya (1974). For an arbitrary choice of $h_E(Z_i)$, the GMM estimator with the 2-step efficient weighting matrix has the following asymptotic distribution:

$$\sqrt{T}(\hat{\phi}-\phi_0) \xrightarrow{d} \mathcal{N}\left(0, (\Gamma(\mathcal{F}_0, h_E)^T \Omega(\mathcal{F}_0, h_E)^{-1}\Gamma(\mathcal{F}_0, h_E))^{-1}\right),$$

with:

$$\Omega(\mathcal{F}_0, h_E) = \mathbb{E}\Big[\Big(h_E(Z_i)m(W_i,\phi_0)\Big)\Big(h_E(Z_i)m(W_i,\phi_0)\Big)^T\Big]$$
$$\text{and} \quad \Gamma(\mathcal{F}_0, h_E) = \mathbb{E}\left[h_E(Z_i)\frac{\partial m(W_i,\phi_0)}{\partial \phi^T}\right].$$

Following Chamberlain (1987) and Amemiya (1974), the following Lemma gives the form of the optimal instruments in our setting.

Lemma 2.1 *Optimal instruments.*

The optimal instruments $h_E^*(Z_i)$ can be written as follows:

$$h_E^*(Z_i) = \left(\mathbb{E}\left[m(W_i,\phi_0)m(W_i,\phi_0)^T|Z_i\right]^{-1}\mathbb{E}\left[\frac{\partial m(W_i,\phi_0)}{\partial\phi^T}\middle| Z_i\right]\right)^T.$$

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

$$V^* = \mathbb{E}\left[\mathbb{E}\left[\frac{\partial m(W_i,\phi_0)}{\partial\tilde{\phi}^T}\middle| Z_i\right]^T \mathbb{E}\left[m(W_i,\phi_0)m(W_i,\phi_0)^T|Z_i\right]^{-1}\mathbb{E}\left[\frac{\partial m(W_i,\phi_0)}{\partial\phi^T}\middle| Z_i\right]\right]$$

Proof. To shorten the notations, we denote: $\Sigma(Z_i) = \mathbb{E}\left[m(W_i,\phi_0)m(W_i,\phi_0)^T|Z_i\right]$, $D(Z_i) = \mathbb{E}\left[\frac{\partial m(W_i,\phi_0)}{\partial\phi^T}\middle| Z_i\right]$ and $\Omega_0(h_E) = \Omega(\mathcal{F}_0, h_E)$.

We want to prove that for any set of instruments $h_E(Z_i)$ that $V^* - \Gamma_0(h_E)^T\Omega_0(h_E)^{-1}\Gamma_0(h_E)$ matrix is semi-definite positive.

$$\begin{aligned} V^* - \Gamma_0(h_E)^T\Omega_0(h_E)^{-1}\Gamma_0(h_E) &= \mathbb{E}\Big[D(Z_i)^T\Sigma(Z_i)^{-1}D(Z_i)\Big] - \mathbb{E}\left[h_E(Z_i)D(Z_i)\right]^T\Omega_0(h_E)^{-1}\mathbb{E}\left[h_E(Z_i)D(Z_i)\right] \\ &= \mathbb{E}\Big[\tilde{D}(Z_i)^T\tilde{D}(Z_i)\Big] - \mathbb{E}\Big[\tilde{D}(Z_i)^T\tilde{h}_E(Z_i)\Big]\mathbb{E}\Big[\tilde{h}_E(Z_i)^T\tilde{h}_E(Z_i)\Big]^{-1}\mathbb{E}\Big[\tilde{h}_E(Z_i)^T\tilde{D}(Z_i)\Big]. \end{aligned}$$

with $\tilde{D}(Z_i) = \Sigma(Z_i)^{1/2}D(Z_i)$ and $\tilde{h}_E(Z_i) = \Sigma(Z_i)^{1/2}h_E(Z_i)$. It follows that we have:

$$V^*(Z_i) - \Gamma_0(h_E)\Omega_0(h_E)^{-1}\Gamma_0(h_E) = \tilde{M}^T\mathbb{E}[\tilde{X}'\tilde{X}]\tilde{M}$$

with $\tilde{X} = \begin{pmatrix} \tilde{D}(Z_i) & \tilde{h}_E(Z_i) \end{pmatrix}$ and $\tilde{M} = \begin{pmatrix} I_{|\theta_0|} & -\mathbb{E}\Big[\tilde{D}(Z_i)^T\tilde{h}_E(Z_i)\Big]\mathbb{E}\Big[\tilde{h}_E(Z_i)^T\tilde{h}_E(Z_i)\Big]^{-1} \end{pmatrix}^T$

The matrix above is semi-definite positive. □

B.3 A feasible approximation of the MPI

B.3.1 A theoretical basis for the interval instruments

Proof of Proposition 4.1.

Let $h \in \mathcal{H}_0$ and we assume homoscedasticity (i.e., under $\overline{H}_0$, we assume: $\mathbb{E}\left[m(W_i,\theta_0)m(W_i,\theta_0)^T|Z_i\right] = \sigma^2 I_p$), then the slope associated with h exists and can be expressed as follows:

$$c(h,\theta_0) = \sigma^{-2}\mathbb{E}\left[h(Z_i)m(W_i;\theta_0)\right]^T\mathbb{E}\left[h(Z_i)h(Z_i)^T\right]^{-1}\mathbb{E}\left[h(Z_i)m(W_i,\theta_0)\right]$$

On the other hand, for any $h \in \mathcal{H}_0$, the best linear predictor of $m(W_i,\theta_0)^T$ on $h(Z_i)^T = (h_1(Z_i),...,h_L(Z_i))$ exists and can be expressed as:

$$BLP(m(W_i,\theta_0)^T|h(Z_i)^T) = \Gamma_h^T h(Z_i) = \mathbb{E}\left[h(Z_i)m(W_i,\theta_0)\right]^T\mathbb{E}\left[h(Z_i)h(Z_i)^T\right]^{-1}h(Z_i)$$

with $\Gamma_h = \underset{\tilde{\Gamma}\in\mathbb{R}^L}{min}\{\mathbb{E}[\|m(W_i,\theta_0) - h(Z_i)^T\tilde{\Gamma})\|_2^2]\}$

Thus, one can easily show that the expectation of the squared norm of this projection is equal to:

$$\mathbb{E}\big[\|BLP(m(W_i,\theta_0)^T|h(Z_i)^T)\|_2^2] = \mathbb{E}\left[h(Z_i)m(W_i;\theta_0)\right]^T\mathbb{E}\left[h(Z_i)h(Z_i)^T\right]^{-1}\mathbb{E}\left[h(Z_i)m(W_i,\theta_0)\right]$$

□

Proof of Proposition 1.5.

Let $h \in \mathcal{H}_0$, then the slope associated with h exists and can be expressed as follows:

$$c(h,\theta_0) = \mathbb{E}\left[h(Z_i)m(W_i;\theta_0)\right]^T\mathbb{E}\left[h(Z_i)m(W_i,\theta_0)m(W_i,\theta_0)^T h(Z_i)^T\right]^{-1}\mathbb{E}\left[h(Z_i)m(W_i,\theta_0)\right]$$

Next, we define define $\Sigma(Z_i) = \mathbb{E}\left[m(W_i,\theta_0)m(W_i,\theta_0)^T|Z_i\right]$. For any $h \in \mathcal{H}_0$, the best linear predictor of $\Sigma(Z_i)^{-1/2}m(W_i,\theta_0)$ on $\Sigma(Z_i)^{1/2}h(Z_i)^T$ exists and can be expressed as:

$$BLP((\Sigma(Z_i)^{-1/2}m(W_i,\theta_0))^T|\Sigma(Z_i)^{1/2}h(Z_i)^T) = \mathbb{E}\left[h(Z_i)m(W_i,\theta_0)\right]^T \mathbb{E}\left[h(Z_i)\Sigma(Z_i)h(Z_i)^T\right]^{-1} h(Z_i)\Sigma(Z_i)^{1/2}$$

Thus, one can easily show that the expectation of the squared norm of this projection is equal to:

$$\begin{aligned}\mathbb{E}\left[\left|\left|BLP(\Sigma(Z_i)^{-1/2}m(W_i,\theta_0))^T|\Sigma(Z_i)^{1/2}h(Z_i)^T)\right|\right|_2^2\right] =& \mathbb{E}\left[h(Z_i)m(W_i;\theta_0)\right]^T \mathbb{E}\left[h(Z_i)\Sigma(Z_i)h(Z_i)^T\right]^{-1}\mathbb{E}\left[h(Z_i)m(W_i,\theta_0)\right]\\ =& \mathbb{E}\left[h(Z_i)m(W_i;\theta_0)\right]^T \mathbb{E}\left[h(Z_i)m(W_i,\theta_0)m(W_i,\theta_0)^T h(Z_i)^T\right]^{-1}\\ & \mathbb{E}\left[h(Z_i)m(W_i,\theta_0)\right]\end{aligned}$$

□

B.3.2 A feasible approximation of the MPI when the correction term is not discretizable

Before we turn to the proof of Proposition 1.6, we first prove the following two Lemmas. The 1st lemma establishes the smoothness of ρ^{-1} and the second Lemma shows the invertibility of the Jacobian matrix of ρ with respect to δ.

Lemma 2.2 *ρ^{-1} is $\mathcal{C}^\infty$*

Proof. We know that the demand function ρ is $\mathcal{C}^\infty$ and from Berry (1994), we know ρ is invertible on $\mathbb{R}^J$. Moreover,$\forall \delta \in \mathbb{R}^J$, $\forall\ X_{2t}$, $\forall F$, $\frac{\partial\rho(\delta,X_{2t},F)}{\partial\delta} \neq 0$. 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:

$$\frac{\partial\rho_j\left(\delta_t,X_{2t},F\right)}{\partial\delta_{kt}} = \begin{cases} \int \mathcal{T}_{jt}(v)\left(1-\mathcal{T}_{kt}(v)\right)F(v)dv & if\ \ j=k \\ -\int \mathcal{T}_{jt}(v)\mathcal{T}_{kt}(v)F(v)dv & if\ \ \ j\neq k \end{cases}$$

with $\mathcal{T}_{jt}(v) \equiv \frac{exp\{\delta_{jt}+X_{2jt}^T v\}}{1+\sum_{j'=1}^J exp\{\delta_{j't}+X_{2j't}^T v\}}$.

One can easily check that $\frac{\partial \rho}{\partial \delta}$ is strictly diagonally dominant. Indeed for each row j:

$$\left|\frac{\partial \rho_j\left(\delta_t, X_{2t}, F\right)}{\partial \delta_{kt}}\right| - \sum_{k \neq j}\left|\frac{\partial \rho_j\left(\delta_t, X_{2t}, F\right)}{\partial \delta_{kt}}\right| = \int \mathcal{T}_{jt}(v) \underbrace{\left(1 - \sum_{k=1}^{J} \mathcal{T}_{kt}(v)\right)}_{>0} F(v)dv > 0.$$

$\square$

Proof of Proposition 1.6.

The proof is in three steps.

- **Step 1:** First, we define $Y_t^0 = \rho(\delta_t, X_{2t}, F_0)$ *a.s.* 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 $\rho^{-1}(Y_t^0, X_{2t}, F_0)$ around Y_t can be expressed as:

$$\rho^{-1}(Y_t^0, X_{2t}, F_0) = \rho^{-1}(Y_t, X_{2t}, F_0) + \left.\frac{\partial \rho^{-1}(Y_t, X_{2t}, F_0)}{\partial y}\right|_{y=Y_t} (Y_t^0 - Y_t) + o\big(||Y_t^0 - Y_t||\big)$$
$$\delta_t = \rho^{-1}(Y_t, X_{2t}, F_0) + \left.\frac{\partial \rho^{-1}(Y_t, X_{2t}, F_0)}{\partial y}\right|_{y=Y_t} (Y_t^0 - Y_t) + o\big(||Y_t^0 - Y_t||\big).$$

We now derive an expression for the first derivative of the inverse function. We make use of lemma 2.3: for any $\delta \in \mathbb{R}^J$, $\frac{\partial \rho(\delta, X_{2t}, F_0)}{\partial \delta}$ is invertible.

$$\frac{\partial \rho(\rho^{-1}(Y_t, X_{2t}, F_0), X_{2t}, F_0)}{\partial y} = I_J \iff \frac{\partial \rho^{-1}(Y_t, X_{2t}, F_0)}{\partial y}\left(\frac{\partial \rho(\rho^{-1}(Y_t, X_{2t}, F_0), X_{2t}, F_0)}{\partial \rho^{-1}(Y_t, X_{2t}, F_0)}\right) = I_J$$
$$\iff \frac{\partial \rho^{-1}(Y_t, X_{2t}, F_0)}{\partial y} = \left(\frac{\partial \rho(\delta_t^0, X_{2t}, F_0)}{\partial \delta}\right)^{-1}$$

with $\delta_t^0 = \rho^{-1}(Y_t, X_{2t}, F_0)$. Consequently,

$$\rho^{-1}(Y_t, X_{2t}, F_0) - \rho^{-1}(Y_t, X_{2t}, F) = -\left(\frac{\partial \rho(\delta_t^0, X_{2t}, F_0)}{\partial \delta}\right)^{-1} (Y_t^0 - Y_t) + o\big(||Y_t^0 - Y_t||\big) \quad a.s., \tag{B.14}$$

with $\delta_t^0 = \rho^{-1}(Y_t, X_{2t}, F_0)$.

- **Step 2:** now let us show that there exists a constant M such that $||Y_t^0 - Y_t||_1 \leq M\tau(F_0 - F_a)$ with $\tau(F_0 - F_a) = \int_{\mathbb{R}^{K_2}} |dF_0(v) - dF_a(v)|$. Let us observe that we derive the results with the L_1 norm. However, this is without loss of generality as norms are equivalent in a finite vectorial space. By definition:

$$Y_t^0 - Y_t = \int_{\mathbb{R}^{K_2}} \frac{\exp\{\delta_t + X_{2t}v\}}{1 + \sum_{k=1}^J \exp\left\{\delta_{kt} + X_{2jk}^T v\right\}}(dF_0(v) - dF_a(v)).$$

Taking the L_1 norm of this vector:

$$\begin{aligned}
||Y_t^0 - Y_t||_1 &= \sum_{j=1}^J \left| \int_{\mathbb{R}^{K_2}} \frac{\exp\{\delta_{jt} + X_{2jt}^T v)}{1 + \sum_{k=1}^J \exp\left\{\delta_{kt} + X_{2jk}^T v\right\}}(dF_0(v) - dF_a(v))\right|, \\
&\leq \sum_{j=1}^J \int_{\mathbb{R}^{K_2}} \underbrace{\left| \frac{\exp\{\delta_{jt} + X_{2jt}^T v)}{1 + \sum_{k=1}^J \exp\left\{\delta_{kt} + X_{2jk}^T v\right\}}\right|}_{\leq 1} |dF_0(v) - dF_a(v)|, \\
&\leq J \int_{\mathbb{R}^{K_2}} |dF_0(v) - dF_a(v)| = J\tau(F_0 - F_a).
\end{aligned}$$

This proves the statement. As a consequence, we have: $||Y_t^0 - Y_t||_1 = O(\tau(F_0 - F_a))$ and $o(||Y_t^0 - Y_t||) = o(\tau(F_0 - F_a))$.

- **Step 3:** We cannot directly use $Y_t^0 - Y_t$ because its expression depends on δ_t which we do not know under misspecification. 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 the equation B.14, we have that:

$$||\delta_t - \delta_t^0|| = ||\delta_t - \rho^{-1}(Y_t, X_{2t}, F_0)|| = O(||Y_t^0 - Y_t||) = O(\tau(F_0 - F_a)).$$

Now let us take the Taylor expansion of $Y_t^0 - Y_t$ around δ_t^0:

$$\begin{aligned}
Y_t^0 - Y_t &= \int_{\mathbb{R}^{K_2}} \frac{\exp\{\delta_t^0 + X_{2t}v\}}{1 + \sum_{k=1}^J \exp\left\{\delta_t^0 + X_{2jk}^T v\right\}}(dF_0(v) - dF_a(v)) \\
&+ \underbrace{\int_{\mathbb{R}^{K_2}} \frac{\partial}{\partial \delta'}\left\{\frac{\exp\{\delta_t^0 + X_{2t}v\}}{1 + \sum_{k=1}^J \exp\left\{\delta_{kt}^0 + X_{2jk}^T v\right\}}\right\}(\delta_t - \delta_t^0)(dF_0(v) - dF_a(v))}_{B} + o\big(||\delta_t - \delta_t^0||\big).
\end{aligned}$$

From what precedes, we know that $o(||\delta_t - \delta_t^0||) = o(\tau(F_0 - F_a))$. Now, let us show that term B in the previous expansion is also $o(\tau(F_0 - F_a))$. Again taking the L_1 norm:

$$\begin{aligned}
||B||_1 &= \sum_{j=1}^{J} \left| \sum_{l=1}^{J} \int_{\mathbb{R}^{K_2}} \frac{\partial}{\partial \delta_l} \left\{ \frac{\exp\{\delta_{jt}^0 + X_{2jt}^T v\}}{1 + \sum_{k=1}^{J} \exp\left\{\tilde{\delta}_{kt} + X_{2jk}^T v\right\}} \right\} (\delta_{lt} - \delta_{lt}^0)(dF_0(v) - dF_a(v)) \right| \\
&\leq \sum_{j=1}^{J} \sum_{l=1}^{J} \int_{\mathbb{R}^{K_2}} \underbrace{\left| \frac{\partial}{\partial \delta_l} \left\{ \frac{\exp\{\delta_{jt}^0 + X_{2jt}^T v\}}{1 + \sum_{k=1}^{J} \exp\left\{\delta_{kt}^0 + x_{2jk}^T v\right\}} \right\} \right|}_{\leq 1} |\delta_{lt} - \tilde{\delta}_{lt}||F_0(v) - F_a(v)| dv \\
&\leq J^2 \tau(F_0 - F_a) O(\tau(F_0 - F_a)) = O(\tau(F_0 - F_a)^2) = o(\tau(F_0 - F_a)) \,.
\end{aligned}$$

Thus, $||B||_1 = o(\tau(F_0 - F_a))$ and by combining all the results together, we get the final result.

- **Conclusion:** a first-order expansion of $\Delta_{0,a}^{m_3}(W_t)$ around F_0 is expressed as:

$$\Delta_{0,a}^{m_3}(Y_t, X_t) = X_{1t}(\beta_0 - \beta_a) + \left(\frac{\partial \rho(\delta_t^0, X_{2t}, F_0)}{\partial \delta} \right)^{-1} \int_{\mathcal{V}} \left[\frac{\exp\{\delta_t^0 + X_{2t} v\}}{1 + \sum_{k=1}^{J} \exp\{\delta_{kt}^0 + X_{2kt}' v\}} - \rho(\delta_t^0, X_{2t}, F_0) \right] dF_a(v) + \mathcal{R}_0,$$

with $\delta_t^0 = \rho^{-1}(Y_t, X_{2t}, F_0)$ and $\mathcal{R}_0 = o\big(\int_{\mathcal{V}} |dF_a(v) - dF_0(v)|\big)$.

□

B.4 Model selection

Proof of Proposition 5.1.

We assume that for $k \in \{1, 2\}$, $h_k \in \mathcal{H}_k^{\perp}$, $\hat{\phi}_k$ satisfies Assumption 2, m satisfies Assumption 3, and $\sigma_{RV}^2 > 0$.

First, let us show that

$$\sqrt{N} \left(\begin{pmatrix} \hat{C}_1 \\ \hat{C}_2 \end{pmatrix} - \begin{pmatrix} C_1 \\ C_2 \end{pmatrix} \right) \xrightarrow[N \to +\infty]{d} \mathcal{N}(0, \tilde{R} \Omega \tilde{R}^T) \quad (*)$$

with:

$$\tilde{R} = \begin{pmatrix} \mathbb{E}[h_1(Z_i)m(W_i,\phi_1)]^T\Omega_{1,1}^{-1} & 0 \\ 0 & \mathbb{E}[h_2(Z_i)m(W_i,\phi_2)]^T\Omega_{2,2}^{-1} \end{pmatrix} \quad and \quad \Omega = \begin{pmatrix} \Omega_{1,1} & \Omega_{1,2} \\ \Omega_{2,1} & \Omega_{2,2} \end{pmatrix}$$

where $\Omega_{k,j} = \mathbb{E}[h_k(Z_i)m(W_i,\phi_k)m(W_i,\phi_j)^T h_j(Z_i)^T]$ and ϕ_k the (pseudo)-true value under specification $\mathcal{F}_k$ (as defined in Assumption 2). By definition, we have:

$$\sqrt{N}\begin{pmatrix} \hat{C}_1 \\ \hat{C}_2 \end{pmatrix} = \hat{\tilde{R}}\sqrt{N}\begin{pmatrix} \frac{1}{N}\sum_{i=1}^N h_1(Z_i)m(W_i,\hat{\phi}_1) \\ \frac{1}{N}\sum_{i=1}^N h_2(Z_i)m(W_i,\hat{\phi}_2) \end{pmatrix}$$

with

$$\hat{\tilde{R}} = \begin{pmatrix} \left(\frac{1}{N}\sum_{i=1}^N h_1(Z_i)m(W_i,\hat{\phi}_1)\right)^T \hat{\Omega}_{1,1}^{-1} & 0 \\ 0 & \left(\frac{1}{N}\sum_{i=1}^N h_2(Z_i)m(W_i,\hat{\phi}_2)\right)^T \hat{\Omega}_{2,2}^{-1} \\ . & \end{pmatrix}$$

To prove this result, we first show that, under the set of assumptions stated in the formal proposition, $\hat{\tilde{R}} \xrightarrow[N\to+\infty]{P} \tilde{R}$ (**).

- From the proof of proposition 3.4 part (ii), we have for $k \in \{1,2\}$:

$$\frac{1}{N}\sum_{i=1}^N h(Z_i)m(W_i,\hat{\phi}) \xrightarrow[N\to+\infty]{P} \mathbb{E}[h_k(Z_i)m(W_i,\phi_k)]$$

- From the proof of proposition 3.4 part (iii), we have for $k \in \{1,2\}$: $\hat{\Omega}_{k,k} \xrightarrow[N\to+\infty]{P} \Omega_{k,k}$. Statement (**) follows from the continuous mapping theorem.

Next, we want to show that:

$$A_n = \sqrt{N}\left(\begin{pmatrix} \frac{1}{N}\sum_{i=1}^N h_1(Z_i)m(W_i,\hat{\phi}_1) \\ \frac{1}{N}\sum_{i=1}^N h_2(Z_i)m(W_i,\hat{\phi}_2) \end{pmatrix} - \begin{pmatrix} \mathbb{E}\left[h_1(Z_i)m(W_i,\phi_1)\right] \\ \mathbb{E}\left[h_2(Z_i)m(W_i,\phi_2)\right] \end{pmatrix}\right) \xrightarrow[N\to+\infty]{d} \mathcal{N}(0,\Omega).$$

We concatenate the moment as follows. We define the following matrices:

$$\tilde{h}_{1,2}(Z_i) \equiv \begin{pmatrix} h_1(Z_i) & 0_{L_1\times p} \\ 0_{L_2\times p} & h_2(Z_i) \end{pmatrix} \qquad \text{and} \qquad \tilde{m}_{1,2}(W_i,\tilde{\phi}_1,\tilde{\phi}_2) \equiv \begin{pmatrix} m(W_i,\tilde{\phi}_1) \\ m(W_i,\tilde{\phi}_2) \end{pmatrix}$$

By construction,

$$A_n = \sqrt{N}\left(\frac{1}{N}\sum_{i=1}^{N}\tilde{h}_{1,2}(Z_i)\tilde{m}_{1,2}(W_i,\hat{\phi}_1,\hat{\phi}_2) - \mathbb{E}\left[\tilde{h}_{1,2}(Z_i)\tilde{m}_{1,2}(W_i,\phi_1,\phi_2)\right]\right)$$

Now, we can easily show that $\mathbb{E}\left[\tilde{h}_{1,2}(Z_i)\tilde{m}_{1,2}(W_i,\phi_1,\phi_2)\right]$ is Neyman orthogonal with respect to $\phi_{1,2} = (\phi_1^T,\phi_2^T)^T$:

$$\mathbb{E}\left[\tilde{h}_{1,2}(Z_i)\frac{\partial \tilde{m}_{1,2}(W_i,\phi_1,\phi_2)}{\partial \phi_{1,2}^T}\right] = \begin{pmatrix} \mathbb{E}\left[h_1(Z_i)\frac{\partial m(W_i,\phi_1)}{\partial \phi_1^T}\right] & 0_{L_1\times \dim(\phi_2)} \\ 0_{L_2\times \dim(\phi_1)} & \mathbb{E}\left[h_2(Z_i)\frac{\partial m(W_i,\phi_2)}{\partial \phi_2^T}\right] \end{pmatrix} = 0$$

where the second inequality comes from the fact that for $k \in \{1,2\}$, $h_k \in \mathcal{H}_k^{\perp}$.We assume the same regularity assumptions as in proposition 3.4. Thus, following the same steps as in proof of proposition 3.4 part (i), we have $A_n \xrightarrow[N\to+\infty]{d} \mathcal{N}(0,\Omega)$. Finally, Slutsky's lemma yields statement $(*)$.

Second, by combining $(*)$ and Slutsky's lemma, we have: $\sqrt{N}(\hat{C}_1 - \hat{C}_2 - (C_1 - C_2)) \xrightarrow[N\to+\infty]{d} \mathcal{N}(0, R\Omega R^T)$ with $R = (1\ \ -1)\tilde{R}$.

Third, by a same line of argument as in the proof of proposition 3.4 part (iii), we have $\hat{\Omega}_{k,l} \xrightarrow[N\to+\infty]{P} \Omega_{k,l}$ for $(k,l) \in \{1,2\}^2$. Thus, $\hat{\Omega} \xrightarrow[N\to+\infty]{P} \Omega$. From $(**)$ we have: $\hat{\tilde{R}} \xrightarrow[N\to+\infty]{P} \tilde{R}$. Hence, by the CMT, we have $\hat{\sigma}_{RV} \xrightarrow[N\to+\infty]{P} \sigma_{RV}$

Finally, as we assume $\sigma_{RV} > 0$, by an application of Slutsky's lemma, we recover the intended asymptotic distribution:

- Under $H_0^{RV}: C_1 = C_2 \qquad T_N^{RV} \xrightarrow[N\to+\infty]{d} \mathcal{N}(0,1)$,

- Under $H_1^{RV}: C_1 < C_2 \qquad T_N^{RV} \xrightarrow[N\to+\infty]{d} -\infty$,

- Under $H_2^{RV} : C_1 > C_2 \qquad T_N^{RV} \xrightarrow[N \to +\infty]{d} +\infty.$

$\square$

C Monte Carlo experiments

C.1 Finite sample performance of our test in the mixed logit model: additional results

Simulation design. The exogenous products attributes x_a, x_b, x_c and x_d are distributed according to the following joint normal distribution:

$\forall m \in \{a, b, c, d\}$, $x_{mij} = \frac{1}{\sqrt{2}}(\tilde{x}_{mij} + \bar{x}_j)$ with $\bar{x}_j \sim \mathcal{N}(0, 1)$ and

$$\begin{bmatrix} \tilde{x}_{aij} \\ \tilde{x}_{bij} \\ \tilde{x}_{cij} \\ \tilde{x}_{dij} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & -0.8 & 0.3 & -0.5 \\ -0.8 & 1 & 0.3 & 0.6 \\ 0.3 & 0.3 & 1 & 0.2 \\ -0.5 & 0.6 & 0.2 & 1 \end{bmatrix} \right).$$

Power against Gamma alternatives. The simulation design is the same as the one presented in Section 6.1. We continue to test the null hypothesis $H_0 : F \in \mathcal{F}_0$, where $\mathcal{F}_0$ is the family of normal distributions. Our objective is now to assess the finite-sample power of our test against alternatives from the Gamma family. To this end, we generate data with the random coefficient distributed according to the Gamma distributions described below and which we plot in Figure 2. The sets of instruments is identical to the ones used to evaluate the power of our specification test against Gaussian mixtures. We report the results in Table 7. The trends that we observed in the Gaussian mixture case also appear here. Once again, the MPI generates the most powerful test, followed by the sets FMPI (6) and FMPI (10) and the polynomial instruments. Overall, it seems that it is harder to reject the normality assumptions against the gamma alternatives considered here than against the Gaussian mixtures considered in Section C.1. As a result, the power differences at smaller sample sizes (N = 500) are less pronounced. Additionally, the limited power of the infeasible MPI under the standard variance correction procedure becomes more pronounced in small sample sizes.

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

Figure 2: Densities, Gamma alternatives

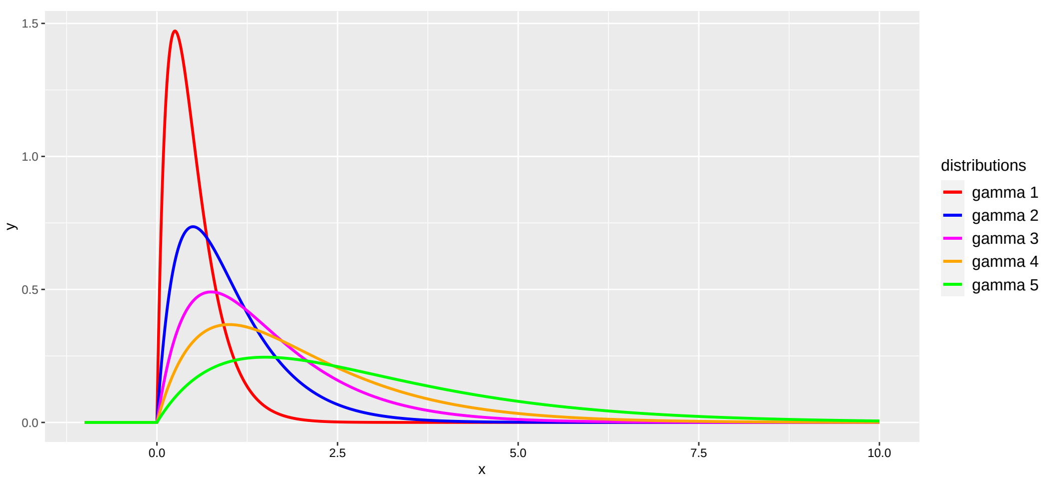


Table 7: Empirical power, Gamma alternatives (1000 replications)

Number of individuals	N=500							
2-step adjustment	orthogonalization				std. variance correction			
Instruments	MPI	FMPI (6)	FMPI (10)	Pol	MPI	FMPI (6)	FMPI (10)	Pol
Gamma 1	0.063	0.046	0.038	0.081	0	0.016	0.016	0.051
Gamma 2	0.057	0.041	0.046	0.058	0	0.009	0.015	0.035
Gamma 3	0.114	0.054	0.04	0.06	0.001	0.017	0.015	0.036
Gamma 4	0.113	0.061	0.052	0.067	0	0.018	0.021	0.041
Gamma 5	0.183	0.099	0.08	0.062	0.005	0.027	0.032	0.039
Number of individuals	N=1000							
2-step adjustment	orthogonalization				std. variance correction			
Instruments	MPI	FMPI (6)	FMPI (10)	Pol	MPI	FMPI (6)	FMPI (10)	Pol
Gamma 1	0.058	0.053	0.042	0.077	0	0.018	0.013	0.047
Gamma 2	0.099	0.057	0.045	0.067	0	0.012	0.019	0.039
Gamma 3	0.162	0.094	0.065	0.075	0	0.023	0.026	0.045
Gamma 4	0.228	0.127	0.106	0.087	0.004	0.039	0.034	0.041
Gamma 5	0.337	0.192	0.147	0.083	0.041	0.072	0.065	0.039
Number of individuals	N=5000							
2-step adjustment	orthogonalization				std. variance correction			
Instruments	MPI	FMPI (6)	FMPI (10)	Pol	MPI	FMPI (6)	FMPI (10)	Pol
Gamma 1	0.065	0.052	0.056	0.056	0	0.01	0.012	0.038
Gamma 2	0.245	0.151	0.115	0.063	0.002	0.054	0.054	0.04
Gamma 3	0.559	0.346	0.295	0.07	0.167	0.199	0.17	0.034
Gamma 4	0.747	0.537	0.463	0.073	0.337	0.338	0.316	0.035
Gamma 5	0.917	0.841	0.774	0.085	0.703	0.616	0.645	0.038

C.2 Finite sample performance of our test in the BLP demand model

In this section, we conduct Monte Carlo simulations to evaluate the empirical size and power of our test in the BLP demand model. We use different sets of testing instruments including the infeasible most powerful instrument (MPI) as a benchmark, a feasible MPI that uses the local approximation of the MPI derived in Proposition 1.6 and a polynomial transformation of the exogenous variables, using Hermite polynomials with decaying tails as in the mixed logit case. For the construction of the infeasible MPI and the feasible approximation of the MPI, we ignore the conditional variance term. We provide more details on the construction of each set of instruments in Appendix C.6. As in the mixed logit case, we consider two procedures to adjust the test for parameter uncertainty: the orthogonalization procedure we developed in Section A.4 and a classical variance correction approach described in Appendix C.5, which serves as a benchmark. Here, we follow the alternative orthogonalization procedure because it is extremely challenging to derive $\mathbb{E}\left[\frac{\partial m(W_t,\phi_0)}{\partial \phi^T}\middle| Z_t\right]$ in the BLP demand model.

The estimator in the BLP demand model is computed using the nested fixed-point SMM procedure detailed in Appendix C.4.2. For estimation, we use the two sets of 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 from Chamberlain (1987). In both cases, the model is over-identified, allowing us to compare the performance of our test with the standard Sargan-Hansen J test (which uses the same instruments for both testing and estimation). We ensure that the number of tested restrictions is similar across the different sets of instruments to allow a fair comparison.

The simulation design closely follows the simulation design used in Dubé et al. (2012), Reynaert and Verboven (2014). It also shares many common features with the one used in the mixed logit example. 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. Only x_c exhibits heterogeneity in preferences, 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:

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

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

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

The 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. The hypothesis tested 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%.

Empirical size. Below in Table 8, we report the empirical sizes of the test with the different sets of instruments described previously for the different sample sizes $T \in \{50, 100, 200\}$ and for different distributions of the RC such that $v_i \sim F \in \mathcal{F}_0$. In the Table below, J corresponds to the usual Sargan-Hansen J test, $FMPI$ corresponds to the feasible MPI or equivalently the interval instruments (with 8 instruments), Pol corresponds to the polynomial transformation of the exogenous variables. Under H_0, the MPI is difficult to define as all the valid instruments have a slope that is equal to 0. Therefore, we do not report it. The letter o or c in parenthesis indicates respectively that the instruments have been orthogonalized ot that the variance has been corrected to account for parameter uncertainty.

Table 8: Empirical size for nominal size 5% (1000 replications)

Estimation instruments	"differentiation"											
Number of markets	T=50				T=100				T=200			
Instruments	J	FMPI (o.)	FMPI (c.)	Pol (o.)	J	FMPI (o.)	FMPI (c.)	Pol (o.)	J	FMPI (o.)	MPI (c.)	Pol (o.)
$v_i \sim \mathcal{N}(-1, 0.5^2)$	0.025	0.016	0.012	0.011	0.039	0.03	0.013	0.017	0.042	0.028	0.021	0.019
$v_i \sim \mathcal{N}(0, 0.75^2)$	0.025	0.016	0.009	0.015	0.032	0.015	0.008	0.015	0.048	0.026	0.011	0.021
$v_i \sim \mathcal{N}(1, 1^2)$	0.025	0.014	0.006	0.015	0.035	0.015	0.01	0.015	0.046	0.028	0.018	0.021
$v_i \sim \mathcal{N}(2, 2^2)$	0.025	0.01	0.008	0.016	0.03	0.013	0.005	0.015	0.035	0.011	0.006	0.02
$v_i \sim \mathcal{N}(3, 3^2)$	0.022	0.01	0.006	0.015	0.032	0.009	0.005	0.015	0.035	0.02	0.012	0.018
Estimation instruments	"optimal"											
Number of markets	T=50				T=100				T=200			
Instruments	J	FMPI (o.)	FMPI (c.)	Pol (o.)	J	FMPI (o.)	FMPI (c.)	Pol (o.)	J	FMPI (o.)	FMPI (c.)	Pol (o.)
$v_i \sim \mathcal{N}(-1, 0.5^2)$	0.039	0.012	0.005	0.013	0.052	0.029	0.01	0.016	0.05	0.028	0.011	0.016
$v_i \sim \mathcal{N}(0, 0.75^2)$	0.038	0.007	0.008	0.016	0.044	0.018	0.01	0.017	0.042	0.025	0.01	0.021
$v_i \sim \mathcal{N}(1, 1^2)$	0.038	0.015	0.005	0.015	0.046	0.017	0.008	0.018	0.036	0.023	0.015	0.022
$v_i \sim \mathcal{N}(2, 2^2)$	0.039	0.007	0.005	0.016	0.045	0.016	0.014	0.016	0.047	0.009	0.008	0.019
$v_i \sim \mathcal{N}(3, 3^2)$	0.034	0.006	0.007	0.016	0.045	0.006	0.005	0.015	0.045	0.02	0.012	0.015

We observe that our test provides a good size control across the different sets of instruments and for both procedures used to adjust for parameter uncertainty. In particular, with a moderate sample size ($T = 50, J = 12$), all the tests are slightly undersized. However, as the sample size increases, the empirical size converges to the nominal level for most configurations, providing strong evidence of the test's asymptotic validity. Additionally, it is noteworthy that, for most configurations, the alternative orthogonalization procedure yields an empirical size that more closely aligns with the nominal size compared to the classical variance correction.

Empirical power. 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 Gaussians or a Gamma distribution. The sets of instruments remain unchanged, with the addition of the infeasible MPI. As the dimension of the infeasible MPI is one and exceeds the number of non-linear parameters involved in the minimization, we cannot apply the alternative orthogonalization strategy developed in Section A.4 and we have to use the classical variance correction for this specific instrument. 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 main takeaway from our results is that, as in the mixed logit case, the interval instruments substantially outperform the polynomial instruments and the traditional Sargan-Hansen J-test in terms of power against all the alternative distributions considered in our simulations and their performance is relatively close to that of the infeasible MPI.

Power against Gaussian mixture alternatives. We simulate data with the random coefficients distributed according to the Gaussian mixtures used in the mixed logit case (see Figure 3 below). We report the results in Table 9. As anticipated, the MPI generates the most powerful test, followed by the feasible MPIs—FMPI (o) and FMPI (c)—the J test, and the polynomial instruments. The feasible tests perform slightly better under the orthogonalization procedure than under the conventional variance correction. The difference in terms of power between the interval instruments (or equivalently the FMPIs) and the other instruments—the J test and polynomial instruments—is striking. These observations hold whether we consider the differentiation instruments or the "optimal instruments" as estimation instruments.

Figure 3: Densities, Gaussian mixture alternatives

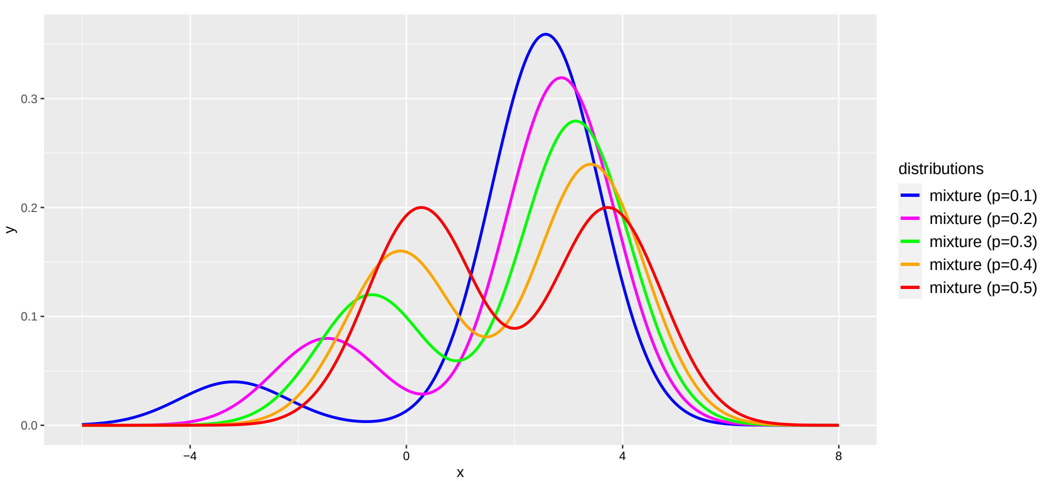


Table 9: Empirical power, Gaussian mixture alternatives (1000 replications)

Estimation instruments	“differentiation”														
Number of markets	T=50					T=100					T=200				
Instruments	MPI(c.)	J	FMPI (o.)	FMPI (c.)	Pol (o.)	MPI(c.)	J	FMPI (o.)	FMPI (c.)	Pol (o.)	MPI(c.)	J	FMPI (o.)	MPI (c.)	Pol (o.)
Mixture 1	0.989	0.111	0.938	0.896	0.014	1	0.306	1	1	0.023	1	0.746	1	1	0.024
Mixture 2	0.998	0.158	0.967	0.949	0.011	1	0.429	1	1	0.024	1	0.883	1	1	0.025
Mixture 3	0.998	0.166	0.933	0.893	0.01	1	0.444	1	1	0.021	1	0.887	1	1	0.023
Mixture 4	0.961	0.147	0.796	0.735	0.009	1	0.409	0.997	0.999	0.018	1	0.865	1	1	0.022
Mixture 5	0.757	0.116	0.471	0.413	0.009	0.981	0.347	0.946	0.935	0.02	1	0.753	1	1	0.025

Estimation instruments	“optimal”														
Number of markets	T=50					T=100					T=200				
Instruments	MPI(c.)	J	FMPI (o.)	FMPI (c.)	Pol (o.)	MPI(c.)	J	FMPI (o.)	FMPI (c.)	Pol (o.)	MPI(c.)	J	FMPI (o.)	MPI (c.)	Pol (o.)
Mixture 1	0.992	0.456	0.944	0.884	0.013	1	0.934	1	1	0.022	1	1	1	1	0.024
Mixture 2	0.999	0.338	0.978	0.953	0.012	1	0.818	1	1	0.022	1	0.997	1	1	0.026
Mixture 3	0.997	0.195	0.944	0.907	0.015	1	0.558	1	1	0.023	1	0.965	1	1	0.027
Mixture 4	0.965	0.078	0.806	0.724	0.015	1	0.239	0.999	0.997	0.022	1	0.655	1	1	0.024
Mixture 5	0.742	0.034	0.5	0.408	0.014	0.982	0.056	0.957	0.929	0.021	1	0.124	1	1	0.025

Power against Gamma alternatives. We simulate data with the random coefficients distributed according to the Gamma distribution used in the mixed logit simulations. We report the results in Table 10. These results follow the same pattern as in the Gaussian mixture case. Once again, the MPI generates the most powerful test, followed by the feasible MPIs—FMPI (o) and FMPI (c)—the J test, and the polynomial instruments. It appears that all the instruments have limited power against the first Gamma distribution.

Table 10: Empirical power, Gamma alternatives (1000 replications)

Estimation instruments	"differentiation"														
Number of markets	T=50					T=100					T=200				
Instruments	MPI(c.)	J	FMPI (o.)	FMPI (c.)	Pol (o.)	MPI(c.)	J	FMPI (o.)	FMPI (c.)	Pol (o.)	MPI(c.)	J	FMPI (o.)	MPI (c.)	Pol (o.)
Gamma 1	0.001	0.029	0.014	0.007	0.013	0.002	0.047	0.037	0.015	0.018	0.001	0.063	0.09	0.047	0.022
Gamma 2	0.76	0.118	0.525	0.375	0.014	0.987	0.317	0.953	0.899	0.018	0.999	0.732	1	1	0.019
Gamma 3	0.969	0.159	0.878	0.787	0.014	1	0.456	0.999	0.996	0.018	1	0.868	1	1	0.025
Gamma 4	0.995	0.161	0.921	0.852	0.014	1	0.495	1	0.998	0.015	1	0.874	1	1	0.027
Gamma 5	0.996	0.217	0.903	0.891	0.02	1	0.6	0.999	0.999	0.02	1	0.924	1	1	0.027

Estimation instruments	"optimal"														
Number of markets	T=50					T=100					T=200				
Instruments	MPI(c.)	J	FMPI (o.)	FMPI (c.)	Pol (o.)	MPI(c.)	J	FMPI (o.)	FMPI (c.)	Pol (o.)	MPI(c.)	J	FMPI (o.)	MPI (c.)	Pol (o.)
Gamma 1	0	0.036	0.013	0.005	0.013	0	0.039	0.044	0.018	0.015	0	0.052	0.096	0.05	0.022
Gamma 2	0.763	0.041	0.53	0.38	0.017	0.985	0.056	0.958	0.913	0.015	0.998	0.053	1	1	0.018
Gamma 3	0.977	0.039	0.862	0.764	0.018	1	0.077	0.999	0.996	0.015	1	0.115	1	1	0.024
Gamma 4	0.996	0.069	0.905	0.826	0.015	1	0.157	1	0.999	0.017	1	0.346	1	1	0.027
Gamma 5	0.997	0.159	0.87	0.853	0.017	1	0.428	0.997	0.998	0.022	1	0.858	1	1	0.026

C.3 Finite-sample performance of our model selection procedure

In this section, we assess the finite-sample performance of our model selection procedure for the mixed logit model (Example 2). To conduct this evaluation, we simulate data based on various parametric distributions for the distribution of RCs, including a normal distribution, a mixture of normals distribution, a uniform distribution and a gamma distribution. Subsequently, we estimate the model under several competing parametric specifications on the distribution of RCs. These specifications include a degenerate random coefficient (i.e. no heterogeneity in preferences), a normal distribution, a mixture of normal distribution (with 2 components), a uniform distribution, a log-normal distribution, and a gamma distribution.

For each candidate specification, we write a specific code on R to estimate the model. Our estimation procedure follows the guidelines outlined in Appendix C.4.1. We then apply the model selection procedure proposed in Section 5 to discriminate between two specifications. The slope is approximated using interval instruments, with eight instruments for each specification. We benchmark our procedure against the likelihood ratio test introduced in Vuong (1989), which is applicable with models estimated via maximum likelihood, including the mixed logit model.

In accordance with our theoretical results, we assume that, for all specifications, the pseudo-true values ϕ_k lie in the interior of the parameter space. This assumption is a key regularity condition to derive the asymptotic properties of our test as well as those of the likelihood ratio test introduced in Vuong (1989). This assumption also ensures that the specification with a degenerate RC does not overlap with the other specifications as it rules out the case where the dipersion parameter σ is zero. However, this assumption also makes the LR test redundant as it implies that the dispersion parameter $\sigma = 0$ does not maximize the population log-likelihood. Consequently, we do not report the likelihood ratio test for the degenerate RC case against other distributions. Lastly, since the normal distribution is nested within the mixture of Gaussians specification, we follow the sequential procedure outlined in Appendix A.7 to pre-test whether the means and variances of the mixture components are equal.

Finally, in the case of the Gaussian mixture, rather than jointly maximizing the log-likelihood over both the mixing probabilities and the regular parameters, we perform a grid-based profiled maximum likelihood estimation. This method helps mitigate several numerical and theoretical challenges associated with estimating the mixing probabilities, including numerical instability, multiple local minima, slow convergence of the EM algorithm, and degenerate solutions. The grid is designed to ensure that no mixing probability is exactly zero, thereby preventing degenerate cases where the mean and variance corresponding to a zero-mixing probability cannot be estimated. Further details on this approach are provided in Appendix C.4.1.

The simulation design is the same as the one used to assess the finite-sample performance of our test in the mixed logit model. The sample size is $N = 1000$. The RCs are distributed according to the following distributions: $v_{1i} \sim \mathcal{N}(0, 2)$, $v_{2i} \sim D_i\mathcal{N}(-2, 2.5)+(1-D_i)\mathcal{N}(2, 1.5)$ with $D_i \sim \text{Bernoulli}(0.5)$, $v_{3i} \sim \text{Uniform}(-3, 3)$, $v_{4i} \sim \text{Log-Normal}(0, 1)$. We report the results in the Table 11. The reported values represent the proportion of times each specification is selected, or when no specification is chosen, which is the case when H_0 is not rejected.

Table 11: Model selection for nominal size 5% (1000 replications)

True specification		Normal						Gaussian mixture					
Model selection procedure		Slope			Likelihood ratio			Slope			Likelihood ratio		
Specification under H_1	Specification under H_2	H_0	H_1	H_2	H_0	H_1	H_2	H_0	H_1	H_2	H_0	H_1	H_2
Degenerate RC	Normal	0.004	0	0.996	-	-	-	0	0	1	-	-	-
Degenerate RC	Gaussian mixture	0.011	0	0.989	-	-	-	0	0	1	-	-	-
Degenerate RC	Uniform	0.011	0	0.989	-	-	-	0	0	1	-	-	-
Degenerate RC	Log-normal	0.011	0.001	0.988	-	-	-	0.008	0.003	0.989	-	-	-
Degenerate RC	Gamma	0.009	0	0.991	-	-	-	0	0	1	-	-	-
Normal	Gaussian mixture	0.927	0.039	0.034	1	0	0	0.851	0.045	0.104	1	0	0
Normal	Uniform	0.718	0.247	0.035	1	0	0	0.75	0.118	0.132	1	0	0
Normal	Log-normal	0.409	0.585	0.006	0.999	0.001	0	0	1	0	0.183	0.817	0
Normal	Gamma	0.502	0.49	0.008	1	0	0	0.006	0.994	0	0.182	0.818	0
Gaussian mixture	Uniform	0.836	0.127	0.037	0.982	0.018	0	0.842	0.109	0.049	0.985	0.015	0.001
Gaussian mixture	Log-normal	0.482	0.511	0.007	0.462	0.538	0	0.001	0.999	0	0	1	0
Gaussian mixture	Gamma	0.609	0.377	0.013	1	0	0	0.002	0.998	0	0.003	0.997	0
Uniform	Log-normal	0.473	0.521	0.006	0.496	0.504	0	0	1	0	0	1	0
Uniform	Gamma	0.572	0.416	0.012	1	0	0	0.002	0.998	0	0.004	0.996	0
Log-normal	Gamma	0.622	0.011	0.366	0.999	0	0.001	0.029	0	0.971	0.894	0.099	0.008
True specification		Uniform						Log-normal					
Model selection procedure		Slope			Likelihood ratio			Slope			Likelihood ratio		
Specification under H_1	Specification under H_2	H_0	H_1	H_2	H_0	H_1	H_2	H_0	H_1	H_2	H_0	H_1	H_2
Degenerate RC	Normal	0	0	1	-	-	-	0.077	0	0.923	-	-	-
Degenerate RC	Gaussian mixture	0.001	0	0.999	-	-	-	0.089	0	0.911	-	-	-
Degenerate RC	Uniform	0	0	1	-	-	-	0.082	0.001	0.917	-	-	-
Degenerate RC	Log-normal	0.002	0.007	0.991	-	-	-	0.064	0.002	0.933	-	-	-
Degenerate RC	Gamma	0	0	1	-	-	-	0.063	0	0.937	-	-	-
Normal	Gaussian mixture	0.825	0.034	0.142	1	0	0	0.635	0.005	0.36	1	0	0
Normal	Uniform	0.674	0.039	0.286	1	0	0	0.728	0.046	0.225	1	0	0
Normal	Log-normal	0.27	0.725	0.005	0.992	0.008	0	0.534	0.013	0.453	0.997	0.003	0
Normal	Gamma	0.449	0.534	0.016	1	0	0	0.462	0.007	0.531	1	0	0
Gaussian mixture	Uniform	0.887	0.039	0.075	0.999	0.001	0	0.674	0.313	0.013	0.821	0.179	0
Gaussian mixture	Log-normal	0.189	0.808	0.002	0.124	0.876	0	0.866	0.047	0.087	0.995	0.005	0
Gaussian mixture	Gamma	0.376	0.619	0.004	0.948	0.052	0	0.865	0.074	0.061	1	0	0
Uniform	Log-normal	0.169	0.829	0.002	0.232	0.768	0	0.563	0.014	0.423	0.779	0.003	0.218
Uniform	Gamma	0.303	0.693	0.004	0.958	0.042	0	0.495	0.011	0.494	1	0	0
Log-normal	Gamma	0.306	0.004	0.691	0.992	0	0.008	0.836	0.124	0.041	0.997	0	0.003

Note: we discard simulations where a boundary solution occurs in the first-stage estimation of the mixture specification, typically when one of the variances reaches the lower bound of the parameter space. To ensure at least 1000 valid simulations, we increase the initial number of simulations.

Our results lead to several observations. First, the model selection method developed in this paper either predominantly selects the correct specification over both the alternative specification and the null hypothesis. In some rare instances, it favors the null hypothesis. Both outcomes strongly indicate that our procedure works as intended. The only case that allows us to empirically evaluate the size of our test is when the true specification is normal, and we seek to choose between the normal and Gaussian mixture specifications. In this case, both procedures exhibit rejection rates close to the nominal level. Our test rejects H_0 in 7.3% of cases, indicating a slight over-rejection, whereas the LR test never rejects, suggesting it is conservative. Comparing our test to the likelihood ratio test, we observe that both tests consistently move in the same direction. However, our test appears to be substantially more powerful than the LR test, which is surprising given the optimality properties typically associated with the latter.

C.4 Details on the first-stage estimator

In this section, we provide details on the derivation of the first-stage estimator in both the mixed logit model and the BLP demand model.

C.4.1 Mixed logit

We estimate the parameter $\phi_0 = (\beta_0^T, \lambda_0^T)^T$ using a simulated maximum likelihood procedure under the parametric assumption $H_0 : F \in \mathcal{F}_0$. The unit of observation is the individual i. The SML estimator maximizes the following objective function:

$$\hat{\phi} = \underset{\tilde{\phi} \in \mathbb{R}^{\dim(\phi)}}{\operatorname{argmax}} \sum_{i=1}^{N} L(W_i, \tilde{\phi}) \text{ with } L(W_i, \tilde{\phi}) = \sum_{j=0}^{J} \mathbf{1}\{Y_{ij} = 1\} \log(\rho_j(X_i, \tilde{\phi}))$$

where the demand function is defined as $\rho_j(\cdot)$ is defined as in Example 2.

- **Estimation procedure.** The demand functions must be computed through numerical integration. In this paper, we use Gaussian quadrature rules to approximate the integral. To be more specific, we use Gaussian or Gauss-Hermite rules to approximate the integrals depending on the parametric specification. We use $k = 10$ nodes, which implies that the approximation rule will be exact for polynomial up to total order $2k - 1 = 19$. We assume that the approximation error is sufficiently small so that it can be ignored in the statistical analysis.

- **Parametrization of the model.** In the simulations to assess the performance of our test, we only estimate the model under the assumption that F is a Gaussian. However, in simulations to evaluate the performance of the model selection procedure, the model is estimated under various specifications for F, including uniform, log-normal, Gaussian mixture, and Gamma distributions. Below, we describe the model specification in the case of the Gaussian mixture (which includes the simple Gaussian as a special case). The parametrization of the other distributions is close to the one presented below. For more details on the parametrization of the other distributions, we refer the reader to our codes, which we made publicly available at the following link: https://sites.google.com/view/max-lesellier/home. Hence, we focus on the case where the estimation is performed under $H_0 : \; F \in \mathcal{F}_0$ with $\mathcal{F}_0$ the family of Gaussian mixtures with L components. We want to estimate the pseudo-true value $\phi_0 = (\beta_0, \lambda_0)$ where λ_0 characterizes the mixture up to permutations of indexes: $\lambda_0 = (\mu_{10}, \ldots, \mu_{L0}, \sigma_{10}^2, \ldots, \sigma_{L0}^2, p_{10}, \ldots, p_{L0})$. The c.d.f. of a Gaussian mixture can be expressed as follows:

$$\forall x \in \mathbb{R} \; , \; F_0(x|\tilde{\lambda}) = \sum_{l=1}^{L} p_l G_0(x|\tilde{\lambda}_l) \qquad \sum_{l=1}^{L} p_l = 1 \qquad L \geqslant 1$$

where $G_0(\cdot|\tilde{\lambda})$ is the c.d.f. of a $\mathcal{N}(\tilde{\mu}_l, \tilde{\sigma}_l{}^2)$.

To estimate the mixed logit model with a Gaussian mixture, we adopt the following parametrization of the model: $v_i = \sum_{l=1}^{L} \mathbf{1}\{D_i = l\} v_{il}$ where v_{il} are i.i.d. and have density $G_0(\cdot|\tilde{\lambda}_l)$ known up to $\tilde{\lambda}_l$ for $l = 1, \ldots, L$, and where $(D_i)_{i=1}^n$ are i.i.d categorically distributed with p.m.f. $\mathbb{P}(D_i = l) = \tilde{p}_l$. For ease of exposition, we assume that for each component of the mixture, there is no correlation between the components of v_l. For every individual i and product j, the demand functions are as follows.

$$\begin{aligned}
\rho_j(X_i, \tilde{\phi}) &= \int_{\mathbb{R}} \frac{\exp\{X_{1ij}'\tilde{\beta} + X_{2ij}'v\}}{1 + \sum_{j'=1}^{J} \exp\{X_{1ij}'\tilde{\beta} + X_{2j't}'v\}} dF_0(v|\tilde{\lambda}) dv, \\
&= \sum_{l=1}^{L} \tilde{p}_l \int_{\mathbb{R}} \frac{\exp\{X_{1ij}'\tilde{\beta} + X_{2ij}'v\}}{1 + \sum_{j'=1}^{J} \exp\{X_{1ij'}'\tilde{\beta} + X_{2j't}'v\}} dG_0(v|\tilde{\lambda}_l) dv, \\
&= \sum_{l=1}^{L} \tilde{p}_l \int_{\mathbb{R}} \frac{\exp\{X_{1ij}'\tilde{\beta} + X_{2ij}'(\tilde{\mu}_l + \sqrt{\tilde{\sigma}_l^2} v)\}}{1 + \sum_{j'=1}^{J} \exp\{X_{1ij'}'\tilde{\beta} + X_{2ij'}'(\tilde{\mu}_l + \sqrt{\tilde{\sigma}_l^2} v)\}} d\Phi_0(v) dv
\end{aligned}$$

where Φ_0 is the c.d.f. of the standard $\mathcal{N}(0,1)$. Let us emphasize that following Ketz (2019), the minimization is performed with respect to $\{\sigma_l^2\}_{l=1}^L$ instead and $(\sigma_l)_{l=1}^L$ directly. As highlighted by Ketz (2019), this parametrization of the model ensures that the gradient of both the sample and population log-likelihood remains full rank at the boundary of the parameter space (i.e. when $\tilde{\sigma}_l = 0$, or when $\tilde{\sigma}_l$ is close to zero). We find that this reparametrization significantly improves the convergence rate of numerical maximization while also ensuring the invertibility conditions necessary for our orthogonalization strategy.

- **Numerical maximization.** We now provide details on how we perform the numerical maximization of the log-likelihood.
  - Since the variances associated with each component must be positive, we incorporate this constraint by maximizing the log-likelihood with respect to $\nu = \exp(\tilde{\sigma}^2)$. The invariance principle ensures that $\widehat{\tilde{\sigma}}^2 = \log(\hat{\nu})$.
  - **A grid-based profiled maximum likelihood estimation.** Instead of jointly maximizing the log-likelihood over both the mixing probabilities and the regular parameters, we adopt a grid-based profiling approach. This method helps mitigate several numerical and theoretical challenges associated with estimating the mixing probabilities, including numerical instability, multiple local minima, and degenerate solutions. The grid is designed to ensure that no mixing probability is exactly zero, thereby preventing degenerate cases where the mean and variance corresponding to a zero-mixing probability cannot be estimated. The grid-based profiled maximum likelihood estimation works as follows.

    1. We fix the mixing probability at a given value.
    2. We then maximize the log-likelihood with respect to the remaining parameters.
    3. This process is repeated for several probability values over a predefined grid.
    4. We choose the mixing probability that the maximizes the log-likelihood.

    The grid is chosen so that no mixing probability is exactly zero, thereby avoiding degenerate solutions, where the mean and the variance associated with the 0-mixing probability cannot be estimated. From a theoretical perspective, this can be seen as a restriction of the space of Gaussian mixtures, where we only consider the mixing probabilities in the grid. It can be readily shown that the estimator of p is a super-consistent estimator of the pseudo-true value p_0, which is defined as the mixing probability that maximizes the population objective function within the grid. As a result, all theoretical results concerning the asymptotic validity of our test and model selection procedure remain applicable. The only difference is that the asymptotic distribution of $\hat{p}$ is no longer normal. To mitigate the risk of local minima, we perform each minimization three times with different starting values.

  - The maximization of the Log-likelihood is performed on the statistical software R using the package **nloptr** and more specifically NLOPT-LD-LBFGS algorithm. We provide an analytical gradient . The tolerance level for the minimization problem is set to 1e-9 .

- **Derivation of the Score and the Hessian.** Next, we provide an analytical expression for the Score and the Hessian of the log-likelihood. Providing a closed form expression for the score and the Hessian is useful for several reasons. First, the score corresponds the gradient of the log-likelihood maximization problem, and supplying a closed form the gradient significantly improves the algorithm's performance. Second, both the score and the Hessian are required to derive the asymptotic distribution of $\sqrt{N}(\hat{\phi} - \phi_0)$ as we will show shortly after; this asymptotic distribution is useful to conduct the pre-test in Section A.7. Furthermore, the score is necessary to derive the standard variance adjustment to correct for parameter uncertainty outlined in Section C.5.

Some of the derivations below build on some results in the seminal paper McFadden (1974). As previously, the index i refers the individual and j denotes the product. By construction, the choice probability implied by the parameter $\tilde{\phi}$ is equal to:

$$P_{ij} = \rho_j(X_i, \tilde{\phi}), \quad i = 1, 2, \ldots, N \quad \text{and} \quad j = 0, 1, \ldots, J,$$

where, as usual, we assume $x_{i0} = 0$. The log-likelihood function is given by:

$$\log L = \sum_{i=1}^{n} \sum_{j=0}^{J} Y_{ij} \log P_{ij}.$$

with $Y_{ij} = 1$ if individual i chooses product j and 0 otherwise. Differentiating the log-likelihood with respect to ϕ, we obtain the Score function

$$\frac{\partial \log L}{\partial \beta} = \sum_{i=1}^{N} \sum_{j=0}^{N} \frac{Y_{ij}}{P_{ij}} \frac{\partial P_{ij}}{\partial \phi},$$

Differentiating the Score further yields the Hessian of the log-likelihood:

$$\frac{\partial^2 \log L}{\partial \phi \partial \phi^T} = \sum_i \sum_j \frac{Y_{ij}}{P_{ij}} \left[\frac{\partial^2 P_{ij}}{\partial \phi \partial \phi^T} - \frac{1}{P_{ij}} \frac{\partial P_{ij}}{\partial \phi} \frac{\partial P_{ij}}{\partial \phi^T} \right].$$

After some manipulation, one can show that the derivative of the likelihood is equal to:

$$\begin{aligned}
\frac{\partial P_{ij}}{\partial \beta} &= \frac{\partial \rho_j(X_i, \tilde{\phi})}{\partial \beta} = \sum_{l=1}^{L} p_l \int \mathcal{T}_{ijl}(v) \left(X_{1ij} - \sum_{j'} \mathcal{T}_{ij'l}(v) X_{1ij'} \right) \phi(v) dv, \\
\frac{\partial P_{ij}}{\partial \mu_l} &= \frac{\partial \rho_j(X_i, \tilde{\phi})}{\partial \mu_l} = p_l \int \mathcal{T}_{ijl}(v) \left(X_{2ij} - \sum_{j'} \mathcal{T}_{ij'l}(v) X_{2ij'} \right) \phi(v) dv, \\
\frac{\partial P_{ij}}{\partial \sigma_l^2} &= \frac{\partial \rho_j(X_i, \tilde{\phi})}{\partial \sigma_l^2} = p_l \frac{1}{2} \frac{1}{\sqrt{\tilde{\sigma}_l^2}} \circ \int \mathcal{T}_{ijl}(v) \left(X_{2ij} - \sum_{j'} \mathcal{T}_{ij'l}(v) X_{2j'} \right) v \phi(v) dv, \\
\frac{\partial P_{ij}}{\partial p_l} &= \frac{\partial \rho_j(X_i, \tilde{\phi})}{\partial p_l} = \int \mathcal{T}_{ijl}(v) \phi(v) dv.
\end{aligned}$$

To maintain conciseness, we just provide closed form formulas for the diagonal terms of the Hessian of the likelihood. The rest of the terms can be derived in a similar way. First, we need to define the following quantities for characteristic k:

$$\overline{X}_{ikl}(v) = \sum_{j'} \mathcal{T}_{ij'l}(v) X_{ij'k}\ , \qquad \overline{X^2}_{ikl}(v) = \sum_{j'} \mathcal{T}_{ij'l}(v) X_{ij'k}^2\ , \qquad \overline{\tilde{X}}_{ikl}(v) = \sum_{j'} \mathcal{T}_{ij'l}(v) X_{ij'k} v.$$

We can show after rearanging the terms:

$$
\begin{aligned}
\frac{\partial^2 P_{ij}}{\partial \beta_k^2} &= \int \sum_{l=1}^{L} \tilde{p}_l \left[\mathcal{T}_{ijl}(v)(X_{1ijk}(v) - \bar{X}_{1ik}(v))^2 - \mathcal{T}_{ijl}(v)\overline{X^2}_{ikl}(v) + \mathcal{T}_{ij}(v)\overline{X}^2_{1ikl(v)} \right] \phi(v)\, dv, \\
\frac{\partial^2 P_{ij}}{\partial \mu_{lk}^2} &= \tilde{p}_l \int \left[\mathcal{T}_{ijl}(v)(X_{2ijk} - \overline{X}^l_{2ik}(v))^2 - \mathcal{T}_{ijl}(v)\overline{X^2}_{2ikl}(v) + \mathcal{T}_{ijl}(v)(\overline{X}_{2ikl}(v))^2 \right] \phi(v)\, dv, \\
\frac{\partial^2 P_{ij}}{\partial \sigma_{lk}^2} &= \frac{1}{4} \frac{1}{\tilde{\sigma}_{lk}^2} \tilde{p}_l \int \left[\mathcal{T}_{ijl}(v)(X_{2ijk}v - \overline{X}_{2ikl}(v))^2 - \mathcal{T}_{ijl}(v)\overline{X^2}_{2ikl}(v) + \mathcal{T}_{ijl}(v)(\overline{X}_{2ikl}(v))^2 \right] \phi(v)\, dv, \\
&- \frac{1}{4} \frac{1}{(\tilde{\sigma}_{lk}^2)^{3/2}} p_l \left[\frac{\partial P_{ij}}{\partial \tilde{\sigma}_{lk}^2} \right]_k \\
\frac{\partial^2 P_{ij}}{\partial p_l^2} &= 0.
\end{aligned}
$$

- **Asymptotic distribution of $\sqrt{N}(\hat{\phi} - \phi_0)$.** White (1982) establishes that in the context of the Maximum Likelihood estimator and under misspecification (i.e. $F \notin \mathcal{F}_0$), the pseudo true value is the parameter that minimizes the Kullback-Leibler distance between the likelihood under the parametric assumption and the true likelihood. We now use usual arguments for M-estimators to study the asymptotic behavior of the SML estimator under potential misspecification. First, we assume that pseudo-true value ϕ_0 is unique: $Q_0(\phi_0) < Q_0(\phi),\ \forall \phi \in \Phi \setminus \{\phi_0\}$. Moreover, when the log-likelihood is differentiable and $\hat{\phi}$ is in the interior of the parameter set Φ, the first-order condition implies that the maximum likelihood estimator satisfies:

$$0 = \sum_{i=1}^{N} \nabla_{\phi} \log l(W_i, \hat{\phi}).$$

Likewise, when the population Log-likelihood is differentiable and ϕ_0 is in the interior of the parameter set Φ, the first-order condition implies:

$$\mathbb{E}\left[\nabla_{\phi} \log l(W_i, \phi_0)\right] = 0. \tag{C.15}$$

Assuming twice continuous differentiability of the log-likelihood, the mean-value theorem guarantees the existence of $\overline{\phi}$ satisfying$||\overline{\phi} - \phi_0||_2 \leqslant ||\hat{\phi} - \phi_0||_2$ and such that:

$$0 = \sum_{i=1}^{N} \nabla_{\phi} \log l(W_i, \phi_0) + \sum_{i=1}^{N} \Xi(W_i, \overline{\phi})(\hat{\phi} - \phi_0)$$

where $\Xi(W_i, \phi) = \left(\frac{\partial^2 \log l(W_i, \phi)}{\partial \phi_i \partial \phi_j} \right)_{i,j}$. Asssuming that $\sum_{i=1}^N \Xi(W_i, \tilde{\phi})$ is invertible, the previous equality entails:

$$\sqrt{N}(\hat{\phi} - \phi_0) = -\left(\frac{1}{N} \sum_{i=1}^N \Xi(W_i, \overline{\phi}) \right)^{-1} \frac{1}{\sqrt{N}} \sum_{i=1}^N \nabla_\phi \log l(W_i, \phi_0).$$

Let

$$I_0 = \mathbb{E}\left[\nabla_\theta \log l(W_i, \phi_0) \nabla_\theta \log l(W_i, \phi_0)^\top \right]$$

be the information matrix and

$$\Xi_0 = \mathbb{E}\left[\Xi(W_i, \phi_0) \right]$$

the expected Hessian. Next, since $\overline{\phi}$ is between $\hat{\phi}$ and ϕ_0, the consistency of $\hat{\phi}$ implies the consistency of $\overline{\phi}$. Furthremore, assuming that the uniform law of large number holds, the Hessian term converges in probability to Ξ_0. Then the inverse Hessian converges in probability to Ξ_0^{-1} by continuity of the inverse at a nonsingular matrix (and assuming that Ξ_0 is full rank). Hence the asymptotic expansion becomes:

$$\sqrt{N}(\hat{\phi} - \phi) = \Xi_0^{-1} \frac{1}{\sqrt{N}} \sum_{i=1}^N \nabla_\phi \log l(W_i, \phi_0) + o_p(1).$$

Finally, using C.15, and applying the Central Limit Theorem along with the continuous mapping theorem, we obtain:

$$\sqrt{N}(\hat{\phi} - \phi_0) \xrightarrow[T \to +\infty]{d} \Xi_0^{-1} \mathcal{N}(0, I_0).$$

Under H_0, the information matrix equality leads to $\Xi_0 = -I_0$, further simplifying the asymptotic distribution established above.

C.4.2 BLP demand model

- **Estimation procedure.** We estimate the parameter $\phi_0 = (\beta_0^T, \lambda_0^T)^T$ using a nested fixed-point SMM procedure under the parametric assumption $H_0 : F \in \mathcal{F}_0$. For the choice of the weighting matrix, we perform a two-step GMM procedure. In the first stage, we use the identity matrix and in the second stage, the optimal weighting matrix. We denote $h_E(Z_t)$ the matrix of instruments used for estimation. Each market $t = 1, ..., T$ corresponds to an independent observation. The GMM estimator minimizes the following objective function:

$$\hat{\phi} \equiv \underset{\tilde{\phi}}{\operatorname{argmin}} \left(\sum_{t=1}^{T} h_E(Z_t) m_3(W_t, \tilde{\phi}) \right)^T W \left(\sum_{t=1}^{N} h_E(Z_t) m_3(W_t, \tilde{\phi}) \right)$$

where the function m_3 is defined in Example 3 as follows:

$$m_3 : (W_t, \tilde{\theta}) \mapsto \rho^{-1}(Y_t, X_{2t}, \tilde{F}) - X_{1t}\tilde{\beta}.$$

We know that there are no closed form for the inverse demand functions.[43] The construction of the feasible inverse demand functions requires the following 2 numerical approximations:

1. The demand functions $\rho_j(., X_{2t}, \tilde{F})$ must be computed through numerical integration. In this paper, we use Gaussian quadrature rules to approximate the integral. To be more specific, we use $k = 8$ nodes, which implies that the approximation rule will be exact for polynomial up to total order $2k-1 = 15$.
2. Following Berry et al. (1995), we derive the inverse of the demand functions $\rho^{-1}(Y_t, X_{2t}, \tilde{F})$ by solving the following contraction for each candidate parameter in the minimization procedure:

$$C : (\cdot, Y_t, X_{2t}, \tilde{F}) : \delta \mapsto \delta + \log(Y_t) - \log(\rho(Y_t, X_{2t}, \tilde{F})).$$

For the contraction, we use a threshold of 10^{-13}. We assume that both of the approximations above are sufficiently precise so that they can be ignored in the statistical analysis.

- **Parametrization of the model.** In the simulations, we only estimate the model under the assumption that F is a Gaussian. However, in the empirical application, the model is estimated under various specifications for F, including

[43] In this paper, we neglect the error arising from the fact that market shares Y_t are typically estimated rather than directly observed.

a normal distribution, a triweight distribution, a log-normal distribution, and a Gaussian mixture distribution. Below, we describe the model specification in the case of the Gaussian mixture (which includes the simple Gaussian as a special case). The parametrization of the other distributions is close to the one presented below. For more details on the parametrization of the other distributions, we refer the reader to our codes, which we made publicly available at the following link: https://sites.google.com/view/max-lesellier/home. Hence, we focus on the case where the estimation is performed under $H_0 :\ F \in \mathcal{F}_0$ with $\mathcal{F}_0$ the family of Gaussian mixtures with L components. We want to estimate the pseudo-true value $\phi_0 = (\beta_0, \lambda_0)$ where λ_0 characterizes the mixture up to permutations of indexes: $\lambda_0 = (p_{10}, \dots, p_{L0}, \mu_{10}, \dots, \mu_{L0}, \sigma^2_{10}, \dots, \sigma^2_{L0})$. The c.d.f. of a Gaussian mixture can be expressed as follows:

$$\forall x \in \mathbb{R}\ ,\ F_0(x|\tilde{\lambda}) = \sum_{l=1}^{L} p_l G_0(x|\tilde{\lambda}_l) \qquad \sum_{l=1}^{L} p_l = 1 \qquad L \geqslant 1$$

where $G_0(\cdot|\tilde{\lambda})$ is the c.d.f. of a $\mathcal{N}(\tilde{\mu}_l, \tilde{\sigma}_l{}^2)$.

To estimate the BLP demand model with a Gaussian mixture, we adopt the following parametrization: $v_i = \sum_{l=1}^{L} \mathbf{1}\{D_i = l\} v_{il}$ where v_{il} are i.i.d. and have density $G_0(\cdot|\tilde{\lambda}_l)$ known up to $\tilde{\lambda}_l$ for $l = 1, \dots, L$, and where $(D_i)_{i=1}^{n}$ are i.i.d categorically distributed with pmf $\mathbb{P}(D_i = l) = \tilde{p}_l$. For ease of exposition, we assume that for each component of the mixture, there is no correlation between the components of v_l. For all market t and product j, the demand functions are as follows. For any $\delta_t \in \mathbb{R}^{J^t}$,

$$\begin{aligned}
\rho_j(\delta_t, X_{2t}, F_0(.|\tilde{\lambda})) &= \int_{\mathbb{R}} \frac{\exp\{\delta_{jt} + X'_{2jt} v\}}{1 + \sum_{j'=1}^{J} \exp\{\delta_{j't} + X'_{2j't} v\}} dF_0(v|\tilde{\lambda}) dv \\
&= \sum_{l=1}^{L} \tilde{p}_l \int_{\mathbb{R}} \frac{\exp\{\delta_{jt} + X'_{2jt} v\}}{1 + \sum_{j'=1}^{J} \exp\{\delta_{j't} + X'_{2j't} v\}} dG_0(v|\tilde{\lambda}_l) dv \\
&= \sum_{l=1}^{L} \tilde{p}_l \int_{\mathbb{R}} \frac{\exp\{\delta_{jt} + X'_{2jt}(\tilde{\mu}_l + \sqrt{\tilde{\sigma}^2_l} v)\}}{1 + \sum_{j'=1}^{J} \exp\{\delta_{j't} + X'_{2j't}(\tilde{\mu}_l + \sqrt{\tilde{\sigma}^2_l} v)\}} d\Phi_0(v) dv.
\end{aligned}$$

where Φ_0 is the c.d.f. of the standard $\mathcal{N}(0,1)$. The linear part of the model is parametrized in the usual way: $\delta_{jt} = X_{2jt}\tilde{\beta} + \xi_{jt}$. As we did in the mixed logit case, the minimization is performed with respect to $\{\sigma_l^2\}_{l=1}^L$ instead and $(\sigma_l)_{l=1}^L$ directly. As highlighted by Ketz (2019), this parametrization of the model ensures that the gradient of both the sample and population log-likelihood remains full rank at the boundary of the parameter space, that is when $\tilde{\sigma}_l = 0$, or when $\tilde{\sigma}_l$ is close to zero.

- **Numerical minimization.**
  - As we did in the mixed logit case, we adopt a grid-based profiled maximum likelihood estimation. Rather than jointly maximizing the log-likelihood over both the mixing probabilities and the regular parameters, we use a grid-based profiling approach. This method helps address various numerical and theoretical challenges associated with estimating the mixing probabilities. For further details on our procedure and the benefits of this approach, we refer the reader to Section C.4.1.
  - The minimization of the GMM objective function is performed on the statistical software R using the package **nloptr** and more specifically NLOPT-LD-LBFGS algorithm. We provide an analytical gradient . The tolerance level for the minimization problem is set to 1e-9 while the threshold for the fixed-point is set to 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 between 7 and 10 independent cores.

- **Derivation of the first derivative of the empirical moment.** Next, we provide an analytical expression for the first derivative of the empirical moment:

$$\Gamma_T(\tilde{\phi}) \equiv \frac{1}{T}\sum_{t=1}^{T} h_E(Z_t)\frac{\partial m_3(W_t,\tilde{\phi})}{\partial \phi^T}.$$

For simplicity, we omit the dependence of $\Gamma_T(\tilde{\phi})$ in $h_E(Z_t)$. Providing a closed form expression for this object is useful for several reasons. First, as we show below, it enables us to supply the gradient of the objective function to the minimization algorithm, significantly enhancing the algorithm's performance. Second, the first derivative of the population moment is required to derive the asymptotic distribution of $\sqrt{T}(\hat{\phi} - \phi_0)$ as we will show shortly after; this asymptotic distribution is necessary to conduct the pre-test in Section A.7. Furthermore, this quantity is necessary to derive the standard variance adjustment to correct for parameter uncertainty outlined in Section C.5.

The gradient of GMM of objective function is expressed as follows:

$$\frac{\partial \mathcal{Q}(\tilde{\phi})}{\partial \phi} = 2\Gamma_T(\tilde{\phi})' W \left(\sum_{t=1}^{T} h_E(Z_t) m_3(W_t, \tilde{\phi}) \right).$$

First, we have: $\frac{\partial m_3(W_t, \tilde{\phi})}{\partial \beta} = -X_{1t}$ and for all t, $\rho(\delta_t, X_{2t}, \lambda) - Y_t = 0$ implies by the implicit function theorem:

$$\frac{\partial m_3(W_t, \tilde{\phi})}{\partial \lambda} = \frac{\partial \rho^{-1}(Y_t, X_{2t}, \tilde{F})}{\partial \lambda} = -\left[\frac{\partial \rho(\delta_t, X_{2t}, \tilde{\lambda})}{\partial \delta} \right]^{-1} \frac{\partial \rho(\delta_t, X_{2t}, \lambda)}{\partial \lambda}.$$

with $\delta_t = \rho^{-1}(Y_t, X_{2t}, \tilde{F})$. Moreover,

- $\frac{\partial \rho(\delta_t, X_{2t}, \tilde{\lambda})}{\partial \delta}$ is a $J \times J$ matrix such that:

$$\frac{\partial \rho_j\left(\delta_t, X_{2t}, \tilde{\lambda}\right)}{\partial \delta_{kt}} = \begin{cases} \sum_l \tilde{p}_l \int \mathcal{T}_{jlt}(v) \left(1 - \mathcal{T}_{klt}(v)\right) \phi_l(v) dv & if \quad j = k \\ -\sum_l \tilde{p}_l \int \mathcal{T}_{jlt}(v) \mathcal{T}_{klt}(v) \phi_l(v) dv & if \quad j \neq k \end{cases}$$

  with $\mathcal{T}_{jlt}(v) \equiv \frac{exp\{\delta_{jt} + x'_{2jt} v_l\}}{1 + \sum_{j'=1}^{J} exp\{\delta_{j't} + x'_{2j't} v_l\}}$.

- $\frac{\partial \rho(\delta_t, X_{2t}, tilde\lambda)}{\partial \lambda}$ is a $J \times 2L$ matrix such that:

$$\frac{\partial \rho_j\left(\delta_t, X_{2t}, \tilde{\lambda}\right)}{\partial \mu_l} = \tilde{p}_l \int \mathcal{T}_{jlt}(v)\left(X_{2jt} - \sum_{j'} \mathcal{T}_{j'lt}(v) X_{2j't}\right)\phi(v)dv,$$

$$\frac{\partial \rho_j\left(\delta_t, X_{2t}, \tilde{\lambda}\right)}{\partial \sigma_l^2} = \frac{1}{2}\frac{1}{\sqrt{\tilde{\sigma}_l^2}} p_l \int \mathcal{T}_{jlt}(v)\left(X_{2jt} - \sum_{j'} \mathcal{T}_{j'lt}(v) X_{2j't}\right) v\phi(v)dv,$$

$$\frac{\partial \rho_j\left(\delta_t, X_{2t}, \tilde{\lambda}\right)}{\partial p_l} = \int \mathcal{T}_{jlt}(v)\phi(v)dv.$$

- **Asymptotic distribution of $\sqrt{T}(\hat{\phi} - \phi_0)$.** To study the behavior of the GMM estimator under potential misspecification (i.e. $F \notin \mathcal{F}_0$) we partially rely on findings in Hall and Inoue (2003) as well as usual arguments for M-estimators. First, as in Hall and Inoue (2003), that the pseudo-true value ϕ_0 is unique : $Q_0(\phi_0) < Q_0(\phi),\ \forall \phi \in \Phi \setminus \{\phi_0\}$. As noted in Hall and Inoue (2003), when the model is misspecified, the pseudo-true value ϕ_0 depends on the weighting matrix W and the instruments $h_E(Z_t)$. For exposition, in what follows, we omit the dependence of ϕ_0 in $h_E(Z_t)$ and W. When the GMM objective function is differentiable and $\hat{\phi}$ is in the interior of the parameter set Φ, the first-order condition implies:

$$\left(\frac{1}{T}\sum_{t=1}^{T} h_E(Z_t)\frac{\partial m_3(W_t, \hat{\phi})}{\partial \phi^T}\right)^T W \frac{1}{T}\sum_{t=1}^{T} h_E(Z_t) m(W_t, \hat{\phi}) = 0. \tag{C.16}$$

Likewise, when the population GMM objective function is differentiable and ϕ_0 is in the interior of the parameter set Φ, the first-order condition implies:

$$\mathbb{E}\left[h_E(Z_t)\frac{\partial m_3(W_t, \phi_0)}{\partial \phi^T}\right]^T W \mathbb{E}\left[h_E(Z_t) m_3(W_t, \phi_0)\right] = 0. \tag{C.17}$$

In the following, we define Γ_0 as:

$$\Gamma_0 = \mathbb{E}\left[h_E(Z_t)\frac{\partial m(W_t, \phi_0)}{\partial \phi^T}\right].$$

Again, we omit the dependence of Γ_0 in $h_E(Z_t)$ and W. By the mean-value theorem, there exists some $\overline{\phi}$ such that $||\overline{\phi}-\phi_0||_2 \leqslant ||\hat{\phi}-\phi_0||_2$ and

$$\sum_{t=1}^{T} h_E(Z_t) m(W_t,\hat{\phi}) = \sum_{t=1}^{T} h_E(Z_t) m_3(W_t,\phi_0) + \sum_{t=1}^{T} h_E(Z_t)\frac{\partial m_3(W_t,\overline{\phi})}{\partial \phi}(\hat{\phi}-\phi_0). \tag{C.18}$$

Combining (C.16) and (C.18), and assuming $\Gamma_T(\overline{\phi})$, $\Gamma_T(\hat{\phi})$ and W are full rank, after rearranging, we obtain:

$$\sqrt{T}(\hat{\phi}-\phi_0) = -\left(\frac{1}{T}\sum_{t=1}^{T}\frac{\partial m_3(W_t,\hat{\phi})}{\partial \phi} h_E(Z_t)^T W \frac{1}{T}\sum_{T=1}^{T} h_E(Z_t)\frac{\partial m_3(W_t,\overline{\phi})}{\partial \phi^T}\right)^{-1} \frac{1}{T}\sum_{t=1}^{T}\frac{\partial m_3(W_t,\hat{\phi})}{\partial \phi} h_E(Z_t)^T W \frac{1}{\sqrt{T}}\sum_{t=1}^{T} h_E(Z_t) m_3(W_t,\phi_0)$$

Under the usual regularity conditions, $\hat{\phi}$ is a consistent estimator of ϕ_0 and by construction $||\overline{\phi}-\phi_0||_2 \leqslant ||\hat{\phi}-\phi_0||_2$. Therefore, assuming that the uniform law of large numbers holds, we obtain:

$$\frac{1}{T}\sum_{t=1}^{T} h_E(Z_t)\frac{\partial m_3(W_t,\overline{\phi})}{\partial \phi^T} \xrightarrow[T\to+\infty]{P} \Gamma_0$$

Furthermore, we assume that W and Γ_0 are both full rank. Then, the continuous mapping theorem yields:

$$\sqrt{T}(\hat{\phi}-\phi_0) = -\left(\Gamma_0^T W \Gamma_0\right)^{-1}\Gamma_0^T W \frac{1}{\sqrt{T}}\sum_{t=1}^{T} h_E(Z_t) m_3(W_t,\phi_0) + o_p(1).$$

Finally, using C.17, and applying the Central Limit Theorem along with the continuous mapping theorem, we obtain:

$$\sqrt{T}(\hat{\phi}-\phi_0) \xrightarrow[T\to+\infty]{d} \left(\Gamma_0^T W \Gamma_0\right)^{-1}\Gamma_0^T W \mathcal{N}(0,\Omega_0).$$

where $\Omega_0 = \text{var}\left(h_E(Z_t) m_3(W_t,\phi_0)\right)$. Let us observe that the asymptotic distribution of $\sqrt{T}(\hat{\phi}-\phi_0)$ is simpler than the one derived in Hall and Inoue (2003). This simplification arises from using the identity in C.17.

- **Intruments used for estimation.** Here, we describe the different sets of instruments used for estimation: the differentiation instruments based on Gandhi and Houde (2019) and the "optimal instruments" based on Reynaert and Verboven

(2014). Both sets of instruments are built upon the following baseline set:

$$(\mathbf{1}, x_{aj}, x_{bj}, x_{cj}, x_{aj}^2, x_{bj}^2, x_{cj}^2, x_{aj}x_{bj}, x_{aj}x_{cj}, x_{bj}x_{cj}).$$

Each set also incorporates specific instruments:

- **Differentiation instruments**: these are defined following Gandhi and Houde (2019) and include:

$$\sum_{k=1}^{J}(x_{ck}-x_{cj})^2, \quad \sum_{k=1}^{J}(x_{ck}-x_{cj})^3, \quad \sum_{k=1}^{J}(x_{ck}-x_{cj})^4,$$

$$\left(\sum_{k=1}^{J}(x_{ck}-x_{cj})^2\right)^3, \quad \left(\sum_{k=1}^{J}(x_{ck}-x_{cj})^2\right)^4, \quad \left(\sum_{k=1}^{J}(x_{ck}-x_{cj})^3\right)^2.$$

- **Optimal instruments**: these instruments are computed in two stages as the derivation of the "optimal instruments" require a preliminary estimate of λ:

  1. The *first-stage instruments* consist of the differentiation instruments combined with the baseline set of instruments.
  2. The *second-stage instruments* consist of the baseline set and the approximation of optimal instruments proposed in Reynaert and Verboven (2014), which approximates:

$$\mathbb{E}\left[\frac{\partial \rho_j^{-1}(S_t, X_{2t}, \lambda)}{\partial \lambda}\middle| Z_t\right].$$

C.5 Standard variance adjustments to correct for parameter uncertainty

In this section, we provide a detailed explanation of the alternative procedures used to adjust the variance-covariance matrix, accounting for the fact that the parameters are estimated in the first stage. These adjustments are referred to as standard variance corrections in our simulation results. Furthermore, these alternative procedures enable us to assess the effectiveness of the orthogonalization approach developed in this paper.

Mixed logit model. In the mixed logit model, the parameter of interest ϕ_0 is estimated using a nested-fixed point SMM procedure, which we describe in Section C.4.1. To adjust the variance-covariance matrix to account for parameter uncertainty, we simply derive the asymptotic distribution of our test statistic (without orthogonalization) under $H_0 : F \in \mathcal{F}_0$. From Section C.4.2, we have under the same regularity conditions:

$$\sqrt{N}(\hat{\phi} - \phi) = \Xi_0^{-1} \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \nabla_\phi \log l(W_i, \phi_0) + o_p(1)$$

Moreover, under H_0, the information matrix equality $\Xi_0 = -I_0$ holds and consequently, the asymptotic expansion becomes:

$$\sqrt{N}(\hat{\phi} - \phi) = I_0^{-1} \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \nabla_\phi \log l(W_i, \phi_0) + o_p(1)$$

Next, we can derive the asymptotic distribution of our test statistic under H_0. By a mean-value expansion around ϕ_0, for any $h \in \mathcal{H}_0$,

$$\begin{aligned}
\frac{1}{\sqrt{N}} \sum_{i=1}^{N} h(Z_i) m(W_i, \hat{\phi}) &= \frac{1}{\sqrt{N}} \sum_{i=1}^{N} h(Z_i) m(W_i, \phi_0) + \frac{1}{N} \sum_{i=1}^{N} h(Z_i) \frac{\partial m(W_i, \overline{\phi})}{\partial \phi} \sqrt{N}(\hat{\phi} - \phi_0), \\
&= \frac{1}{\sqrt{N}} \sum_{i=1}^{N} h(Z_i) m(W_i, \phi_0) + \mathbb{E}\left[h(Z_i) \frac{\partial m(W_i, \phi_0)}{\partial \phi} \right] I_0^{-1} \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \nabla_\phi \log l(W_i, \phi) + o_p(1), \\
&\equiv (I_L, \Gamma_0^* I_0^{-1}) \sqrt{N} \begin{pmatrix} \frac{1}{N} \sum_{i=1}^{N} h(Z_i) m(W_i, \phi_0) \\ \frac{1}{N} \sum_{i=1}^{N} \nabla_\phi \log l(W_i, \phi_0), \end{pmatrix} \\
&\xrightarrow[N \to +\infty]{d} (I_L, \Gamma_0^* I_0^{-1}) \mathcal{N}(0, \Omega_0).
\end{aligned}$$

where $||\overline{\phi} - \phi_0||_2 \leqslant ||\hat{\phi} - \phi_0||_2$. Furthermore, the second line stems from the asymptotic distribution of $\sqrt{N}(\hat{\phi} - \phi_0)$ derived previously and the uniform law of large numbers. The last line is implied by Slutky's lemma and the central limit theorem, where $\Omega_0 = \text{var}((h(Z_i) m(W_i, \phi_0))^T, (\nabla_\phi \log l(W_i, \phi))^T)^T)$.

BLP demand model. In the BLP demand model, the parameter of interest ϕ_0 is estimated using a nested-fixed point SMM procedure, which we describe in Section C.4.2. To adjust the variance-covariance matrix to account for parameter uncertainty, we simply derive the asymptotic distribution of our test statistic (without orthogonalization) under $H_0 : F \in \mathcal{F}_0$. Here, unlike under misspecification, ϕ_0 does not depend on $h_E(Z_t)$ nor W. From Section C.4.2 (under the same regularity conditions), we directly obtain:

$$\sqrt{T}(\hat{\phi} - \phi_0) = -\left(\Gamma_0^T W \Gamma_0\right)^{-1} \Gamma_0^T W \frac{1}{\sqrt{T}} \sum_{t=1}^{T} h_E(Z_t) m_3(W_t, \phi_0) + o_p(1).$$

Next, by a mean-value expansion of the tested empirical moment around ϕ_0, there exists $\phi^\dagger$ such that $||\phi^\dagger - \phi_0||_2 \leqslant ||\hat{\phi} - \phi_0||_2$ and

$$\begin{aligned}
\sqrt{T}\frac{1}{T}\sum_{t=1}^{T} h(Z_t) m(W_t, \hat{\phi}) &= \sqrt{T}\frac{1}{T}\sum_{t=1}^{T} h(Z_t) m(W_t, \phi_0) + \frac{1}{N}\sum_{t=1}^{T} h(Z_t) \frac{\partial m(W_t, \phi^\dagger)}{\partial \phi^T} \sqrt{T}(\hat{\phi} - \phi_0) \\
&= \sqrt{T}\frac{1}{T}\sum_{t=1}^{T} h(Z_t) m(W_t, \phi_0) + \Gamma_0^* \sqrt{T}(\hat{\phi} - \phi_0) + o_p(1) \\
&= \sqrt{T}\frac{1}{T}\sum_{t=1}^{T} h(Z_t) m(W_t, \phi_0) - \Gamma_0^* \left(\Gamma_0^T W \Gamma_0\right)^{-1} \Gamma_0^T W \sqrt{T}\frac{1}{T}\sum_{t=1}^{N} h_E(Z_t) m(W_t, \phi_0) + o_p(1) \\
&= (I_L \quad - \Gamma_0^* \left(\Gamma_0^T W \Gamma_0\right)^{-1} \Gamma_0^T W) \begin{pmatrix} \frac{1}{\sqrt{T}} \sum_{t=1}^{T} h(Z_t) m(W_t, \phi_0) \\ \frac{1}{\sqrt{T}} \sum_{t=1}^{T} h_E(Z_t) m(W_t, \phi_0) \end{pmatrix} + o_p(1) \\
&\xrightarrow[T \to +\infty]{d} (I_L \quad - \Gamma_0^* \left(\Gamma_0^T W \Gamma_0\right)^{-1} \Gamma_0^T W) \mathcal{N}(0, \Sigma_0).
\end{aligned}$$

where $\Gamma_0^* = \mathbb{E}\left[h(Z_i)\frac{\partial m(W_i, \phi_0)}{\partial \phi^T}\right]$. The second equality uses the uniform law of large numbers and the third equality uses the asymptotic distribution of $\sqrt{T}(\hat{\phi} - \phi_0)$ derived previsouly. The last line is implied by Slutky's lemma and the central limit theorem, where $\Sigma_0 = \text{var}((h(Z_i)m(W_i, \phi_0))^T, (h_E(Z_i)m(W_i, \phi_0))^T)^T)$.

C.6 Details on the instruments used for the specification test

Infeasible MPI. The infeasible MPI corresponds to the intrument defined in Proposition 3.2, which can only be derived if the econometrician knows F_a. In our simulations, we ignore the variance term. In the case of the BLP model, even if one knows (F_a, β_a), there is no closed form for $\mathbb{E}[\Delta_{0,a}^m(W_i)|Z_i]$. Therefore, we use the local approximation $\tilde{\Delta}_{0,a}^m(W_i)$ outlined in Proposition 1.6. However, computing the conditional expectation remains challenging even with this approach. Therefore, following Reynaert and Verboven (2014), we use the approximation $\mathbb{E}[\tilde{\Delta}_{0,a}^m(W_i)|Z_i] \approx \tilde{\Delta}_{0,a}^m(\mathbb{E}[W_i|Z_i])$.

Feasible MPI (i.e interval instruments). The feasible MPIs correspond to the instruments whose construction is outlined in Section 4.1. In the case of the BLP demand model, the derivation of the interval instruments requires to take the conditional expectation over known functions $\mathbb{E}[\psi_0(W_i, v_l, \beta_a)|Z_i]$ which can be quite challenging. Instead, we follow the approach of Reynaert and Verboven (2014), we use the approximation $\mathbb{E}[\psi_0(W_i, v_l, \beta_a)|Z_i] \approx \psi_0(\mathbb{E}[W_i|Z_i], v_l, \beta_a)$.

In both the mixed logit and BLP demand models, constructing the interval instruments requires selecting L points in the support of F_a. In the Mixed logit case, we perform the test with $L = 6$ and $L = 10$ whereas in the BLP demand model, we perform the test with $L = 8$ instruments. In our simulations, we use a straightforward procedure to select points in the support of F_a to construct the interval instruments. Importantly, this procedure is uniform across all specifications considered, and it does not exploit knowledge of the alternative when choosing the points. We proceed as follows. Let L denote the number of instruments, or equivalently, the number of points chosen in the support. Define $U = \{-0.25, 0.25\}$. The choice of points v_k depends on whether $L \leq 6$ or $L > 6$, as outlined below:

- **Case 1:** $L \leq 6$ In this case, the points v_k are chosen as:

  $$v_k \in U \cup \left\{ -2 + \frac{4k}{L-2} : k = 0, 1, \ldots, L-2 \right\}.$$

  Here, the points are evenly spaced between -2 and 2, with a step size of $\frac{4}{L-2}$, and these points are combined with the fixed set U.

- **Case 2:** $L > 6$ In this case, the range of the interval is symmetrically extended to accommodate the larger number of instruments. The points v_k are chosen as:

  $$v_k \in U \cup \left\{ -2 - 0.75 \cdot \frac{L-6}{2} + \frac{Rk}{L-2} : k = 0, 1, \ldots, L-2 \right\},$$

  where the total range R is given by

  $$R = 4 + 0.75 \times (L - 6).$$

  The interval extends symmetrically by $0.75 \cdot \frac{L-6}{2}$ on both ends, and the step size is calculated as $\frac{R}{L-2}$.

For the choice of β_a, which is required in the mixed logit model, we simply take $\hat{\beta}_0$. This choice is motivated by the observation that the distribution of RCs affects only marginally the estimation of the parameter β.

Polynomial instruments. We use Hermite polynomials with a tail decaying at an exponential rate. Namely, the formula for the k^{th} polynomial is:

$$p_k(x) = H_{e,n}(x) \left(1\{|x| > z_{0.975}\} \exp\left(\frac{-(|x| - z_{0.975})^2}{2} \right) + 1\{|x| \leq z_{0.975}\} \right)$$

where $H_{e,k}$ is the Hermite polynomial of order k and $z_{0.975}$ is the quantile of level 0.975 of the standard normal $\mathcal{N}(0, 1)$.

D Empirical application

D.1 Descriptive statistics

Table 12 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.

Table 12: Summary Statistics (Sales weighted)

	Year						
	2012	2013	2014	2015	2016	2017	2018
Diesel							
Price/income	0.77	0.74	0.74	0.74	0.74	0.71	0.70
Size (m2)	8.32	8.32	8.33	8.37	8.43	8.49	8.55
Horsepower (kW/100)	1.12	1.12	1.13	1.15	1.18	1.20	1.24
Fuel cost (euros/100km)	7.76	7.17	6.68	5.59	5.04	5.45	6.13
Fuel cons. (Lt./100km)	5.25	5.05	4.95	4.78	4.68	4.70	4.78
CO2 emission (g/km)	137.70	132.36	129.34	125.00	122.30	122.79	125.09
Nb. products	163	172	189	202	193	188	186
Gasoline							
Price/income	0.48	0.46	0.46	0.47	0.47	0.46	0.44
Size (m2)	7.26	7.28	7.29	7.33	7.40	7.47	7.54
Horsepower (kW/100)	0.81	0.81	0.83	0.86	0.90	0.92	0.94
Fuel cost (euros/100km)	9.35	8.62	8.18	7.37	6.85	7.31	7.77
Fuel cons. (Lt./100km)	5.81	5.53	5.47	5.37	5.35	5.43	5.44
CO2 emission (g/km)	136.86	129.63	127.03	124.56	123.50	124.70	124.61
Nb. products	206	233	248	265	262	267	258

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

D.2 Details on the estimation procedure and the specification test

Estimation procedure. For the estimation of the model, we closely follow the procedure outlined in Section C.4.2. The model is estimated under various specifications for F, including a degenerate distribution, normal distribution, triweight distribution, log-normal distribution, and Gaussian mixture distribution.

The estimation proceeds in three steps. First, we estimate the model using a baseline set of instruments, which we describe shortly. With this preliminary estimate in hand, we compute an approximation of the optimal instruments based on Reynaert and Verboven (2014). We then re-estimate the model, augmenting the baseline instruments with these optimal instruments. Collinear instruments are removed using a sequential elimination procedure. Finally, in the third step, we re-estimate the model again using the optimal weighting matrix.

Instruments used for estimation. For all specifications, we use a baseline set of instruments that includes cost shifters, namely transformations of the real effective exchange rate between the country of assembly and Germany, a distance dummy (equal to one if the plant is located more than 10,000 km from Germany), and standard BLP instruments related to vehicle size, fuel cost, horsepower, and height (sum of product characteristics for both the firm's own and rival products). In addition, for all specifications except the degenerate distribution, we include *differentiation instruments* to help identify the distribution of random coefficients. These instruments follow the approach of Gandhi and Houde (2019) and are defined as:

$$\sum_{k=1}^{J}(\hat{p}_k-\hat{p}_j)^2, \quad \sum_{k=1}^{J}(\hat{p}_k-\hat{p}_j)^3, \quad \sum_{k=1}^{J}(\hat{p}_k-\hat{p}_j)^4, \sum_{k=1}^{J}|\hat{p}_k-\hat{p}_j|\mathbf{1}\{|\hat{p}_k-\hat{p}_j|<0.2\}, \quad \sum_{k=1}^{J}|\hat{p}_k-\hat{p}_j|\mathbf{1}\{|\hat{p}_k-\hat{p}_j|<0.2\}$$

Here, $\hat{p}$ denotes the predicted price obtained from regressing price on the instruments, used as a proxy for the conditional expectation of price. The approximation of the optimal instruments used in the second stage follows the procedure in Section C.4.2.

Construction of the interval instruments. The construction of the interval instruments follows Section C.6. The only modification concerns the selection of points in the support of F. For all specifications, we use 15 equally spaced points in the interval $[-1, -21]$. Collinear instruments are removed using a sequential elimination procedure.

D.3 Preliminary analysis

In Table 13, we present the first-stage regression results for price, which is an equilibrium outcome and therefore endogenous. The explanatory variables include exogenous product characteristics, along with excluded instruments. These instruments include cost shifters, transformations of the real effective exchange rate between the country of assembly and Germany, a distance dummy (equal to one if the plant is located more than 10,000 km from Germany), and standard BLP instruments related to vehicle size, fuel cost, horsepower, and height (sum of product characteristics for both the firm's own and rival products). We find that the excluded instruments are jointly significant, with an F-statistic of 38,780, indicating strong relevance. The results show that a higher exchange rate between the Euro and the local currency at the place of assembly is associated with a lower vehicle price, consistent with a reduction in production costs. Additionally, we observe that the absence of an exchange rate, which in most cases implies that the car is manufactured in Europe, is negatively correlated with price. This may reflect the impact of shipping expenses being incorporated into the final price. The distance dummy also increases the price.

Table 13: First-stage regression for price

	Dependent variable: Price
Constant	-0.226^{***}
	(0.081)
Power (KW/100)	0.377^{***}
	(0.002)
Diesel	0.090^{***}
	(0.002)
Fuel cost	0.007^{***}
	(0.001)
Size (m^2)	0.054^{***}
	(0.002)
Foreign	0.002
	(0.002)
Height (m)	0.126^{***}
	(0.007)
Diesel × Trend Post-2015	-0.004^{***}
	(0.001)
Exchange rate	-0.011^{***}
	(0.001)
Missing exchange rate	-0.547^{***}
	(0.069)
Exchange rate^2	0.0001^{***}
	(0.00001)
Distance	0.010^{***}
	(0.003)
BLP Product 0	0.001
	(0.001)
BLP Firm 0	-0.006^{***}
	(0.001)
BLP Product 1	-0.00002^{**}
	(0.00001)
BLP Product 2	0.0002
	(0.0004)
BLP Firm 2	0.002^{***}
	(0.001)
BLP Product 3	−0.002
	(0.001)
BLP Firm 3	-0.033^{***}
	(0.002)
BLP Product 4	0.001
	(0.002)
BLP Firm 4	0.042^{***}
	(0.003)
Observations	47,753
R^2	0.982
Adjusted R^2	0.982
Residual Std. Error	0.105 (df = 47,684)
F Statistic	$38{,}781^{***}$ (df = 69; 47,684)

D.4 Additional empirical results

Here, we present the estimates of the homogeneous parameters in the utility function. The signs of all parameters align with economic intuition. Furthermore, the estimates of β show little variation across specifications, indicating robustness to the choice of distribution for the random coefficients.

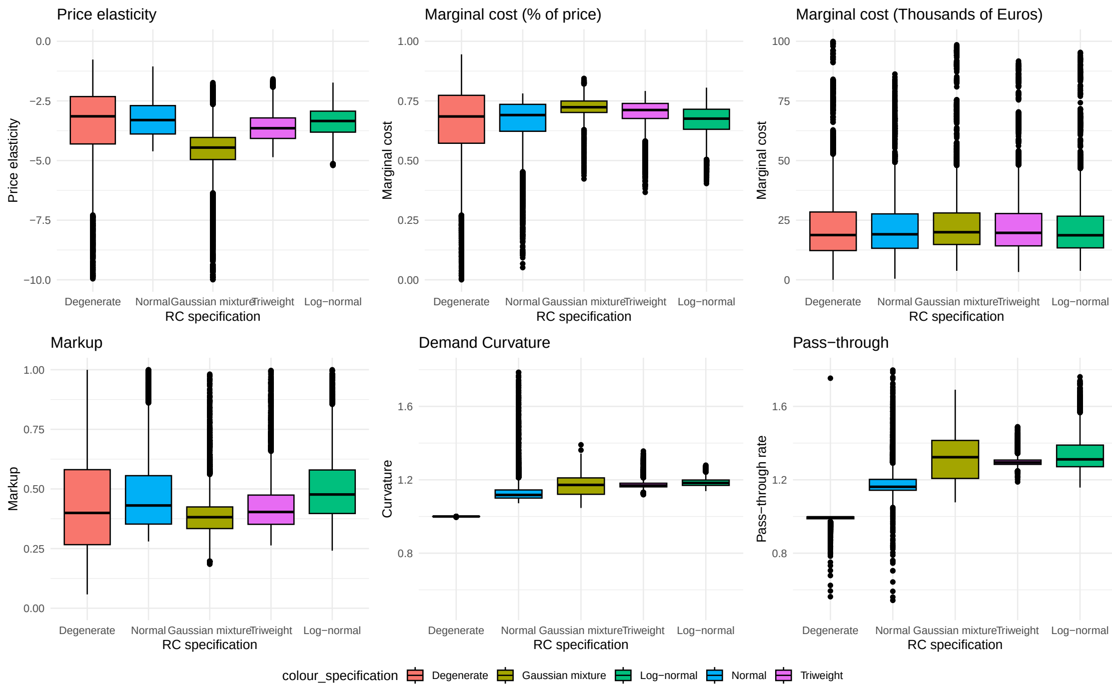
Table 14: Estimation results under different distributional assumptions for the price RC

		Specification				
Variable		Degenerate	Normal	Gaussian Mixture	Triweight	Log-normal
Constant	Estimate	-15.198	-14.219	-13.414	-13.422	-11.709
	Std. Error	(0.367)	(0.256)	(0.285)	(0.234)	(0.260)
Power (KW/100)	Estimate	2.221	1.197	1.603	1.375	1.327
	Std. Error	(0.152)	(0.075)	(0.092)	(0.092)	(0.069)
Diesel	Estimate	0.168	0.235	0.302	0.269	0.230
	Std. Error	(0.046)	(0.037)	(0.048)	(0.043)	(0.040)
Fuel cost (euro/100 km)	Estimate	-0.153	-0.138	-0.141	-0.140	-0.141
	Std. Error	(0.010)	(0.007)	(0.010)	(0.009)	(0.010)
Size (m^2)	Estimate	0.288	0.509	0.564	0.540	0.479
	Std. Error	(0.029)	(0.030)	(0.030)	(0.028)	(0.027)
Foreign	Estimate	-0.247	-0.298	-0.277	-0.284	-0.285
	Std. Error	(0.015)	(0.014)	(0.015)	(0.014)	(0.014)
Height (m)	Estimate	3.326	2.596	2.843	2.763	2.737
	Std. Error	(0.114)	(0.076)	(0.080)	(0.084)	(0.076)
Diesel × Trend Post-2015	Estimate	-0.191	-0.202	-0.212	-0.207	-0.211
	Std. Error	(0.015)	(0.014)	(0.016)	(0.015)	(0.016)

Note: Brand, Class, and State fixed effects, as well as a time trend, are included but not reported.

In Figure 4, we plot the empirical distributions of the counterfactual quantities. We observe that the triweight and Gaussian mixtures generate lower price elasticities and higher pass-through rates than the normal or degenerate specifications.

Figure 4: Empirical distribution of quantities of interest under different specifications



Note: to improve the readability of the box plots, some extreme values have been truncated.

D.5 Model selection

Next, we follow the procedure outlined in Section 5 to test for equality of the slopes across different specifications. This enables us to rank specifications, including those rejected by our specification test. If a specification yields a significantly larger slope, it indicates that the specification is less supported by the data. The null hypothesis corresponds to the equality of slopes.

We report the results in Table 15. The triweight specification is consistently selected across comparisons. It is followed by the Gaussian mixture and log-normal distributions, for which we cannot reject slope equality at the 5% level. The normal and degenerate distributions are the least supported by the data.

Table 15: RV Test for Comparison of Slopes

	Distribution under H_2			
Distribution under H_1	Normal	Gaussian Mixture	Triweight	Log-normal
Degenerate	1.132	4.949	7.212	3.496
Normal		4.162	6.497	2.543
Gaussian Mixture			3.234	-1.509
Triweight				-4.163

Critical values: $c_{0.10} = \pm 1.65$, $c_{0.05} = \pm 1.96$, $c_{0.01} = \pm 2.57$.

Finally, since the normal specification is nested within the Gaussian mixture, we apply the pre-test outlined in Section A.7 to verify whether the two distributions differ. The results are presented in Table 16. We reject the null hypothesis $H_0 : F_{\text{mixture}} \in \mathcal{F}_{\text{normal}}$ at the 5% level, which implies that we can the test for equality of slopes is valid.

Table 16: Pre-test $H_0 : F_{mixture} \in \mathcal{F}_{normal}$ $(\alpha = 5\%)$

Stat.	Critical val.	DF
164.009	3.841	1

D.6 Details on derivation of the quantities of interest

In this section, we define the quantities of interest—price elasticity, marginal cost, markup, demand curvature, and pass-through rate—that we compute under various specifications for the distribution of random coefficients. 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.

Price elasticities. For the calculation of the price elasticities,we use the following expression:

$$\eta_j^j = \frac{p_j}{s_j}\frac{\partial s_j}{\partial p_j} = \frac{p_j}{s_j}\int v\left(1 - \frac{\exp\{\delta_j + p_j v\}}{1 + \sum_{j'=1}^{J}\exp\{\delta_{j'} + p_{j'} v\}}\right)\frac{\exp\{\delta_j + p_j v\}}{1 + \sum_{j'=1}^{J}\exp\{\delta_{j'} + p_{j'} v\}}\,dF(v).$$

Demand curvature. The demand curvature is defined as follows: $\eta_j^2(\mathbf{p}) = s_j(\mathbf{p}) \frac{\partial^2 s_j(\mathbf{p})}{\partial p_j^2} \left(\frac{\partial s_j(\mathbf{p})}{\partial p_j} \right)^{-2}$.

Marginal costs and mark-ups. To recover the marginal costs and the implied markups, 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. Furthermore, we assume that pricing decisions are made independently across markets (we remove the index t for simplicity). The profit of each firm f is equal to:

$$\Pi_f(\mathbf{p}) = \sum_{j \in J_f} (p_j - c_j)\, s_j(\mathbf{p})$$

where J_f is the set of goods produced by firm f, c_j is the marginal cost for good j, $s_j(p)$ is the market share of product j. The first-order condition with respect to price p_j yields:

$$s_j(\mathbf{p}) + \sum_{j' \in J_f} (p_{j'} - c_{j'}) \frac{\partial s_{j'}(\mathbf{p})}{\partial p_j} = 0.$$

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 $\mathbf{\Delta}(\mathbf{p}) = \frac{\partial s_{j'}(\mathbf{p})}{\partial p_j}$ 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:

$$\mathbf{c} = \mathbf{p} - (\Delta(\mathbf{p}))^{-1}\, \mathbf{s}(\mathbf{p}). \tag{D.19}$$

The mark-up for product j equals: $(p_j - c_j)/c_j$.

Pass-through. 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. In our simulations, we consider a macro shock that increases the marginal cost of all the products by 1%. We calculate the new equilibrium price using Eqn. D.19 using fixed point iteration. In each iteration the price is updated following the rule: $p_{k+1} = \gamma p^* + (1 - \gamma)p_{k+1}$. We set $\gamma = 0.5$. The pass-through rate corresponds to the proportion of the cost increase that is transmitted to the price.

Finally, note that for the normal distribution, we truncate the distribution of the random coefficient at -0.05 to exclude consumers with positive price elasticities (who prevent our pass-through algorithm to converge).

Appendix References

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

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

Andrews, Donald WK (1987): "Asymptotic results for generalized Wald tests," *Econometric Theory*, 3 (3), 348–358.

——— (1999): "Estimation when a parameter is on a boundary," *Econometrica*, 67 (6), 1341–1383.

Berry, Steven T (1994): "Estimating discrete-choice models of product differentiation," *The RAND Journal of Economics*, 25 (2), 242–262.

Berry, Steven T and Philip 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 (5), 1749–1797.

Berry, Steven T, James Levinsohn, and Ariel Pakes (1995): "Automobile prices in market equilibrium," *Econometrica*, 63 (4), 841–890.

Chamberlain, Gary (1987): "Asymptotic efficiency in estimation with conditional moment restrictions," *Journal of Econometrics*, 34 (3), 305–334.

Chen, Xiaohong (2007): "Large sample sieve estimation of semi-nonparametric models," *Handbook of econometrics*, 6, 5549–5632.

Cox, David R (1961): "Tests of separate families of hypotheses," in *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability*, vol. 1, 96.

Davidson, Russell and James G MacKinnon (1981): "Several tests for model specification in the presence of alternative hypotheses," *Econometrica*, 49 (3), 781–793.

Dubé, Jean-Pierre, Jeremy T Fox, and Che-Lin Su (2012): "Improving the numerical performance of static and dynamic aggregate discrete choice random coefficients demand estimation," *Econometrica*, 80 (5), 2231–2267.

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

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

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

Fox, Jeremy T., Kyoo il Kim, Stephen P. Ryan, and Patrick Bajari (2012): "The random coefficients logit model is identified," *Journal of Econometrics*, 166 (2), 204–212.

Fox, Jeremy T., Kyoo il Kim, and Chenyu Yang (2016): "A simple nonparametric approach to estimating the distribution of random coefficients in structural models," *Journal of Econometrics*, 195 (2), 236–254.

Gandhi, Amit and Jean-François Houde (2019): "Measuring substitution patterns in differentiated-products industries," *NBER working paper 26375*.

Geweke, John (1981): "The approximate slopes of econometric tests," *Econometrica*, 49 (6), 1427–1442.

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

Hall, Alastair R. and Atsushi Inoue (2003): "The large sample behaviour of the generalized method of moments estimator in misspecified models," *Journal of Econometrics*, 114 (2), 361–394.

Hansen, Bruce (2022): *Probability and statistics for economists*, Princeton University Press.

Ichimura, Hidehiko and T Scott Thompson (1998): "Maximum likelihood estimation of a binary choice model with random coefficients of unknown distribution," *Journal of Econometrics*, 86 (2), 269–295.

Ketz, Philipp (2019): "On asymptotic size distortions in the random coefficients logit model," *Journal of Econometrics*, 212 (2), 413–432.

Lewbel, Arthur (2000): "Semiparametric qualitative response model estimation with unknown heteroscedasticity or instrumental variables," *Journal of Econometrics*, 97 (1), 145–177.

McFadden, D. (1974): "Conditional logit analysis of qualitative choice behavior," Tech. Rep. 105-142, Frontiers in Econometric, New York.

Newey, Whitney K and Daniel McFadden (1994): "Large sample estimation and hypothesis testing," *Handbook of econometrics*, 4, 2111–2245.

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

Rivers, Douglas and Quang Vuong (2002): "Model selection tests for nonlinear dynamic models," *The Econometrics Journal*, 5 (1), 1–39.

Silvey, Samuel D (1959): "The Lagrangian multiplier test," *The Annals of Mathematical Statistics*, 30 (2), 389–407.

Vuong, Quang H (1989): "Likelihood ratio tests for model selection and non-nested hypotheses," *Econometrica*, 57 (2), 307–333.

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

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

Windmeijer, Frank (2005): "A finite sample correction for the variance of linear efficient two-step GMM estimators," *Journal of econometrics*, 126 (1), 25–51.