Regression Under Endogeneity: Bernstein-von Mises Theory and Bayes Factors Testing[*]

Siddhartha Chib[†]
Minchul Shin[‡]
Anna Simoni[§]

September 19, 2023

Abstract

A standard assumption in the Bayesian estimation of linear regression models is that the regressors are exogenous (uncorrelated with the error). In practice, however, this assumption can be invalid. In this paper, under the rubric of the exponentially tilted empirical likelihood, we derive the consequences of neglected endogeneity. We derive a Bernstein-von Mises theorem for the posterior distribution of a (default) base model that assumes that the regressors are exogenous when that assumption, in fact, is false. We also develop a Bayes factor test for endogeneity that compares the base model with an extended model that is immune from the problem of neglected endogeneity. We prove that this test is a consistent selection procedure: as the sample becomes large, it almost surely selects the base model if the regressors are exogenous, and the extended model otherwise. The methods are illustrated with simulated data, and problems concerning the causal effect of automobile prices on automobile demand, and the causal effect of potentially endogenous airplane ticket prices on passenger volume.

Keywords: Bayesian inference; Causal inference; Exponentially tilted empirical likelihood; Endogeneity; Exogeneity; Instrumental variables; Marginal likelihood; Posterior consistency.

[*] The views expressed here are our own and do not necessarily represent the views of the Federal Reserve Bank of Philadelphia or the Federal Reserve System.

[†] Olin Business School, Washington University in St. Louis, Campus Box 1133, 1 Bookings Drive, St. Louis, MO 63130. e-mail: chib@wustl.edu.

[‡] Research Department, Federal Reserve Bank of Philadelphia, 10 Independence Mall, Philadelphia, PA 19106, e-mail: visiblehand@gmail.com.

[§] CREST, CNRS, ENSAE, Ecole Polytechnique, Institut Polytechnique de Paris, 5, Avenue Henry Le Chatelier, 91120 Palaiseau - France, e-mail: simoni.anna@gmail.com.

1 Introduction

Consider the semiparametric linear regression model $y = x'\beta + z_1'\gamma + \varepsilon$, where y is the scalar response, $x \in \mathbb{R}^{d_x}$ is the treatment vector of interest, z_1 is a vector of controls and ε is the idiosyncratic noise with an unknown distribution. A standard assumption in the Bayesian estimation of such models is that the regressors x are exogenous (uncorrelated with the error). In many practical applications, however, this assumption is not satisfactory and is likely to be at odds with the data. Provided one has a vector of valid instruments z_2, at least of the same dimension as x, it is possible to develop a prior-posterior analysis of the parameters based on those instruments, from both the parametric and semiparametric Bayesian viewpoints, see for example, Drèze (1976), Kleibergen and van Dijk (1998), Chao and Phillips (1998), Kleibergen and Zivot (2003), Hoogerheide, Kleibergen and van Dijk (2007), Schennach (2005), Liao and Jiang (2011), Florens and Simoni (2012, 2016, 2021), Kato (2013), Shin (2014), and, of particular relevance to the current paper, Chib, Shin and Simoni (2018).

In this paper we contribute to this extensive literature in two dimensions. Under the rubric of the exponentially tilted empirical likelihood, we first derive the consequences of fitting (what we call) the base model from moment conditions that neglect endogeneity. Specifically, given the moment conditions, $\mathbf{E}[\varepsilon(\theta)x] = 0$, $\mathbf{E}[\varepsilon(\theta)z_1] = 0$ and $\mathbf{E}[\varepsilon(\theta)z_2] = 0$, where $\varepsilon(\theta) := (y - x'\beta + z_1'\gamma)$ and $\theta := (\beta', \gamma')'$, we derive a Bernstein-von Mises (BvM) theorem for the posterior of the scaled θ in this model when the assumption of exogeneity is false in the population. This is the first such result in the literature. As a corollary we also derive the limiting posterior distribution of the scaled θ when the exogeneity assumption is true. By comparing these two posterior distributions we show that the consequences of neglecting endogeneity can be severe. This result gives pause to the standard practice of fitting Bayesian regression models under the default assumption of exogeneity.

Another missing element in the existing literature is a test for the exogeneity/endogeneity of the regressors. Such a test would be desirable given the consequences of neglected endogeneity that we document. To fill this gap, we derive the first Bayesian test for the endogeneity of regressors.

This test is based on the marginal likelihood of the base model, and the marginal likelihood of an extended model that has the same z_1 and z_2 orthogonality conditions as the base model, but has the exogeneity condition amended to $\mathbf{E}[\varepsilon(\theta)x] = v$, where v is a vector of new parameters. We show that the log of the Bayes factor (ratio of marginal likelihoods) of the base vs the extended model is a consistent Bayesian test of endogeneity. Specifically, with probability approaching one as the sample size increases, the test picks the base model when x is exogenous, and the extended model when x is endogenous. Interestingly, the idea of comparing two models to detect endogeneity, one which is misspecified under endogeneity and the other which is not, is similar in spirit to the frequentist Hausman (1978) test where the comparison is based on estimators (rather than models) that are inconsistent and consistent under endogeneity. In this sense, the log-Bayes factor test we provide is a Bayesian analogue of the Hausman test for endogeneity.

The rest of the paper is organized as follows. In Section 2 we specify the base model and analyze it when x is endogenous and when it is not. We derive BvM theorems for θ in each case and discuss the consequences of neglected endogeneity. In Section 3 we consider the extended model that is robust to endogeneity of x and derive the corresponding BvM theorem for the augmented parameter $\psi := (\theta', v')'$. In Section 4 we develop our Bayes factor test for endogeneity and establish its large-sample model consistency. In Section 5 we illustrate the practical value of the methodology with real data applications. Concluding remarks are made in Section 6. Proofs of the theorems are collected in the appendix (and an online supplementary appendix).

2 Base model

Consider a random vector $w_i = (y_i, x_i, z_i) \in \mathbb{R}^{d+1}$ from an unknown probability distribution P, where $y_i \in \mathbb{R}$, $x_i \in \mathbb{R}^{d_x}$, $z_i = (z'_{1,i}, z'_{2,i})' \in \mathbb{R}^{d_z}$, $d_z = d_{z_1} + d_{z_2}$ and $d = d_x + d_z$. Suppose that under P, w_i follows the regression model

$$y_i = \beta'_{\circ} x_i + \gamma'_{\circ} z_{1,i} + \varepsilon_i \ , \tag{2.1}$$

where the focus is on the causal effect of x on y, captured by the parameter $\beta_\circ$, and z_1 is a vector of controls. Letting

$$\begin{aligned}\varepsilon_i(\theta) : &= y_i - \beta' x_i - \gamma' z_{1,i} \\ &\equiv y_i - \theta' \tilde{w}_{1,i} \ ,\end{aligned}$$

where $\tilde{w}_{1,i} := (x_i', z_{1i}')'$, $\theta := (\beta', \gamma')' \in \Theta \subset \mathbb{R}^p$ $(p := d_x + d_{z_1})$, $\theta_\circ := (\beta_\circ', \gamma_\circ')'$ and $\mathbf{E}[\cdot] := \mathbf{E}^P[\cdot]$ denotes the expectation with respect to the true distribution P, suppose also that $z_{j,i}$, $j = 1, 2$, satisfy the (exogeneity) restrictions $\mathbf{E}[\varepsilon_i(\theta_\circ) z_{j,i}] = 0$. Assume that the intercept (if any) is contained in z_1. In this set-up, z_2 is a vector of instrumental variables (variables that are correlated with x and do not directly affect the outcome). Under the assumption that $d_{z_2} \geq d_x$, the instruments help to identify $\beta_\circ$ when $\mathbf{E}[\varepsilon_i(\theta_\circ) x_i] \neq 0$.

2.1 Moment restrictions

Now suppose that one has prior information about θ that is summarized in a prior density $\pi(\theta)$, which we assume has positive mass around $\theta_\circ$, for any $\theta_\circ$. Generally, in a default Bayesian analysis, given data $w_{1:n} = \{w_i\}_{i=1}^n$ iid from P, this prior would be updated assuming that ε_i follows a parametric distribution, or an unknown distribution modeled by a Dirichlet process prior with a Gaussian base measure, for example, Chib and Greenberg (2010). Moreover, the assumption that x_i is exogenous would be maintained.

In this paper we consider an even more assumption-light approach and consider updating the prior based solely on moment restrictions, free of additional assumptions about the distribution of ε_i. In the base model, Bayesian updating is made under the assumption that x_i is exogenous, that is, $\mathbf{E}[\varepsilon_i(\theta_\circ) x_i] = 0$. To define the moment restrictions, define the d vector of functions for a generic value of θ

$$g(w_i, \theta) := \varepsilon_i(\theta) \begin{pmatrix} x_i \\ z_{1i} \\ z_{2i} \end{pmatrix}.$$

Then, the moment restrictions can be expressed as

$$\mathbf{E}[g(w_i, \theta_\circ)] = 0. \tag{2.2}$$

Note that the assumption of exogeneity is potentially erroneous. If this is the case, then the moments restrictions (2.2) would be misspecified in the sense that these restrictions are not satisfied for the true P for any value of θ. To benchmark the cost of misspecification, and to develop a test of endogeneity, we specify below in Section 3 an extended model that is immune from this problem.

2.2 ETEL posterior

Suppose now that the data distribution P conditional of θ is given the nonparametric prior in Schennach (2005). Then, following arguments in Schennach (2005), the nonparametric marginal posterior of θ, after marginalization over P, is given by the truncated distribution

$$\pi^n(\theta | w_{1:n}) \propto \pi(\theta)\, \widehat{p}(w_{1:n}|\theta)\, I[\theta \in H], \tag{2.3}$$

where $\pi(\theta)$ is the prior density, $P^n(\theta) = \{\widehat{p}_i(\theta)\}$ is the exponentially tilted likelihood (ETEL) function, and $I[A]$ denotes the indicator function of the event A. Formally, the ETEL function is the discrete probability distribution with support on the atoms w_i, $i = 1, \ldots, n$, that is nearest in the Kullback-Leibler (KL) discrepancy to the empirical $\{\frac{1}{n}\}$ distribution satisfying the moment restrictions, that is,

$$\{\widehat{p}_i(\theta)\} := \arg\max_{p_1,\ldots,p_n} \sum_{i=1}^{n} [-p_i \log(np_i)]$$
$$\text{subject to } \sum_{i=1}^{n} p_i = 1, \text{ and } \sum_{i=1}^{n} p_i g(w_i, \theta) = 0. \tag{2.4}$$

The region of truncation H is given by $H := \{\theta : \sum_{i=1}^{n} \widehat{p}_i(\theta) g(w_i, \theta) = 0\}$. Equivalently, since the maximizer in (2.4) is unique it can be shown that H is the set of θ for which the convex hull of $\bigcup_{i=1}^{n} g(w_i, \theta)$ contains zero. If H is empty, there is no solution in θ to (2.4).

In practice, a convenient way to compute the ETEL function is from the dual of (2.4). In particular, one can derive that

$$\widehat{p}_i(\theta) := \frac{e^{\widehat{\lambda}(\theta)' g(w_i,\theta)}}{\sum_{j=1}^{n} e^{\widehat{\lambda}(\theta)' g(w_j,\theta)}}, \qquad i = 1, \ldots, n \quad , \tag{2.5}$$

where $\widehat{\lambda}(\theta) \equiv \widehat{\lambda}(w_{1:n}, \theta) := \arg\min_{\lambda \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} \exp(\lambda' g(w_i, \theta))$.

The posterior $\pi^n(\theta | w_{1:n})$ is, of course, not in closed form, but it can be summarized by MCMC methods for any n. In the sequel we are interested in the large sample behavior of this posterior distribution when the $\mathbf{E}[\varepsilon(\theta_\circ)x] = 0$ restrictions are assumed but these are restrictions are false. Apparently, there are no results of this type even in a parametric Bayesian set up.

We conclude by defining the population counterpart of $P^n(\theta)$, which is needed for our theoretical work. Let

$$\mathcal{Q}_\theta = \Big\{ Q \in \mathbb{M};\ \mathbf{E}^Q[g(w_i, \theta)] = 0 \Big\} \tag{2.6}$$

denote the set of probability distributions that satisfy the moment conditions for a given θ, where $\mathbb{M}$ is the set of all probability distributions on $\mathbb{R}^{d+1}$. Then, the population counterpart of $P^n(\theta)$ is the element from $\mathcal{Q}_\theta$ that is closest to P in the KL divergence. Denote this best element, for every $\theta \in \Theta$, as

$$Q^*(\theta) := \mathrm{arginf}_{Q \in \mathcal{Q}_\theta} \mathbf{KL}(Q || P) \tag{2.7}$$

where $\mathrm{KL}(Q||P) := \int \log\left(\frac{dQ}{dP}\right) dQ$ is the KL discrepancy of Q from P if Q is absolutely continuous with respect to P; otherwise the discrepancy is equal to $+\infty$. Moreover, when the dual representation of the KL minimization problem (2.7) holds, which is always true if the moments are correctly specified, $Q^*(\theta)$ has a closed form Radon-Nikodyn derivative with respect to P:

$$\frac{dQ^*(\theta)}{dP} = \frac{e^{\lambda_*'(\theta)g(w,\theta)}}{\mathbf{E}[e^{\lambda_*'(\theta)g(W,\theta)}]},$$

where $\lambda_*(\theta) := \arg\min_{\lambda\in\mathbb{R}^d} \mathbf{E}[e^{\lambda' g(w,\theta)}]$, unique by the strict convexity of the integrand in λ.

2.3 Neglected endogeneity

The base model is misspecified if the x-exogenous assumption is false. Let us call this *neglected endogeneity*. Then, there is no value of θ for which the true data generating process P satisfies the moment restriction $\mathbf{E}[g(w_i,\theta)] = 0$, that is, $P \notin \bigcup_{\theta\in\Theta} \mathcal{Q}_\theta$. In this case, we show that the posterior concentrates on a ball centered on the pseudo-true value $\theta_* \neq \theta_\circ$, which we suppose is unique (see Assumption 2.2 below). While we maintain the assumption that $\mathbf{E}[g(w_i,\theta_*)]$ exists, the fact that, for every value of θ, $\mathbf{E}[g(w_i,\theta)] \neq 0$ implies that the tilting parameter $\lambda_*(\theta)$ is nonzero. This becomes a complicating factor in the derivations.

The pseudo-true value is that value of θ that minimizes $KL(P||Q^*(\theta))$, for $Q^*(\theta)$ defined in (2.7). Notice the inversion of the probabilities in the KL discrepancies used to define $Q^*(\theta)$ and θ_*. By definition, $\mathbf{E}^{Q^*(\theta)}[g(w_i,\theta)] = 0$ for every $\theta \in \Theta$, where $\mathbf{E}^{Q^*(\theta)}[\cdot]$ denotes the expectation with respect to $Q^*(\theta)$. The term involving θ in $\mathrm{KL}(P||Q^*(\theta))$ is simply the P expectation of $-\log\frac{dQ^*(\theta)}{dP}$. Thus, when the dual representation holds, we can get the pseudo-true value as

$$\theta_* = \arg\max_{\theta\in\Theta} \mathbf{E}\log\left(\frac{e^{\lambda_*'(\theta)g(w,\theta)}}{\mathbf{E}[e^{\lambda_*'(\theta)g(W,\theta)}]}\right). \tag{2.8}$$

If the moments are misspecified, however, it is possible that $Q^*(\theta)$ and P do not have a common support, for any θ, see Sueishi (2013) for a discussion on this point, in which case, the maximizer in (2.8) would not necessarily equal the θ_* that directly minimizes $\mathrm{KL}(P||Q^*(\theta))$. To avoid this situation, we introduce the following assumption.

Assumption 2.1 (Non-emptyness.) *When $\mathbf{E}[\varepsilon_i x_i] \neq 0$, there exists $Q \in \bigcup_{\theta\in\Theta} \mathcal{Q}_\theta$ such that Q is mutually absolutely continuous with respect to P, where $\mathcal{Q}_\theta$ is defined in (2.6).*

This assumption implies that there is a θ for which $\mathcal{Q}_\theta$ is non-empty and that θ_* is identified by (2.8). We also assume that the pseudo-true value is unique.

Assumption 2.2 (Identification.) *The maximizer θ_* in* (2.8) *is unique and is in the interior of Θ, where the interior is defined with respect to $\mathbb{R}^p$.*

The next assumption is standard when studying the frequentist properties of Bayesian procedures. First, it requires that the prior density of θ is continuous so that $\pi(\theta_* + h/\sqrt{n})$ behaves like the constant $\pi(\theta_*)$ for n large, for very bounded $h \in \mathbb{R}^p$. Second, it requires that θ_* lies in the support of the prior. This is a necessary (though, of course, not sufficient) condition for the posterior distribution to concentrate on θ_* as n becomes large.

Assumption 2.3 *(a) π is a continuous probability measure that admits a density with respect to the Lebesgue measure; (b) π is positive on a neighborhood of θ_*.*

We conclude with three assumptions about the model. Recall the notation $w_i := (y_i, x_i', z_i')'$ and $\tilde{w}_{1,i} := (x_i', z_{1,i}')'$. Moreover, we denote $\tilde{w}_i := (x_i', z_i')'$, $\|\cdot\|_2$ the Euclidean norm and $\|\cdot\|_F$ the Frobenius norm.

Assumption 2.4 *(a) w_i, $i = 1, \dots, n$ are i.i.d. observable random variables each one taking values in a complete probability space $(\mathcal{W}, \mathfrak{B}_\mathcal{W}, P)$, where $\mathcal{W} \subseteq \mathbb{R}^{d+1}$, $\mathfrak{B}_\mathcal{W}$ is the associated σ-field and P is a probability distribution satisfying model* (2.1)*; (b) $\Theta \subset \mathbb{R}^p$ is compact and connected; (c) for every λ in a neighborhood of $\lambda_*(\theta_*)$ the matrix $\mathbf{E}[e^{\lambda' g(w_i,\theta_*)}\varepsilon_i(\theta_*)^2 \tilde{w}_i \tilde{w}_i']$ has smallest (resp. largest) eigenvalue bounded away from zero (resp. infinity).*

Assumption 2.5 *(a) $\mathbf{E}[\|\tilde{w}_i \tilde{w}_{1,i}'\|_F] < \infty$; (b) $\mathbf{E}[\tilde{w}_i \tilde{w}_{1,i}'] < \infty$ with rank p.*

Assumptions 2.4 and 2.5 are standard in the literature, see *e.g.* Schennach (2007). Assumption 2.4 *(c)* guarantees that the asymptotic covariance matrix is invertible. In this paper we show stochastic local asymptotic normality (LAN) of the ETEL function by developing a proof different from the one derived in Chib et al. (2018), see Theorem A.1 in the Appendix. Stochastic LAN is an essential step in the proof of the Bernstein von-Mises theorem (see *e.g.* Van der Vaart (1998)) and it is established based on the following assumption. We denote by $\widetilde{w}_{i,k}$ the k-th component of $\widetilde{w}_i$. Moreover, for any $\delta > 0$ and for some constant $C > 0$ we denote by $B_\delta(\lambda_*(\theta_*)) := \{\lambda \in \mathbb{R}^d; \|\lambda - \lambda_*(\theta_*)\|_2 \leq C\delta\}$ (resp. $B_\delta(\theta_*) := \{\theta \in \mathbb{R}^p; \|\theta - \theta_*\|_2 \leq C\delta\}$) a closed ball centered around $\lambda_*(\theta_*)$ (resp. θ_*) with radius δ.

Assumption 2.6 *(a) For any $\delta > 0$ there is a function $\gamma_0(w_i)$ such that $\left\|e^{\lambda' g(w_i,\theta)} g(w_i, \theta)\right\|_2 \leq \gamma_0(w_i)$ for every $\lambda \in B_\delta(\lambda_*(\theta_*))$ and every $\theta \in B_\delta(\theta_*)$, and $\mathbf{E}[\gamma_0(w_i)] < \infty$; (b) for any $\delta > 0$ there exists a function $\gamma_1(w_i)$ such that $\left|e^{\lambda' g(w_i,\theta)}\right| \leq \gamma_1(w_i)$ for every $\lambda \in B_\delta(\lambda_*(\theta_*))$, every $\theta \in B_\delta(\theta_*)$ and $\mathbf{E}[\gamma_1(w_i)] < \infty$; (c) for $j, \ell, \ell' = 1, 2$, for any $k = 1, \ldots, k$ and every $\delta > 0$ there exists a function $\gamma_2(w_i)$ such that $\left|e^{\lambda' g(w_i,\theta)} \varepsilon_i(\theta)^{j-1} \widetilde{w}_{i,k}^{\ell} (h' \widetilde{w}_{1,i})^{\ell'}\right| \leq \gamma_2(w_i)$ for every $\lambda \in B_\delta(\lambda_*(\theta_*))$, every $\theta \in B_\delta(\theta_*)$ and every h in a compact set, and $\mathbf{E}[\gamma_2(w_i)] < \infty$; (d) the following operator norm*

$$\mathbf{E}\left[\sup_{\lambda \in B_\delta(\lambda_*(\theta_*))} \left\|e^{(\ell-2)\lambda' g(w_i,\theta_*)} \varepsilon_i(\theta_*)^{\ell} \widetilde{w}_{i,k}^{\ell-2} \widetilde{w}_i \widetilde{w}_i'\right\|\right]$$

is bounded away from infinity for every $k = 1, \ldots, d$, any $\delta > 0$ and for $\ell = 3, 4$;
(e) for any $\delta > 0$, $\mathbf{E}\left[\sup_{\lambda \in B_\delta(\lambda_*(\theta_*))} e^{2\lambda' g(w_i,\theta_*)} \varepsilon_i(\theta_*)^2 \|\widetilde{w}_i\|_2^2\right] < \infty$; (f) for every $j, k = 1, \ldots, d$ and every $\delta > 0$ there exists a function $b_{j,k}(w_i)$ such that $\left|e^{\lambda' g_i(w_i,\theta_*)} \widetilde{w}_{i,j} \widetilde{w}_{i,k} \varepsilon_i(\theta_*)^2\right| \leq b_{j,k}(w_i)$ for every $\lambda \in B_\delta(\lambda_*(\theta_*))$ and $\mathbf{E}[b_{j,k}(w_i)] < \infty$; (g) for any $j, k = 1, \ldots, k$ and every $\delta > 0$ there exists a function $\gamma_3(w_i)$ such that $\left|e^{\lambda' g(w_i,\theta)} \varepsilon_i(\theta) \widetilde{w}_{i,k} \widetilde{w}_{i,j} h' \widetilde{w}_{1,i}\right| \leq \gamma_3(w_i)$ for every $\lambda \in B_\delta(\lambda_*(\theta_*))$, every $\theta \in B_\delta(\theta_*)$ and every h in a compact set, and $\mathbf{E}[\gamma_3(w_i)] < \infty$.

Under these assumptions, and an identification-type condition (stated in the next theorem), we are now able to establish that the posterior distribution concentrates on θ_* at the rate $\sqrt{n}$ as the sample size increases. We use the notation $\ell_{n,\theta}(w_i) := \log \widehat{p}_i(\theta)$ for the log-likelihood for one observation w_i.

Theorem 2.1 (Posterior Consistency) *Let Assumptions 2.1 - 2.6 hold. Let*

$$\Theta_n := \{\|\theta - \theta_*\| \leq M_n/\sqrt{n}\},$$

denote a ball around θ_* with radius at most $M_n/\sqrt{n}$, where M_n is any sequence of positive constants diverging to ∞. Assume that there exists a constant $C > 0$ such that

$$P\left(\sup_{\theta \in \Theta_n^c} \frac{1}{n}\sum_{i=1}^{n}\left(\ell_{n,\theta}(w_i) - \ell_{n,\theta_*}(w_i)\right) \leq -\frac{CM_n^2}{n}\right) \to 1 \ , \ \textit{as } n \to \infty. \tag{2.9}$$

Then,

$$\pi\left(\sqrt{n}\|\theta - \theta_*\| > M_n \middle| w_{1:n}\right) \xrightarrow{p} 0 \ , \ \textit{as } n \to \infty \ . \tag{2.10}$$

This theorem states that the posterior concentrates at the parametric rate despite the semiparametric nature of the problem. This result is important in proving the Bernstein-von Mises theorem stated given in Theorem (2.2) below.

Condition (2.9) controls the behaviour of the ETEL function $\theta \mapsto \ell_{n,\theta}(w_i)$ at a distance from θ_* and it ensures that θ_* is well-separated from θs that are at a certain distance from it. A similar condition is in Kleijn and van der Vaart (2012, Lemma 4.2) and it is also related to the classical condition in *e.g.* Lehmann and Casella (1998, Assumption 6.B.3) and Chernozhukov and Hong (2003, Assumption 3). To better understand the meaning of this assumption, note that asymptotically the log-ETEL function is maximized at the pseudo-true value θ_*. Hence, Assumption (2.9) requires that if the parameter θ is far from the pseudo-true value θ_*, that is $\|\theta - \theta_*\| > M_n/\sqrt{n}$, then $\sum_{i=1}^n \ell_{n,\theta}(w_i)$ evaluated at such θ has to be small relative to the close to the maximum value $\sum_{i=1}^n \ell_{n,\theta_*}(w_i)$. Controlling this behavior is important because the posterior involves integration over the whole support of θ. Subsets of Θ that can be distinguished from θ_* uniformly (with probability approaching 1 as $n \to \infty$) based on the ETEL function will receive a posterior probability that is asymptotically negligible. An alternative to this condition would be to require the existence of asymptotically consistent tests ϕ_n that are able to distinguish from the true distribution P in a uniform way, that is, for every $\epsilon > 0$ there exists a sequence of tests $\{\phi_n\}$ such that as $n \to 0$,

$$\mathbf{E}[\phi_n] \to 0, \qquad \text{and} \qquad \sup_{\{\theta; \|\theta-\theta_*\| \geq \epsilon\}} \mathbf{E}\left[e^{\ell_{n,\theta}(w_i)-\ell_{n,\theta_*}(w_i)}(1-\phi_n)\right] \to 0. \tag{2.11}$$

2.4 BvM theory

Since Theorem (2.1) establishes that the posterior concentrates in neighbourhoods of θ_* of size decreasing at the rate $n^{-1/2}$, then we now give a BvM result for the sequence of posterior distributions of the local parameter $h := \sqrt{n}(\theta - \theta_*)$ as the sample size diverges to infinity. We are assuming in this result that endogeneity has been neglected. Thus, this is the BvM result under misspecification of some moments. We denote by $\dot{\ell}_{n,\theta_*}$ the first derivative of $\theta \mapsto \ell_{n,\theta}(w_{1:n})$ evaluated at θ_* and by $\ddot{\lambda}_{*,j}(\theta_*)$ the d-matrix of second derivatives of $\theta \mapsto \lambda_{*,j}(\theta)$ evaluated at θ_*, where $\lambda_{*,j}(\theta)$ denotes the j-th component of the d vector $\lambda_*(\theta)$.

Theorem 2.2 (Bernstein-von Mises under neglected endogeneity) *Assume that the conditions of Theorem 2.1 hold. Then, the sequence of posteriors converge in total variation towards a Normal distribution, that is,*

$$\sup_B \left| \pi(\sqrt{n}(\theta - \theta_*) \in B | w_{1:n}) - \mathcal{N}_{\Delta_{n,\theta_*}, V_{\theta_*}}(B) \right| \xrightarrow{p} 0, \tag{2.12}$$

where $B \subseteq \Theta$ is any Borel set, $\Delta_{n,\theta_*} := \frac{1}{\sqrt{n}} \sum_{i=1}^n V_{\theta_*} \dot{\ell}_{n,\theta_*}(w_i) + o_p(1) = \mathcal{O}_p(1)$, and V_{θ_*} is a positive definite matrix equal to the inverse of:

$$\begin{aligned} V_{\theta_*}^{-1} = \mathbf{E}^{Q^*(\theta_*)}\left[\tilde{w}_{1,i}\tilde{w}_i'(I + \lambda_*(\theta_*)\varepsilon_i\tilde{w}_i')\right] \left(\mathbf{E}^{Q^*(\theta_*)}\left[\varepsilon_i^2 \tilde{w}_i \tilde{w}_i'\right]\right)^{-1} \\ \times \left(2\mathbf{E}[\tilde{w}_i\tilde{w}_{1,i}'] - \mathbf{E}^{Q^*(\theta_*)}\left[\tilde{w}_{1,i}\tilde{w}_i'(I + \lambda_*(\theta_*)\varepsilon_i\tilde{w}_i')\right]\right) \\ - \sum_{j=1}^{d_x} \ddot{\lambda}_{*,j}(\theta_*)\mathbf{E}[\varepsilon_i x_{i,j}] + \mathbb{V}ar_{Q^*(\theta_*)}[\tilde{w}_{1,i}\tilde{w}_i'\lambda_*(\theta_*)], \end{aligned}$$

where $\mathbb{V}ar_{Q^*(\theta_*)}$ denotes the variance taken with respect to the distribution $Q^*(\theta_*)$.

To prove this theorem, in addition to the posterior consistency result stated in Theorem 2.1, we need to show that $\sum_{i=1}^{n}\ell_{n,\theta}(w_i)$ satisfies a stochastic Local Asymptotic Normality (LAN) expansion around the pseudo-true value θ_*. This is shown in the Online Appendix by using a proof that differs from the proof used in Chib et al. (2018) to establish this result. The stochastic LAN expansion means that the loglikelihood ratio $\sum_{i=1}^{n}\ell_{n,\theta}(w_i)-\ell_{n,\theta_*}(w_i)$ evaluated at a random local parameter around the pseudo-true value, is well approximated by a quadratic form.

Theorem 2.2 establishes that as $n\to\infty$ the posterior distribution of $\sqrt{n}(\theta-\theta_*)$ has a shape more and more similar to the one of a Gaussian distribution. The mean of the Gaussian limiting distribution is given by a bounded-in-probability stochastic series Δ_{n,θ_*} which converges in distribution to a zero-mean Gaussian random variable. As it has been shown in Chib et al. (2018), $\Delta_{n,\theta_*}=\sqrt{n}(\widehat{\theta}-\theta_*)+o_P(1)$ with $\widehat{\theta}$ the frequentist ETEL estimator of Schennach (2007). Therefore, Theorem 2.2 can be formulated as in the classical formulation with centering of the limiting distribution $\sqrt{n}(\widehat{\theta}-\theta_*)$.

The asymptotic variance of the posterior distribution of $\sqrt{n}(\theta-\theta_*)$ involves many terms which are non-zero because in the misspecified case the tilting parameter $\lambda_*(\theta_*)$ evaluated at the pseudo-true value is non-zero. To understand how the misspecification affects the asymptotic variance of the posterior distribution, we provide in the corollary below the Bernstein-von Mises theorem for the case when the exogeneity assumption holds. In this case, we denote by $\theta_\circ$ the true value of θ defined as the value that satisfies the moment restriction $\mathbf{E}[g(w_i,\theta_\circ)]=0$. Since $P\in\bigcup_{\theta\in\Theta}\mathcal{Q}_\theta$ then, $\theta_\circ$ is equivalently defined as

$$\theta_\circ=\arg\max_{\theta\in\Theta}\mathbf{E}\log\left(\frac{e^{\lambda_*(\theta)'g(w,\theta)}}{\mathbf{E}[e^{\lambda_*(\theta)'g(W,\theta)}]}\right). \tag{2.13}$$

Corollary 2.1 (Bernstein-von Mises under exogeneity) *Let $\theta_\circ$ denote the true value of θ and let Assumptions 2.2 - 2.6 hold with θ_* replaced by $\theta_\circ$, $\lambda_*(\theta_\circ)$ replaced by zero, and the matrix in Assumption 2.4 (c) replaced by the matrix $\mathbf{E}[\varepsilon_i(\theta_\circ)^2\widetilde{w}_i\widetilde{w}_i']$. Assume that there exists a constant*

$C > 0$ *such that as* $n \to \infty$,

$$P\left(\sup_{\|\theta-\theta_\circ\|>M_n/\sqrt{n}} \frac{1}{n}\sum_{i=1}^{n}\left(\ell_{n,\theta}(w_i) - \ell_{n,\theta_\circ}(w_i)\right) \leq -\frac{CM_n^2}{n}\right) \to 1, \qquad (2.14)$$

where M_n *is any sequence of positive constants diverging to infinity. Then the posteriors converge in total variation towards a Normal distribution, that is,*

$$\sup_{B}\left|\pi(\sqrt{n}(\theta-\theta_\circ)\in B|w_{1:n}) - \mathcal{N}_{\Delta_{n,\theta_\circ},V_{\theta_\circ}}(B)\right| \xrightarrow{p} 0, \qquad (2.15)$$

where $B \subseteq \Theta$ *is any Borel set,* $\Delta_{n,\theta_\circ} := \frac{1}{\sqrt{n}}\sum_{i=1}^{n} V_{\theta_\circ}\mathbf{E}\left[\widetilde{w}_{1,i}\widetilde{w}_i'\right]\left(\mathbf{E}[\varepsilon_i^2\widetilde{w}_i\widetilde{w}_i]\right)^{-1}\varepsilon_i\widetilde{w}_i$*, and* $V_{\theta_\circ}$ *is the inverse of* $V_{\theta_\circ}^{-1} := \mathbf{E}\left[\widetilde{w}_{1,i}\widetilde{w}_i'\right]\left(\mathbf{E}[\varepsilon_i^2\widetilde{w}_i\widetilde{w}_i']\right)^{-1}\mathbf{E}\left[\widetilde{w}_i\widetilde{w}_{1,i}'\right]$.

The main difference between Corollary 2.1 and Theorem 2.2 concerns the expression for the asymptotic variance: V_{θ_0} has a simplified expression compared to V_{θ_*}. This simplification is due to the fact that when the model is correctly specified, that is, when the exogeneity assumption holds, then the tilting parameter λ_* evaluated at the true $\theta_\circ$ is equal to zero: $\lambda_*(\theta_\circ) = 0$. This also implies that $Q^*(\theta_*) = P$. Moreover, $\mathbf{E}[\varepsilon_i\widetilde{w}_i] = 0$. Condition (3.2) has a similar interpretation as (2.9) in the neglected endogeneity case. Because here the moment restrictions are all correctly specified, then the alternative condition based on tests requires that for every $\epsilon > 0$ there exists a sequence of tests $\{\phi_n\}$ such that as $n \to 0$,

$$\mathbf{E}[\phi_n] \to 0, \qquad \text{and} \qquad \sup_{\{\theta;\|\theta-\theta_\circ\|\geq\epsilon\}} \mathbf{E}_n\left[e^{\ell_{n,\theta}(w_i)}(1-\phi_n)\right] \to 0. \qquad (2.16)$$

Contrarily to the case with neglected endogeneity, in this case where one correctly assumes that the covariates x are exogenous, the posterior distribution has the same asymptotic variance as the efficient Generalized Method of Moments (see Hansen (1982) and Chamberlain (1987)). In addition, the fact of using the moment restrictions $\mathbf{E}[\varepsilon_i x_i] = 0$ to construct the posterior distribution allows to reduce the asymptotic variance when x_i is exogenous with respect to the analysis made by using only the moment restrictions $\mathbf{E}[\varepsilon_i z_i] = 0$. Indeed, the difference between the asymptotic precision of the base model $\mathbf{E}[g(w_i, \theta_\circ)] = 0$ and the one of the model $\mathbf{E}[\varepsilon_i z_i] = 0$ is

$$a'\Big(\mathbf{E}[\varepsilon_i^2 x_i x_i'] - \mathbf{E}[\varepsilon_i^2 x_i z_i']\mathbf{E}[\varepsilon_i^2 z_i z_i']^{-1}\mathbf{E}[\varepsilon_i^2 z_i x_i']\Big)^{-1} a,$$

which is positive definite, where $a' := (\mathbf{E}[\widetilde{w}_{1,i} x_i'] - \mathbf{E}[\widetilde{w}_{1,i} z_i']\mathbf{E}[\varepsilon_i^2 z_i z_i']^{-1}\mathbf{E}[\varepsilon_i^2 z_i x_i'])$. We now provide an illustrating example.

Example 1 *To illustrate the consequences of fitting the base model under misspecification, consider the following data generating process (DGP):*

$$\begin{aligned} y_i &= \gamma_0 + x_i\,\beta + z_{1i}\,\gamma_1 + \varepsilon_i \\ x_i &= \delta_0 + z_{1i}\,\delta_1 + z_{2i}\,\delta_2 + u_i \\ z_{1i} &= v_i \\ z_{2i} &= w_i \end{aligned}$$

for $i = 1, \ldots, n$, where $n \in \{250, 500, 1000, 2000\}$. Suppose that the (u_i, v_i, w_i) are marginally Gaussian, that ε_i is marginally a skewed Gaussian mixture $0.5\mathcal{N}(0.5, 0.5^2) + 0.5\mathcal{N}(-0.5, 1.118^2)$, that $(\varepsilon_i, u_i, v_i)$ have a joint distribution induced by a Gaussian copula with covariance matrix $R = \left(\begin{smallmatrix} 1 & 0.6 & 0 \\ 0.6 & 1 & 0 \\ 0 & 0 & 1 \end{smallmatrix}\right)$ and that the covariance of w_i with each of the other errors is zero. Also assume that each parameter is one (except for δ_1, which is .5). Under this DGP, z_{1i} is correlated with ε_i and x_i but since w_i is uncorrelated with the other shocks, z_{2i} is a valid instrument that is also relevant for x_i. For each of the four sample sizes, the posterior of $\theta := (\beta, \gamma_0, \gamma_1)$ is calculated from the four moments

$$\mathbf{E}\Big[(y_i - x_i\,\beta - \gamma_0 - z_{1i}\,\gamma_1)\begin{pmatrix} x_i \\ 1 \\ z_{1i} \\ z_{2i} \end{pmatrix}\Big] = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}.$$

The second moment is counter to fact (this is the problem of neglected endogeneity). The ETEL posterior is sampled by the tailored one-block M-H algorithm for 20000 iterations beyond a burn-in of a 1000 cycles. The marginal posterior density of β for each sample size is computed from these MCMC sampled draws. Kernel smoothed versions of the posterior densities are given in Figure 1. As the sample size increases, the posterior concentrates on a value quite different from the true value of β.

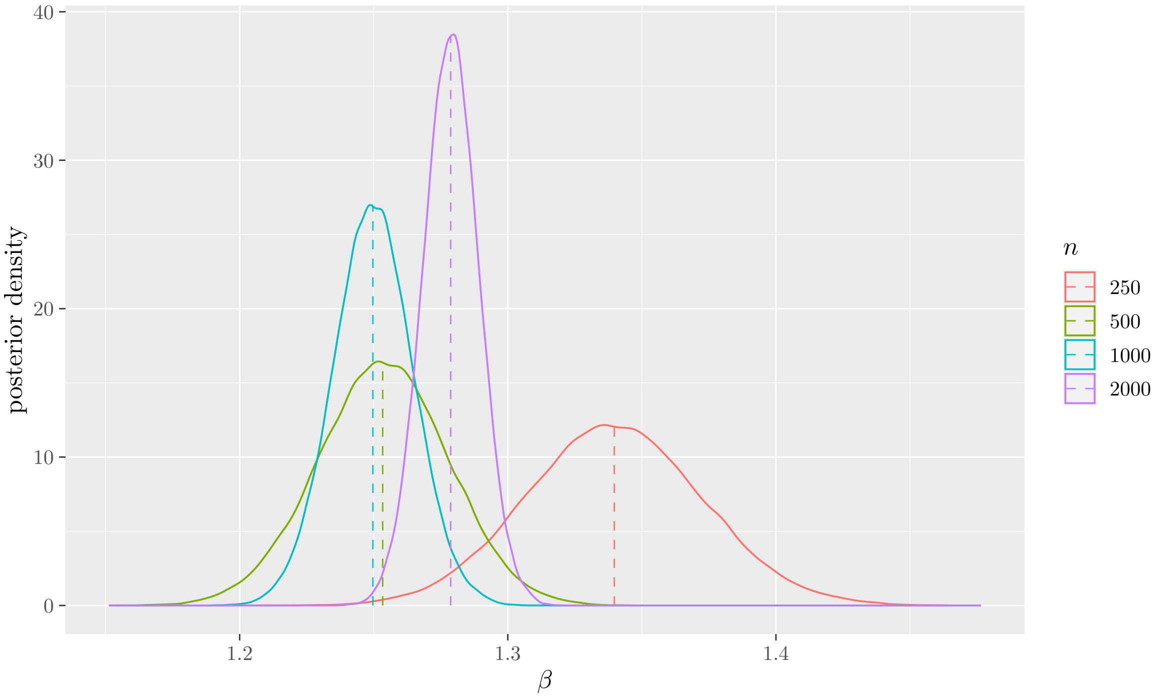

Figure 1: Base model under neglected endogeneity: Marginal posterior densities of β for different sample sizes. Posterior mean is indicated by dashed vertical line.

3 Extended model

We now consider an extended version of the base model that is immune from the problem of neglected endogeneity. One way to define such a model is by omitting the x moment conditions altogether. This has two problems, however. First, if x is exogenous, the asymptotic variance of the parameters is larger than from the base model, as we have shown above. Second, it is not possible to develop a test of endogeneity by omitting the x restrictions. Therefore, it is necessary to define an extended model that achieves immunity from the problem of neglected endogeneity, but does not omit the x restrictions.

Let $v \in \mathcal{V} \subseteq \mathbb{R}^{d_x}$ be a nuisance parameter of the same dimension as x. Now let

$$g(w_i, \theta, v) := \varepsilon_i(\theta) \begin{pmatrix} x_i \\ z_{1i} \\ z_{2i} \end{pmatrix} - \begin{pmatrix} v \\ 0 \\ 0 \end{pmatrix} \quad ,$$

where v serves the purpose of inactivating the x restrictions. The extended model can now be defined by the moment restrictions

$$\mathbf{E}[g(w_i, \psi_\circ)] = 0,$$

where $\psi := (\theta', v')' \in \Psi$ is the vector θ augmented by v, $\Psi := \Theta \times \mathcal{V}$ and $\psi_\circ$ denotes the true value of ψ. This model is correctly specified under both exogeneity and endogeneity of x_i. Importantly, this model can be compared with the base model to develop a test of endogeneity, as we show in the next section.

The ETEL posterior $\pi(\psi|w_{1:n})$ emerges in the same way as above with θ now replaced by ψ, the moment functions $g(x, \theta)$ replaced by $g(x, \psi)$, the prior $\pi(\theta)$ extended to $\pi(\psi) = \pi(\theta)\pi(v)$ and the parameter space Θ enlarged to Ψ. The true value of ψ is denoted by $\psi_\circ$ and is defined similarly to $\theta_\circ$ as the value that satisfies the moment restriction $\mathbf{E}[g(w_i, \psi_\circ)] = 0$ or equivalently as

$$\psi_\circ = (\theta_\circ, v_\circ) = \arg\max_{\psi \in \Psi} \mathbf{E} \log\left(\frac{e^{\lambda_*(\psi)' g(w,\psi)}}{\mathbf{E}\left[e^{\lambda_*(\psi)' g(w,\psi)}\right]} \right) = \arg\max_{\psi \in \Psi} \mathbf{E} \log\left(\frac{e^{\lambda_*(\psi)' g(w,\theta)}}{\mathbf{E}\left[e^{\lambda_*(\psi)' g(w,\theta)}\right]} \right). \tag{3.1}$$

Since the extended model is always correctly specified, a result similar to the one in Corollary 2.1 can be shown to hold. We have the following BvM result.

Theorem 3.1 (Bernstein-von Mises) *Let $\psi_\circ = (\theta_\circ, v_\circ)$ denote the true value of ψ and let Assumptions 2.2 - 2.6 hold with θ_* replaced by $\psi_\circ$, $\lambda_*(\psi_\circ)$ replaced by zero, and the matrix in Assumption 2.4 (c) replaced by the matrix $\mathbf{E}[\varepsilon_i(\psi_\circ)^2 \widetilde{w}_i \widetilde{w}_i']$. Assume that there exists a constant $C > 0$ such that* *as $n \to \infty$,*

$$P\left(\sup_{\|\psi-\psi_\circ\|>M_n/\sqrt{n}} \frac{1}{n}\sum_{i=1}^{n}\Big(\ell_{n,\psi}(w_i)-\ell_{n,\psi_\circ}(w_i)\Big) \leq -\frac{CM_n^2}{n}\right) \to 1, \tag{3.2}$$

where M_n is any sequence of positive constants diverging to infinity. Then the posteriors converge in total variation towards a Normal distribution, that is,

$$\sup_B \left|\pi^n(\sqrt{n}(\psi-\psi_\circ)\in B|w_{1:n}) - \mathcal{N}_{\Delta_{n,\psi_\circ},V_{\psi_\circ}}(B)\right| \xrightarrow{p} 0, \tag{3.3}$$

where $B \subseteq \Psi$ is any Borel set, $\Delta_{n,\psi_\circ} = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} V_{\psi_\circ}\mathbf{E}\left[\frac{dg(w_i,\psi_\circ)'}{d\psi}\right]\Omega_{\psi_\circ}^{-1}(\varepsilon_i\widetilde{w}_i - \widetilde{v}_\circ)$, with $\widetilde{v}_\circ := (v_\circ', 0', 0')'$ and $v_\circ := \mathbf{E}[\varepsilon_i x_i]$, and $V_{\psi_\circ}$ is the inverse of

$$V_{\psi_\circ}^{-1} = \mathbf{E}\left[\frac{dg(w_i,\psi_\circ)'}{d\psi}\right]\Omega_{\psi_\circ}^{-1}\mathbf{E}\left[\frac{dg(w_i,\psi_\circ)}{d\psi'}\right]$$

with $\Omega_{\psi_\circ} := \mathbf{E}\left[g(w_i,\psi_\circ)g(w_i,\psi_\circ)'\right] = \mathbf{E}[\varepsilon_i^2\widetilde{w}_i\widetilde{w}_i'] - \widetilde{v}_\circ\widetilde{v}_\circ'$ and $\mathbf{E}\left[\frac{dg(w_i,\psi_\circ)'}{d\psi}\right] = -\begin{pmatrix}\mathbf{E}[\widetilde{w}_{1i}x_i'] & \mathbf{E}[\widetilde{w}_{1i}z_i'] \\ I_{d_x} & 0\end{pmatrix}$.

Note that both Corollary 2.1 and Theorem 3.1 refer to situations where the model is correctly specified. There is an important difference, however. In the latter case, correctness of the model is achieved by inactivating the x-moment conditions when this is incorrect.

Remark 3.1 *Let us analyze the asymptotic precision matrix $V_{\psi_\circ}^{-1}$. It is interesting to notice that in the case where x is exogenous, then $v_\circ = 0$ and the joint distribution of the first p components of the limiting random vector has variance equal to the asymptotic variance in Corollary 2.1. On the contrary, in the case where x is endogenous and then $v_\circ \neq 0$, the variance of the joint distribution of the first p components of the limiting random vector differs from the asymptotic variance in Theorem 2.2. A proof of this is provided in the Supplementary Material.*

Example 1 (continued) *Consider again the fitting of the data that follows the DGP given in Ex-*

ample 1. The extended (correctly specified) moment restricted model is

$$\mathbf{E}\Big[(y_i - x_i\,\beta - \gamma_0 - z_{1i}\,\gamma_1)\begin{pmatrix} x_i \\ 1 \\ z_{1i} \\ z_{2i} \end{pmatrix}\Big] = \begin{pmatrix} v \\ 0 \\ 0 \\ 0 \end{pmatrix}$$

The parameter of interest is now $\psi = (\beta, \gamma_0, \gamma_1, v)$. We use a default student-t prior on v centered at the GMM estimate and spread given by 4 times the GMM asymptotic variance. The prior of θ is the same as in the base model. The ETEL posterior for each of the four different sample sizes is sampled by the tailored one block M-H method for 20000 iteratations beyond a burn-in of 1000 cycles. The marginal posterior densities of β are given in Figure 2. One can see that the posterior of β, even for $n = 250$, is close to the true value of β, and, for $n = 2000$, is essentially centered around the true value.

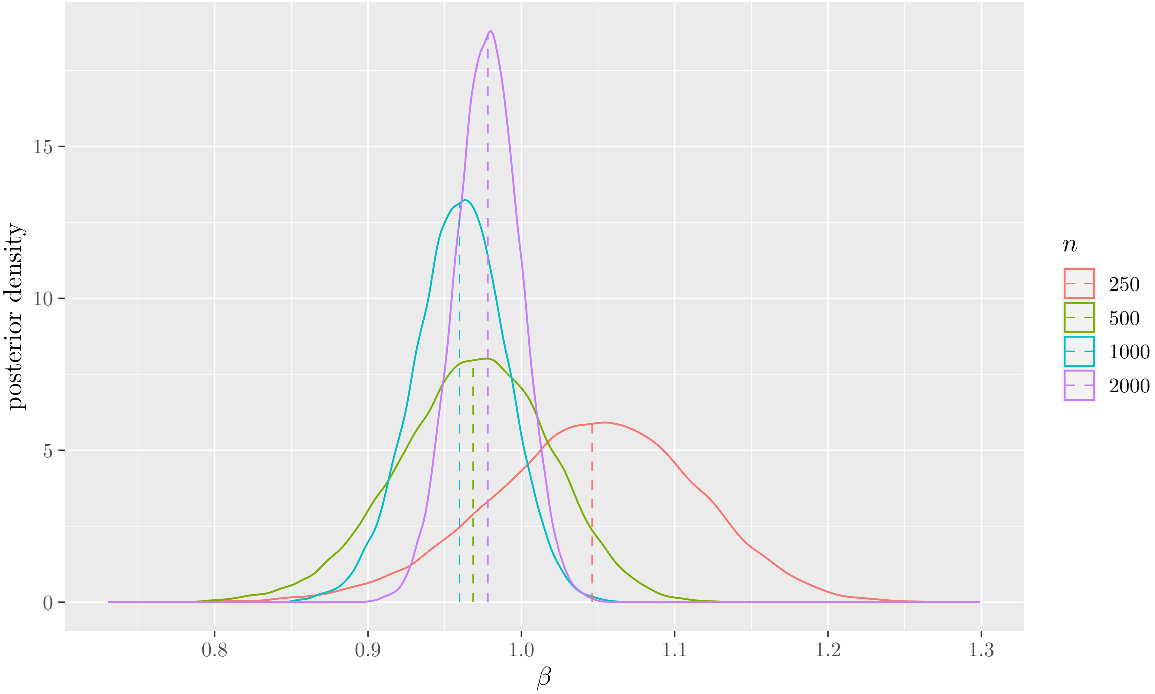

Figure 2: Extended model (x_i moment is inactive): Marginal posterior densities of β for different sample sizes. Posterior mean is indicated by dashed vertical line.

4 Bayes factor testing

We now show that we can develop a Bayesian test of endogeneity by comparing the base and extended models in terms of marginal likelihoods. According to the theory in Chib et al. (2018), for valid comparisons of moment condition models, the contending models must arise from a common encompassing model and should have the same number of moment conditions. We have ensured that this condition is met by including the $\mathbf{E}[\varepsilon_i(\theta)z_{2,i}] = 0$ restriction in the base model, and not excluding the $\mathbf{E}[\varepsilon_i(\theta)x_i] = v$ condition from the extended model.

To describe the test, define the base and extended models of sections 2 and 3 as:

$$\begin{aligned} &\mathcal{M}_b : \qquad && \mathcal{Q}_\theta := \left\{ Q \in \mathbb{M}; \mathbf{E}^Q[g(w_i,\theta)] = 0 \right\}, \qquad && \theta \in \Theta \\ &\mathcal{M}_e : \qquad && \mathcal{Q}_{e,(\theta,v)} := \left\{ Q \in \mathbb{M}; \mathbf{E}^Q[g(w_i,\theta,v)] = 0 \right\}, \qquad && \theta \in \Theta, v \in \mathbb{R}^{d_x} \end{aligned} \tag{4.1}$$

where $\mathbb{M}$ is the set of all probability distributions on $\mathbb{R}^{d+1}$. In addition, as above, define, for every $\theta \in \Theta$, the best element of $\mathcal{Q}(\theta)$ (best in terms of closeness to P) as

$$Q^*(\theta) := \text{arginf}_{Q \in \mathcal{Q}_\theta} \text{KL}(Q||P) \tag{4.2}$$

and, for every $\psi \in \Theta \times \mathcal{V}$, the best element of $\mathcal{Q}_{e,\psi}$ as

$$Q_e^*(\psi) := \text{arginf}_{Q \in \mathcal{Q}_{e,\psi}} \text{KL}(Q||P).$$

Our Bayesian test of endogeneity is given by the Bayes factor of $\mathcal{M}_e$ versus $\mathcal{M}_b$

$$BF_{eb} = \frac{m(w_{1:n}|\mathcal{M}_e)}{m(w_{1:n}|\mathcal{M}_b)} \quad ,$$

where $m(w_{1:n}|\mathcal{M}) = \int \widehat{p}(w_{1:n}|\mathcal{M},\theta)\pi(\theta)d\theta$, for $\mathcal{M} \in \{\mathcal{M}_b,\, \mathcal{M}_e\}$, are the model marginal likelihoods. In our work, we compute these by the method of Chib (1995), as extended to general M-H chains in Chib and Jeliazkov (2001).

Intuitively, to see why this test works, note that $\mathcal{M}_b$ is correctly specified when x is exogenous and misspecified when x is endogenous; however, $\mathcal{M}_e$ is correctly specified in both the cases. Therefore, from Chib et al. (2018), it follows that $\mathcal{M}_b$, which has $(d - p)$ overidentifying restrictions, versus M_e, which has $(d - p - d_x)$ overidentifying restrictions, would be preferred by the Bayes factor when x is exogenous (because it has more overidentifying restrictions than $\mathcal{M}_e$), whereas, $\mathcal{M}_e$ would be preferred when x is endogenous (because $\mathcal{M}_b$ in that case would be misspecified).

The next theorem, which extends Chib et al. (2018) to the particular case considered here, establishes consistency of this Bayes factor test, that is, as the sample size increases, model $\mathcal{M}_b$ is selected when x is exogenous and model $\mathcal{M}_e$ when x is endogenous with probability approaching one.

Theorem 4.1 *Let the Assumptions of Theorems 2.2 and 3.1, and of Corollary 2.1 hold. Let us consider the comparison of models $\mathcal{M}_b$ and $\mathcal{M}_e$ in* (4.1)*. Then,*

$$\lim_{n\to\infty} P\left(\log m(w_{1:n}|\mathcal{M}_e) > \log m(w_{1:n}|\mathcal{M}_b)\right) = 1$$

if and only if there is no θ such that $\mathbf{E}[\varepsilon_i(\theta)x_i] = 0$ holds, and the limit is zero otherwise.

As we show in the proof of this theorem, the failure of the necessary and sufficient condition $\mathbf{E}[\varepsilon_i(\theta)x_i] = 0$ for any θ, is equivalent to the inequality $\mathrm{KL}(P||Q_e^*(\psi)) < \mathrm{KL}(P||Q^*(\theta))$. Thus, as in the general result in Chib et al. (2018, Theorem 3.2) for moment condition models, comparing the log marginal likelihoods of the base and extended models, and selecting the one with the higher value, in the limit, selects the model that is closest in the KL divergence to the true model.

Example 1 (continued) *In the same generating process as Example 1, suppose that $(\varepsilon_i, u_i, v_i)$ have a joint distribution induced by a Gaussian copula with covariance matrix $R = \begin{pmatrix} 1 & \rho & 0 \\ \rho & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$. The parameter ρ controls the degree of endogeneity. We let ρ take values in the set from -.5 to .5, in increments of 0.1. For each value of ρ in this set, we generate 100 samples of size n. For each sample, we compute the the base and extended models, and calculate the log-marginal likelihoods.*

We then count the number of times the log marginal likelihood of $\mathcal{M}_e$ exceeds that of $\mathcal{M}_b$. The results are given Table 1. We can see from this table that even for small values of ρ, our test of endogeneity correctly concludes that the correct model is $\mathcal{M}_e$. In contrast, an incorrect test of endogeneity that compares the base without the z_2 restriction with the extended model without the x condition produces completely erroneous results, with the latter model selected 100% of the times, even when $\rho = 0$.

ρ	-0.5	-0.4	-0.3	-0.2	-0.1	0.0	0.1	0.2	0.3	0.4	0.5
$n = 250$	100	97	73	25	6	1	6	37	76	96	100
$n = 500$	100	100	93	64	8	0	17	72	99	100	100
$n = 1000$	100	100	100	91	28	1	35	98	100	100	100
$n = 2000$	100	100	100	100	65	1	59	100	100	100	100

Table 1: Model selection frequencies from 100 replications of data simulated from the design in Example 1. For each combination of n and $\mathrm{Cov}(\varepsilon, u) = \rho$, the entries give the number of times in 100 replications of the data that the log-marginal likelihood of $\mathcal{M}_e$ exceeds the log-marginal likelihood of $\mathcal{M}_b$.

5 Real data examples

5.1 Causal effect of price on automobile demand

We consider the classic problem of automobile demand dealt in Berry, Levinsohn and Pakes (1995). This problem has recently been revisited by Chernozhukov, Hansen and Spindler (2015), henceforth BLP and CHS, respectively. Apart from its intrinsic value, this problem is worth analyzing because it involves a realistically large number of controls and instruments.

To set up the problem, let y_{ijt} denote the log of the ratio of the market share of product i in market j at time t, relative to an external option, and let x_{ijt} denote the potentially endogenous automobile price variable. In the sample data, this variable is demeaned. For controls, let z_{ijt} denote the observed characteristics of the product. In BLP these are taken to be a constant, an air conditioning dummy (air), horsepower divided by weight ($hpwt$), miles per dollar (mpd), and vehicle size ($space$). In our notation, $y_{ijt} = x_{ijt}\beta + z'_{1ijt}\gamma + \varepsilon_i$, where $z_{1ijt} = (1, mpd_{ijt}, space_{ijt}, hpwt_{ijt}, air_{ijt})$. BLP used ten instruments, five formed by summing the value of these five characteristics over other automobiles produced by the same firm and five formed by summing the above characteristics over automobiles produced by other firms. These form z_{2ijt}. In revisiting this analysis, CHS augment the original controls with quadratics, and cubics in $trend$, mpd, $space$, $hpwt$, and all first order interactions, and then used sums of these characteristics as potential instruments.

In our analysis, we consider both formulations, but in the augmented variant we introduce nonlinear controls by transforming each of $trend$, $hpwt$, mpd and $space$ by natural cubic spline basis functions, each centered at five equally spaced quantile knots (the cubic spline basis functions are taken from Chib and Greenberg (2010)). We opt for this approach to avoid widely different covariate values from parametric quadratic and cubic terms of these covariates. After the imposition of an identification restriction on the basis expansions, which reduces the number of nonlinear terms to four for each continuous covariate, the RHS of the augmented outcome model is defined by x (price) and z_1 (consisting of an intercept, sixteen nonlinear covariates, denoted by $trend_Bj$, mpd_{Bj}, $space_{Bj}$ and $hpwt_{Bj}$, for $j = 1, \ldots, 4$, and the air-conditioning dummy). The set of augmented instruments that form z_2 in this augmented model are then constructed as in BLP.

We fit four models to these data: the base and extended models under the controls and instruments in BLP, and the base and extended models under the augmented set of controls and instruments. In the BLP version, the base and extended models contain six and seven parameters, respectively, estimated with the help of ten instruments, while in the augmented variant, the base and extended models have nineteen and twenty parameters, respectively, estimated from 53 moment restrictions. The $n = 2217$ observations on $(y_{ijt}, x_{ijt}, z_{1ijt})$ are assumed to be a random sample from the population of automobile products across markets and time. Because it is difficult to fix priors on the parameters by a priori considerations, we randomly select 15% of the sample to make training sample priors. In particular, we used the GMM estimate and its standard error fitted on the training data (model by model) as the prior mean and twice the GMM standard error

	Original BLP (Linear)	Augmented BLP (Nonlinear)
Base model (price is exogenous)	-14386.81	-14431.86
Extended model (price is endogenous)	-14364.59	-14397.67

Table 2: Results from the proposed Bayesian test of endogeneity. The log marginal likelihoods for the base and extended models under the original BLP model and its augmented variant. Results based on a training sample prior (using randomly selected 15% of the data) and 10,000 MCMC iterations (beyond a burn-in of 1000) of a tailored single block M-H algorithm. Logarithm of marginal likelihoods are computed by the method of Chib (1995) and Chib and Jeliazkov (2001).

as the prior standard deviation. Each model is fit with these priors and the ETEL constructed from the remaining data by the single block M-H algorithm of Chib and Greenberg (1995). We find that despite the relatively large numbers of parameters and instruments, this algorithm is both fast and efficient. The results show that the posterior mean of the coefficient on $price$ is -0.14, with a 95$\%$ posterior credibility interval running from -.16 to -.13. The posterior mean is larger in magnitude than the OLS estimate originally reported by BLP. Note that the posterior distribution of the augmentation parameter, v, is concentrated to the right of zero, indicating that the $price$ is likely endogenous.

For confirmation, we turn to our formal test of endogeneity. The results are reported in Table 2. We can see that the marginal likelihood is larger for the extended models in both the original BLP and the augmented BLP specifications, supporting the conclusion that price is endogenous.

We conclude this analysis by plotting the posterior distributions of the price coefficient from each model. The estimated effect of price on automobile demand is larger (in absolute value) when endogeneity of price is taken into account. Interestingly, the price effect is smaller and more concentrated in the augmented models, suggesting that some of the excess sensitivity to price observed in the orginal BLP model is due to the omission of the nonlinear controls. In addition, it is worth noting that if we were to only fit the base model (which the marginal likelihood confirms is misspecified in this case) we would miss the fact that incorporating nonlinearities impacts the posterior distribution.

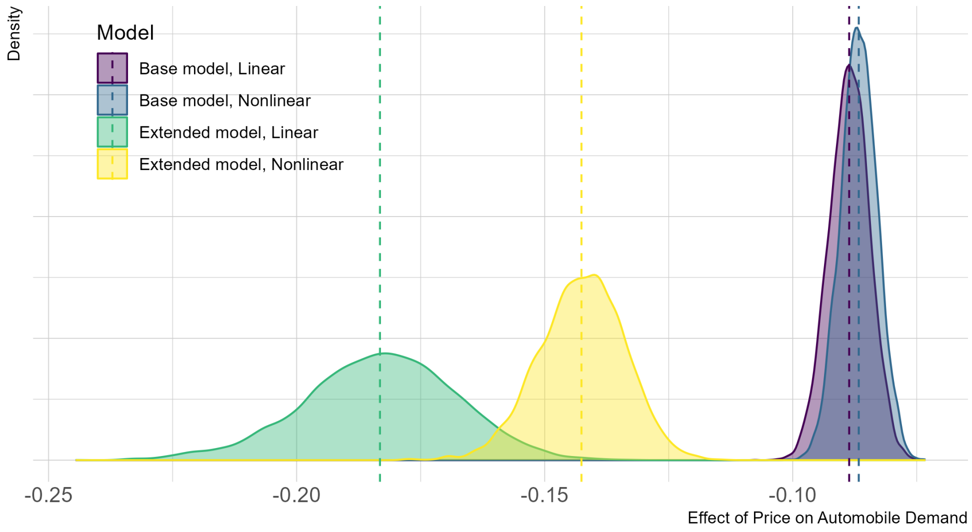

Figure 3: BLP models: Marginal posterior distributions of the coefficient on the price variable, β. Posterior mean and standard deviation of β are -0.089 and 0.004 for the base model with the original BLP (linear) specification while they are -0.087 and 0.004 with the augmented BLP (nonlinear) specification. For the extended model, posterior mean and standard deviation of β are -0.183 and 0.015 for the linear specification and -0.143 and 0.009 for the nonlinear specification.

5.2 Effect of airfares on passenger traffic

The emphasis of the theory and applications in this paper is on situations with a single outcome variable, however, our framework can be applied more broadly. An important example is clustered, longitudinal data. Let $y_i = (y_{i1}, \dots, y_{iT})$ denote T potentially correlated and heteroskedastic measurements on subject i. The outcome is thus a $T \times 1$ vector, rather than a scalar. Adjusting the dimensions of the controls and instruments, respectively, suppose that independently across i, the clustered outcomes follow the linear model $y_i = X_i\beta + Z_{1,i}\gamma + \varepsilon_i$, where X_i is $T \times d_x$, $Z_{1,i}$ is $T \times d_{z_1}$, $Z_{2,i}$ is $T \times d_{z_2}$, and ε_i is $T \times 1$. Now assume that $Z_{1,i}$ and $Z_{2,i}$ satisfy the clustered data exogeneity restrictions $E[Z'_{j,i}\varepsilon_i(\theta)] = 0, j = 1, 2$, but that the clustered data exogeneity restrictions $E[X'_i\varepsilon_i(\theta)] = 0$ related to X_i are in doubt. We can apply our framework to this problem by defining a base model in which the latter restrictions are imposed, and an extended model that contains the inactive restrictions $E[X'_i\varepsilon_i(\theta)] = v$, where v is now a $d_x \times 1$ vector of unknown parameters. In parallel to the approach developed above, the marginal likelihood comparison of these models is a test for the exogeneity of X.

As an illustration of this extended set-up, we consider a $T = 4$ balanced longitudinal data set on airfares and passenger traffic for the years 1997, 1998, 1999, and 2000 from Wooldridge (2010). For each year t, $t \leq 4$, the data is clustered by route i, $i \leq n = 1149$. For each flight route defined by the origin and destination cities, one has the log of the average number of passengers per day ($lpassen$), the log of the average one-way fare in dollars ($lfare$), the log of the distance in miles ($ldist$), and the fraction of the market corralled by the biggest carrier ($concen$). The model of interest is $lpassen_{it} = \beta\, lfare_{it} + \gamma_1 trend_t + \gamma_2 ldist_{it} + \varepsilon_{it}$, where $trend$ is a trend variable taking values $1, 2, 3, 4$, and each of the variables in this regression is mean centered. The goal is to estimate the price elasticity parameter β, but one is concerned that $lfare$ is possibly endogenous. In the estimation we assume that $concen$ is a valid instrument (it does not directly appear in the outcome model and it affects $lfare$, both reasonable assumptions).

Clustered by route i, we have

$$\begin{pmatrix} lpassen_{i1} \\ lpassen_{i2} \\ lpassen_{i3} \\ lpassen_{i4} \end{pmatrix} = \begin{pmatrix} lfare_{i1} & 1 & ldist_{i1} \\ lfare_{i2} & 2 & ldist_{i2} \\ lfare_{i3} & 3 & ldist_{i3} \\ lfare_{i4} & 4 & ldist_{i4} \end{pmatrix} \begin{pmatrix} \beta \\ \gamma_1 \\ \gamma_2 \end{pmatrix} + \begin{pmatrix} \varepsilon_{i1} \\ \varepsilon_{i2} \\ \varepsilon_{i3} \\ \varepsilon_{i4} \end{pmatrix},$$

or compactly as $y_i = \widetilde{W}_{1,i}\theta + \varepsilon_i$, $i = 1, 2, \ldots, 1149$, where $\theta : 7 \times 1$ is the unknown parameter of interest. In this model, the distribution of ε_i is not specified. Moreover, the elements of ε_i can be serially correlated and heteroskedastic in an arbitrary, unknown way.

Now let $Z_i := \left(\widetilde{W}_{1,i}, 1, concen_i\right)$, $i \leq n$, be a 4×5 matrix, where 1 is a vector of ones, and $concen_i = (concen_{i1}, \ldots, concen_{i4})' : 4 \times 1$ is the vector of $concen$ values for route i. In the base model, $lfare$ is exogenous. The model is defined by the five moments

$$\mathcal{M}_b : \quad \mathbf{E}[Z_i'(y_i - X_i\theta)] = 0_{5\times 1}$$

In the extended model, the $lfare$ moment condition is inactive. Specifically,

$$\mathcal{M}_e : \quad \mathbf{E}[Z_i'(y_i - X_i\theta)] = \begin{pmatrix} v \\ 0_{4\times 1} \end{pmatrix}$$

The ETEL-based estimation of these two models makes no assumption about the joint distribution of the cluster-level errors.

We specify the prior from a training sample. We randomly split the sample into a training sample (of say 115 clusters, equal to 10% of the total clusters) and an estimation sample (consisting of the remaining 1034 clusters). We then estimate the base mode on the training sample with a student-t prior centered on the system wide 2SLS estimate from the training data, sd of 10 and 2.5 degrees of freedom. The posterior mean and sd is calculated from these training data under this prior. We then take the posterior mean and twice the sd from the training sample fit as the mean and sd of the prior. This determination of the prior from the training sample is helpful in the fitting, but, due to the thick tails of the prior, the information brought in by the prior pales in comparison with the information from the estimation sample.

We sample the posterior in each model by the one-block tailored MCMC algorithm. In the base model, from 10,000 MCMC draws beyond a burn-in of 1000, we find that the posterior mean of β is -0.551 and its 95% posterior credibility interval is (-0.683, -0.419). Moreover, computation shows that $\log(m(w_{1:n}|\mathcal{M}_b) = -7190.222$ and $\log(m(w_{1:n}|\mathcal{M}_e) = -7191.06$, signalling that $lprice$ in this problem can be viewed as exogenous.

6 Concluding remarks

This paper has developed an analysis of regression models in which the variables of primary interest, the treatment variables, are possibly endogenous (correlated with the regression error). This endogeneity problem is generally assumed away in the Bayesian literature, but this leads to a serious misspecification problem since endogeneity, in practice, is the rule, rather than the exception.

In order to avoid the risk of distributional misspecification, the framework we have developed relies only on moment restrictions. The analysis in the paper revolves around the study of two models: the base model, where the exogeneity assumption is enforced, and an extended model, where the exogeneity moment is included, but is made inactive. The real-data examples discussed in the paper showcase the practical relevance of the methods.

The paper makes two key contributions. The first is in the study of the large sample behavior of the posterior distributions in the base and extended models in cases where the exogeneity assumption in the population is true or false. The second is in the development of a Bayesian test of endogeneity that is based on the marginal likelihoods of the base and extended models. In the former case, BvM theorems are established and, in the latter, the large sample consistency of the Bayes factor test is established.

It is important to mention that the approach proposed here can be extended to situations where the controls are assumed to enter the model nonparameterically. While the finite sample analysis of such models, after approximating the unknown functions by (say) spline basis expansion methods, would proceed in much the same way as discussed in this paper, the large sample analysis would require new developments to account for a growing number of basis function parameters with sample size. We intend to describe the theory in a future paper.

References

Berry, S., Levinsohn, J. and Pakes, A. (1995), ‘Automobile prices in market equilibrium’, *Econometrica* **63**(4), 841–890.

Chamberlain, G. (1987), ‘Asymptotic efficiency in estimation with conditional moment restrictions’, *Journal of Econometrics* **34**(3), 305–334.

Chao, J. and Phillips, P. (1998), ‘Posterior distributions in limited information analysis of the simultaneous equations model using the jeffreys prior’, *Journal of Econometrics* **87**(1), 49–86.

Chernozhukov, V., Hansen, C. and Spindler, M. (2015), ‘Post-selection and post-regularization inference in linear models with many controls and instruments’, *American Economic Review* **105**(5), 486–490.

Chernozhukov, V. and Hong, H. (2003), ‘An MCMC Approach to Classical Estimation’, *Journal of Econometrics* **115**(2), 293–346.

Chib, S. (1995), ‘Marginal likelihood from the Gibbs output’, *Journal of the American Statistical Association* **90**(432), 1313–1321.

Chib, S. and Greenberg, E. (2010), ‘Additive cubic spline regression with Dirichlet process mixture errors’, *Journal of Econometrics* **156**(2), 322–336.

Chib, S. and Jeliazkov, I. (2001), ‘Marginal likelihood from the Metropolis-Hastings output’, *Journal of the American Statistical Association* **96**(453), 270–281.

Chib, S., Shin, M. and Simoni, A. (2018), ‘Bayesian estimation and comparison of moment condition models’, *Journal of the American Statistical Association* **113**(524), 1656–1668.

Drèze, J. H. (1976), ‘Bayesian limited information analysis of the simultaneous equations model’, *Econometrica* **44**(5), 1045–1075.

Florens, J.-P. and Simoni, A. (2012), ‘Nonparametric estimation of an instrumental regression: A quasi-Bayesian approach based on regularized posterior’, *Journal of Econometrics* **170**(2), 458 – 475.

Florens, J.-P. and Simoni, A. (2016), ‘Regularizing priors for linear inverse problems’, *Econometric Theory* **32**(1), 71–121.

Florens, J.-P. and Simoni, A. (2021), ‘Gaussian processes and Bayesian moment estimation’, *Journal of Business & Economic Statistics* **39**(2), 482–492.

Hansen, 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**(6), 1251–1271.

Hoogerheide, L., Kleibergen, F. and van Dijk, H. K. (2007), 'Natural conjugate priors for the instrumental variables regression model applied to the angrist-krueger data', *Journal of Econometrics* **138**(1), 63–103.

Kato, K. (2013), 'Quasi-Bayesian analysis of nonparametric instrumental variables models', *Annals of Statistics* **41**(5), 2359–2390.

Kleibergen, F. and van Dijk, H. K. (1998), 'Bayesian simultaneous equations analysis using reduced rank structures', *Econometric Theory* **14**(6), 701–743.

Kleibergen, F. and Zivot, E. (2003), 'Bayesian and classical approaches to instrumental variable regression', *Journal of Econometrics* **114**(1), 29–72.

Kleijn, B. and van der Vaart, A. (2012), 'The Bernstein-von-Mises theorem under misspecification', *Electronic Journal of Statistics* **6**, 354–381.

Lehmann, E. L. and Casella, G. (1998), *Theory of Point Estimation (Springer Texts in Statistics)*, 2nd edn, Springer.

Liao, Y. and Jiang, W. (2011), 'Posterior consistency of nonparametric conditional moment restricted models', *Annals of Statistics* **39**(6), pp. 3003–3031.

Schennach, S. M. (2005), 'Bayesian exponentially tilted empirical likelihood', *Biometrika* **92**(1), 31–46.

Schennach, S. M. (2007), 'Point Estimation with Exponentially Tilted Empirical Likelihood', *Annals of Statistics* **35**(2), 634–672.

Shin, M. (2014), Bayesian GMM, Technical report, University of Pennsylvania.

Sueishi, N. (2013), 'Identification problem of the exponential tilting estimator under misspecification', *Economics Letters* **118**(3), 509 – 511.

Van der Vaart, A. W. (1998), *Asymptotic Statistics*, Cambridge Series in Statistical and Probabilistic Mathematics.

Wooldridge, J. M. (2010), *Econometric Analysis of Cross Section and Panel Data*, MIT press.

A Proofs of the main results

We denote by $g_i(\theta) := g(w_i, \theta)$ the moment function evaluated at w_i and by $\widehat{\lambda}(\theta) := \widehat{\lambda}(w_{1:n}, \theta)$ the tilting parameter. When we omit y_i from the vector of the i-th observation we use the notation $\widetilde{w}_i := (x_i', z_i')'$, and when in addition we omit $z_{2,i}$ we use the notation $\widetilde{w}_{1,i} := (x_i', z_{1,i}')'$. Moreover, $\widehat{g}(\theta) := \mathbf{E}_n[g_i(\theta)]$, $dg_i(\theta)/d\theta' = -\widetilde{w}_i\widetilde{w}_{1,i}'$. We use the notation $\mathbf{E}_n[\cdot] := \frac{1}{n}\sum_{i=1}^n[\cdot]$ for the empirical mean. For a probability Q we use the notation $\mathbf{E}^Q[\cdot]$ to denote the expectation with respect to Q and $\mathbb{V}ar_Q$ the variance with respect to Q. For the true distribution P: $\mathbf{E}[\cdot] := \mathbf{E}^P[\cdot]$. The log-likelihood function for one observation w_i is denoted by $\ell_{n,\theta}(w_i)$:

$$\ell_{n,\theta}(w_i) := \log \widehat{p}_i(\theta) = \log \frac{e^{\widehat{\lambda}(\theta)'g(w_i,\theta)}}{\sum_{k=1}^n e^{\widehat{\lambda}(\theta)'g_k(\theta)}} = -\log n + \log \frac{e^{\widehat{\lambda}(\theta)'g(w_i,\theta)}}{\frac{1}{n}\sum_{k=1}^n e^{\widehat{\lambda}(\theta)'g_k(\theta)}}$$

so that the log-ETEL function is $\ell_{n,\theta}(w_{1:n}) = \sum_{i=1}^n \ell_{n,\theta}(w_i) = \log \prod_{i=1}^n \widehat{p}_i(\theta) = \log \widehat{p}(w_{1:n}|\theta)$.

By replacing $\widehat{\lambda}(\theta)$ with its true value $\lambda_*(\theta)$ we define:

$$\ell_{*,\theta}(w) := \log \frac{e^{\lambda_*(\theta)'g(w,\theta)}}{\sum_{k=1}^n e^{\lambda_*(\theta)'g_k(\theta)}} =: \log p_w^*(\theta) \qquad \text{and} \qquad \ell_{*,\theta}(w_{1:n}) := \sum_{i=1}^n \ell_{*,\theta}(w_i).$$

The first (resp. second) derivative of $\theta \mapsto \ell_{n,\theta}(w_{1:n})$ evaluated at a point θ_1 is denoted by $\dot{\ell}_{n,\theta_1}(w_{1:n})$ (resp. $\ddot{\ell}_{n,\theta_1}(w_{1:n})$). Moreover, for a function $\lambda(\theta)$ of θ, define $\tau_i(\lambda, \theta) := \frac{e^{\lambda(\theta)'g_i(\theta)}}{\mathbf{E}_n[e^{\lambda(\theta)'g_j(\theta)}]}$, $\tau_i^\dagger(\lambda, \theta) := \frac{e^{\lambda(\theta)'g_i(\theta)}}{\mathbf{E}[e^{\lambda(\theta)'g_j(\theta)}]}$, $\tau_i^\diamond(\lambda, \theta) := e^{\lambda(\theta)'g_i(\theta)}$. So, $\tau_i(\widehat{\lambda}, \theta) = n\widehat{p}_i(\theta)$ and $\tau_i^\dagger(\lambda_*, \theta) = dQ^*(\theta)/dP$. We also use the notation: $\check{\Omega}(\lambda, \theta) := \mathbf{E}_n[\tau_i(\lambda, \theta)\varepsilon_i(\theta)\widetilde{w}_i\widetilde{w}_i']$, $\check{\Omega}^\diamond(\lambda, \theta) := \mathbf{E}_n[\tau_i^\diamond(\lambda, \theta)\varepsilon_i(\theta)\widetilde{w}_i\widetilde{w}_i']$, $\check{\Omega}^\dagger(\lambda, \theta) := \mathbf{E}_n[\tau_i^\dagger(\lambda, \theta)\varepsilon_i(\theta)\widetilde{w}_i\widetilde{w}_i']$, $\Omega_*^\diamond(\theta) := \mathbf{E}[\tau_i^\diamond(\lambda_*, \theta)\varepsilon_i(\theta)\widetilde{w}_i\widetilde{w}_i']$ and $\Omega_*^\dagger(\theta) := \mathbf{E}^{Q^*(\theta)}[\varepsilon_i(\theta)\widetilde{w}_i\widetilde{w}_i']$. Finally, MVT refers to the Mean Value Theorem.

Moreover, we make use of the following identities obtained by taking the total derivative of the first order condition for $\widehat{\lambda}$ and for λ_*: $\forall\theta \in \Theta$,

$$\begin{aligned}\frac{d\widehat{\lambda}(\theta)'}{d\theta} &= \mathbf{E}_n\left[\tau_i(\widehat{\lambda},\theta)\widetilde{w}_{1,i}\widetilde{w}_i'(I+\widehat{\lambda}(\theta)g_i(\theta)')\right]\check{\Omega}(\widehat{\lambda},\theta)^{-1},\\ \frac{d\lambda_*(\theta)'}{d\theta} &= \mathbf{E}^{Q^*(\theta)}\left[\widetilde{w}_{1,i}\widetilde{w}_i'(I+\lambda_*(\theta)g_i(\theta)')\right]\Omega_*^{\dagger}(\theta)^{-1},\end{aligned}$$

respectively.

A.1 Proof of Theorem 2.1

Define the events $A_{n,1} := \left\{\sup_{\theta\in\Theta_n^c}\frac{1}{n}\sum_{i=1}^n(\ell_{n,\theta}(w_i)-\ell_{n,\theta_*}(w_i)) \le -CM_n^2/n\right\}$ and

$$A_{n,2} := \left\{\int_\Theta \frac{\widehat{p}(w_{1:n}|\theta)}{\widehat{p}(w_{1:n}|\theta_*)}\pi(\theta)d(\theta) \ge e^{-CM_n^2/2}\right\}.$$

By (2.9), $P(A_{n,1}^c) \to 0$ and by Lemma B.2 in the Supplementary Material, $P(A_{n,2}^c) \to 0$. Therefore, by the law of total expectation

$$\begin{aligned}\mathbf{E}\left[\pi\left(\Theta_n^c\middle| w_{1:n}\right)\right] &\le \mathbf{E}\left[\pi\left(\sqrt{n}\|\theta-\theta_*\|>M_n\middle| w_{1:n}\right)\middle| A_{n,1}\cap A_{n,2}\right]P(A_{n,1}\cap A_{n,2})+o(1)\\ &= \mathbf{E}\left[\frac{\int_{\Theta_n^c}e^{\sum_{i=1}^n(\ell_{n,\theta}(w_i)-\ell_{n,\theta_*}(w_i))}\pi(\theta)d\theta}{\int_\Theta e^{\sum_{i=1}^n(\ell_{n,\theta}(w_i)-\ell_{n,\theta_*}(w_i))}\pi(\theta)d\theta}\middle| A_{n,1}\cap A_{n,2}\right]P(A_{n,1}\cap A_{n,2})+o(1)\\ &\le e^{-CM_n^2}\pi(\Theta_n^c)\mathbf{E}\left[\left(\int_\Theta\frac{\widehat{p}(w_{1:n}|\theta)}{\widehat{p}(w_{1:n}|\theta_*)}\pi(\theta)d\theta\right)^{-1}\middle| A_{n,1}\cap A_{n,2}\right]+o(1)\\ &\le e^{-CM_n^2}e^{CM_n^2/2}\pi(\Theta_n^c)+o(1)=o(1) \quad \text{(A.1)}\end{aligned}$$

which proves the result of the theorem.

A.2 Proof of Theorem 2.2

The proof of this theorem proceeds as the proof of Chib et al. (2018, Theorem 2.2). It depends on two intermediate results: the posterior consistency result of Theorem 2.1 and the stochastic LAN expansion (A.2) established in Theorem A.1 below. We prove these two theorems by using a strategy of proofs different from the one used in Chib et al. (2018) and that is specific for the setting considered in this paper.

Theorem A.1 (Stochastic LAN under neglected endogeneity.) *Let V_{θ_*} be as defined in the statement of Theorem 2.2 and assume it is nonsingular. Let Assumptions 2.1 - 2.6 hold. For every $h \in \mathbb{R}^p$ let $\theta := \theta_* + h/\sqrt{n}$. Then, for every compact set $K \subset \mathbb{R}^p$,*

$$\sup_{h\in K}\left|\log\frac{\widehat{p}(w_{1:n}|\theta_* + h/\sqrt{n})}{\widehat{p}(w_{1:n}|\theta_*)} - h'V_{\theta_*}^{-1}\Delta_{n,\theta_*} + \frac{1}{2}h'V_{\theta_*}^{-1}h\right| \xrightarrow{p} 0 \qquad \text{as } n \to \infty, \tag{A.2}$$

where θ_* is as defined in (2.8), *$V_{\theta_*}^{-1} := -\text{plim}\ddot{\ell}_{n,\theta_*}(w_{1:n})/n$, $h'V_{\theta_*}^{-1}\Delta_{n,\theta_*} := \frac{h'}{\sqrt{n}}\dot{\ell}_{n,\theta_*}(w_{1:n}) \xrightarrow{d} \mathcal{N}(0, h'H_*h)$ is bounded in probability and H_* is a positive definite matrix defined in Lemma B.3 of the Supplementary Material.*

Proof. We use a second order MVT expansion applied to $\theta \mapsto \ell_{n,\theta}(w_{1:n})$ around θ_* and the first order condition of $\widehat{\lambda}(\theta)$, $\mathbf{E}_n\left[e^{\widehat{\lambda}(\theta)'g_i(\theta)}g_i(\theta)\right] = 0$, to get:

$$\begin{aligned}
\ell_{n,\theta}(w_{1:n}) - \ell_{n,\theta_*}(w_{1:n}) &= (\theta-\theta_*)'\dot{\ell}_{n,\theta_*}(w_{1:n}) + \frac{1}{2}(\theta-\theta_*)'\ddot{\ell}_{n,\widetilde{\theta}}(w_{1:n})(\theta-\theta_*)\\
&= (\theta-\theta_*)'\frac{d\widehat{\lambda}(\theta_*)'}{d\theta}n\widehat{g}(\theta_*) + (\theta-\theta_*)'\frac{d\widehat{g}(\theta_*)'}{d\theta}\widehat{\lambda}(\theta_*)n + \frac{n}{2}(\theta-\theta_*)'\frac{d^2[\widehat{\lambda}(\widetilde{\theta})'\widehat{g}(\widetilde{\theta})]}{d\theta d\theta'}(\theta-\theta_*)\\
&- n(\theta-\theta_*)'\mathbf{E}_n\left[\tau_i(\widehat{\lambda},\theta_*)\frac{dg_i(\theta_*)'}{d\theta}\right]\widehat{\lambda}(\theta) - \frac{n}{2}(\theta-\theta_*)'\mathbf{E}_n\left[\tau_i(\widehat{\lambda},\widetilde{\theta})\frac{dg_i(\widetilde{\theta})'}{d\theta}\right]\frac{d\widehat{\lambda}(\theta)}{d\theta'}(\theta-\theta_*)\\
&\quad - \frac{n}{2}(\theta-\theta_*)'\mathbf{E}_n\left[\tau_i(\widehat{\lambda},\widetilde{\theta})\frac{dg_i(\widetilde{\theta})'}{d\theta}\widehat{\lambda}(\widetilde{\theta})\widehat{\lambda}(\widetilde{\theta})'\frac{dg_i(\widetilde{\theta})}{d\theta'}\right](\theta-\theta_*)\\
&\quad - \frac{n}{2}(\theta-\theta_*)'\frac{d\widehat{\lambda}(\widetilde{\theta})'}{d\theta}\mathbf{E}_n\left[\tau_i(\widehat{\lambda},\widetilde{\theta})g_i(\widetilde{\theta})\widehat{\lambda}(\theta_*)'\frac{dg_i(\widetilde{\theta})}{d\theta'}\right](\theta-\theta_*)\\
&\quad + \frac{n}{2}(\theta-\theta_*)'\mathbf{E}_n\left[\tau_i(\widehat{\lambda},\widetilde{\theta})\frac{dg_i(\widetilde{\theta})'}{d\theta}\right]\widehat{\lambda}(\widetilde{\theta})\widehat{\lambda}(\widetilde{\theta})'\mathbf{E}_n\left[\tau_i(\widehat{\lambda},\widetilde{\theta})\frac{dg_i(\widetilde{\theta})}{d\theta'}\right](\theta-\theta_*)
\end{aligned}$$

for $\widetilde{\theta} = \tau\theta + (1-\tau)\theta_*$ and some $\tau \in [0,1]$. By replacing θ by $\theta_* + h/\sqrt{n}$, so that $\widetilde{\theta} = \theta_* + \tau h/\sqrt{n}$ and by using the expression for $g_i(\theta)$ and its derivative with respect to θ, the previous expression simplifies as:

$$\begin{aligned}\ell_{n,\theta}(w_{1:n}) - \ell_{n,\theta_*}(w_{1:n}) = h'\frac{d\widehat{\lambda}(\theta_*)'}{d\theta}\sqrt{n}\widehat{g}(\theta_*) - h'\sqrt{n}\mathbf{E}_n[\widetilde{w}_{1,i}\widetilde{w}_i']\widehat{\lambda}(\theta_*) + \frac{1}{2}h'\frac{d^2[\widehat{\lambda}(\widetilde{\theta})'\widehat{g}(\widetilde{\theta})]}{d\theta d\theta'}h \\ + \sqrt{n}h'\mathbf{E}_n\left[\tau_i(\widehat{\lambda},\theta_*)\widetilde{w}_{1,i}\widetilde{w}_i'\right]\widehat{\lambda}(\theta_*) + \frac{1}{2}h'\mathbf{E}_n\left[\tau_i(\widehat{\lambda},\widetilde{\theta})\widetilde{w}_{1,i}\widetilde{w}_i'\right]\frac{d\widehat{\lambda}(\widetilde{\theta})}{d\theta'}h \\ - \frac{1}{2}\mathbf{E}_n\left[\tau_i(\widehat{\lambda},\widetilde{\theta})\left(h'\widetilde{w}_{1,i}\widetilde{w}_i'\widehat{\lambda}(\widetilde{\theta})\right)^2\right] + \frac{1}{2}h'\frac{d\widehat{\lambda}(\widetilde{\theta})'}{d\theta}\mathbf{E}_n\left[\tau_i(\widehat{\lambda},\widetilde{\theta})g_i(\widetilde{\theta})\widehat{\lambda}(\widetilde{\theta})'\widetilde{w}_i\widetilde{w}_{1,i}'\right]h \\ + \frac{1}{2}h'\mathbf{E}_n\left[\tau_i(\widehat{\lambda},\widetilde{\theta})\widetilde{w}_{1,i}\widetilde{w}_i'\right]\widehat{\lambda}(\widetilde{\theta})\widehat{\lambda}(\widetilde{\theta})'\mathbf{E}_n\left[\tau_i(\widehat{\lambda},\widetilde{\theta})\widetilde{w}_i\widetilde{w}_{1,i}'\right]h \quad \text{(A.3)}\end{aligned}$$

and $h'\frac{d^2[\widehat{\lambda}(\widetilde{\theta})'\widehat{g}(\widetilde{\theta})]}{d\theta d\theta'}h = h'\sum_{l=1}^{d}\frac{d^2\widehat{\lambda}_l(\widetilde{\theta})}{d\theta d\theta'}\widehat{g}_l(\widetilde{\theta})h - 2h'\frac{d\widehat{\lambda}(\widetilde{\theta})'}{d\theta}\mathbf{E}_n\left[\widetilde{w}_i\widetilde{w}_{1,i}'\right]h$, where $\widehat{\lambda}_l(\widetilde{\theta})$ and $\widehat{g}_l(\theta)$ denote the l-th components of the vectors $\widehat{\lambda}(\widetilde{\theta})$ and $\widehat{g}(\theta)$, respectively. Let us start by considering the terms of first order, to which we add and subtract the first order condition for θ_* (which is $\frac{d\lambda_*(\theta_*)'}{d\theta}\mathbf{E}[g_i(\theta_*)] - \mathbf{E}[\widetilde{w}_{1i}\widetilde{w}_i']\lambda_*(\theta_*) + \mathbf{E}[\tau_i^\dagger(\lambda_*,\theta_*)\widetilde{w}_{1i}\widetilde{w}_i']\lambda_*(\theta_*) = 0$):

$$\begin{aligned}h'\frac{d\widehat{\lambda}(\theta_*)'}{d\theta}\sqrt{n}\widehat{g}(\theta_*) - h'\sqrt{n}\mathbf{E}_n[\widetilde{w}_{1,i}\widetilde{w}_i']\widehat{\lambda}(\theta_*) + \sqrt{n}h'\mathbf{E}_n\left[\tau_i(\widehat{\lambda},\theta_*)\widetilde{w}_{1,i}\widetilde{w}_i'\right]\widehat{\lambda}(\theta) = \\ h'\sqrt{n}\left(\frac{d\widehat{\lambda}(\theta_*)'}{d\theta}\widehat{g}(\theta_*) - \frac{d\lambda_*(\theta_*)'}{d\theta}\mathbf{E}[g_i(\theta_*)]\right) - h'\sqrt{n}\left(\mathbf{E}_n[\widetilde{w}_{1,i}\widetilde{w}_i']\widehat{\lambda}(\theta_*) - \mathbf{E}[\widetilde{w}_{1i}\widetilde{w}_i']\lambda_*(\theta_*)\right) \\ + \sqrt{n}h'\left(\mathbf{E}_n\left[\tau_i(\widehat{\lambda},\theta_*)\widetilde{w}_{1,i}\widetilde{w}_i'\right]\widehat{\lambda}_*(\theta) - \mathbf{E}[\tau_i^\dagger(\lambda_*,\theta_*)\widetilde{w}_{1i}\widetilde{w}_i']\lambda_*(\theta_*)\right) =: h'V_{\theta_*}^{-1}\Delta_{n,\theta_*}.\end{aligned}$$

By Lemma B.3 in the Supplementary Material, $h'V_{\theta_*}^{-1}\Delta_{n,\theta_*}$ is asymptotically normal with zero mean and variance equal to the non-singular matrix $h'H_*h$ whose expression is given in the statement of Lemma B.3.

Now, let us consider the terms of second order in (A.3). Because h is bounded, then $\widetilde{\theta} \to \theta_*$ as $n \to \infty$. Moreover, we use the following limits as $n \to \infty$. (1) By continuity of $\theta \mapsto \lambda_*(\theta)$ (by Lemma B.5 in the Supplementary Material), and continuity of $\theta \mapsto g_i(\theta)$, we have: $\lambda_*(\widetilde{\theta}) \to \lambda_*(\theta_*)$, and $g_i(\widetilde{\theta}) \to g_i(\theta_*)$. (2) By Lemma B.9 then $\mathbf{E}_n\left[\tau_i(\widehat{\lambda},\widetilde{\theta})\widetilde{w}_{1,i}\widetilde{w}_i'\right]$ converges in probability to $\mathbf{E}\left[\tau_i^\dagger(\lambda_*,\theta_*)\widetilde{w}_{1,i}\widetilde{w}_i'\right]$. (3) By Lemma B.10 then $\mathbf{E}_n\left[\tau_i(\widehat{\lambda},\widetilde{\theta})\left(h'\widetilde{w}_{1,i}\widetilde{w}_i'\widehat{\lambda}(\widetilde{\theta})\right)^2\right]$

converges in probability to $\mathbf{E}\left[\tau_i^{\dagger}(\lambda_*,\theta_*)\left(h'\widetilde{w}_{1,i}\widetilde{w}_i'\lambda_*(\theta_*)\right)^2\right]$. (4) By Lemma B.11 then

$$\mathbf{E}_n\left[\tau_i(\widehat{\lambda},\widetilde{\theta})h'\widetilde{w}_{1,i}\widetilde{w}_i'\widehat{\lambda}(\widetilde{\theta})\varepsilon_i(\widetilde{\theta})'\widetilde{w}_i'\right]\xrightarrow{p}\mathbf{E}\left[\tau_i^{\dagger}(\lambda_*,\theta_*)h'\widetilde{w}_{1,i}\widetilde{w}_i'\lambda_*(\theta_*)\varepsilon_i(\theta_*)'\widetilde{w}_i'\right].$$

(5) By combining (2), (4) and Lemma B.6 we have that $h'\frac{d\widehat{\lambda}(\widetilde{\theta})'}{d\theta}\xrightarrow{p}h'\mathbf{E}^{Q^*(\theta_*)}[\widetilde{w}_{1,i}\widetilde{w}_i'(I+\lambda_*(\theta_*)g_i(\theta_*)')]\Omega_*^{\diamond}(\theta_*)^{-1}$. Hence, by using these limits the term of second order in (A.3) is equal to:

$$\begin{aligned}\frac{1}{2}h'\sum_{l=1}^{d}\frac{d^2\lambda_{*,l}(\theta_*)}{d\theta d\theta'}\mathbf{E}\left[g_{i,l}(\theta_*)\right]h-h'\frac{d\lambda_*(\theta_*)'}{d\theta}\mathbf{E}\left[\widetilde{w}_i\widetilde{w}_{1,i}'\right]h+\frac{1}{2}h'\mathbf{E}\left[\tau_i^{\dagger}(\lambda_*,\theta_*)\widetilde{w}_{1,i}\widetilde{w}_i'\right]\frac{d\lambda_*(\theta_*)}{d\theta'}h\\-\frac{1}{2}\mathbf{E}\left[\tau_i^{\dagger}(\lambda_*,\theta_*)\left(h'\widetilde{w}_{1,i}\widetilde{w}_i'\lambda_*(\theta_*)\right)^2\right]+\frac{h'}{2}\frac{d\lambda_*(\theta_*)'}{d\theta}\mathbf{E}\left[\tau_i^{\dagger}(\lambda_*,\theta_*)g_i(\theta_*)\lambda_*(\theta_*)'\widetilde{w}_i\widetilde{w}_{1,i}'\right]h\\+\frac{1}{2}h'\mathbf{E}\left[\tau_i^{\dagger}(\lambda_*,\theta_*)\widetilde{w}_{1,i}\widetilde{w}_i'\right]\lambda_*(\theta_*)\lambda_*(\theta_*)'\mathbf{E}\left[\tau_i^{\dagger}(\lambda_*,\theta_*)\widetilde{w}_i\widetilde{w}_{1,i}'\right]h+o_p(1).\quad\text{(A.4)}\end{aligned}$$

By remarking that $\tau_i^{\dagger}(\lambda_*,\theta_*)=dQ^*(\theta_*)/dP$, and that $\mathbf{E}\left[g_{i,l}(\theta_*)\right]=0$ for every $l>d_x$, the previous expression can be simplified as

$$\begin{aligned}\frac{1}{2}h'\sum_{l=1}^{d_x}\frac{d^2\lambda_{*,l}(\theta_*)}{d\theta d\theta'}\mathbf{E}\left[g_{i,l}(\theta_*)\right]h-h'\frac{d\lambda_*(\theta_*)'}{d\theta}\mathbf{E}\left[\widetilde{w}_i\widetilde{w}_{1,i}'\right]h-\frac{1}{2}h'\mathbb{V}ar_{Q^*(\theta_*)}\left[\widetilde{w}_{1,i}\widetilde{w}_i'\lambda_*(\theta_*)\right]h\\+\frac{1}{2}h'\mathbf{E}^{Q^*(\theta_*)}\left[\widetilde{w}_{1,i}\widetilde{w}_i'\left(I+\lambda_*(\theta_*)g_i(\theta_*)'\right)\right]\frac{d\lambda_*(\theta_*)}{d\theta'}h+o_p(1)\\=:-\frac{1}{2}h'V_{\theta_*}^{-1}h+o_p(1).\end{aligned}$$

By putting all these elements together we get:

$$\ell_{n,\theta}(w_{1:n})-\ell_{n,\theta_*}(w_{1:n})=h'V_{\theta_*}^{-1}\Delta_{n,\theta_*}-\frac{1}{2}h'V_{\theta_*}^{-1}h+o_p(1),$$

where $h'V_{\theta_*}^{-1}\Delta_{n,\theta_*}\xrightarrow{d}\mathcal{N}(0,h'H_*h)$, $\frac{h'}{\sqrt{n}}\dot{\ell}_{n,\theta_*}=h'V_{\theta_*}^{-1}\Delta_{n,\theta_*}+o_p(1)$ and $V_{\theta_*}^{-1}=\text{plim}\ddot{\ell}_{n,\theta_*}/n+o_p(1)$. Thus, we obtain the result of the lemma.

A.3 Proof of Theorem 4.1

In view of Chib et al. (2018, Theorem 3.2) we know that

$$\lim_{n\to\infty} P\left(\log m(w_{1:n}|M_e) > \log m(w_{1:n}|M_b)\right) = 1$$

if and only if $\mathrm{KL}(P||Q_e^*(\psi_\circ)) < \mathrm{KL}(P||Q^*(\theta_*))$ (remark that $\theta_* = \theta_\circ$ when x_i is exogenous). We now prove that the latter inequality holds if and only if there is no θ such that $\mathbf{E}[\varepsilon_i(\theta)x_i] = 0$. The remaining moment restrictions are always satisfied by assumption.

Suppose that $\mathrm{KL}(P||Q_e^*(\psi_\circ)) < \mathrm{KL}(P||Q^*(\theta_*))$ and suppose that there exists a θ such that $\mathbf{E}[\varepsilon_i(\theta)x_i] = 0$ so that $P \in \mathcal{Q}_\theta$. By Assumption 2.2 with θ_* replaced by $\theta_\circ$ then this θ must be equal to $\theta_\circ$ which in turn equals θ_*. It follows that $P \in \mathcal{Q}_{\theta_*}$ and by definition of $Q^*(\theta_*)$: $Q^*(\theta_*) = P$ since $Q^*(\theta_*)$ is the closest to P, in the KL sense, among all the distributions in $\mathcal{Q}_{\theta_*}$. Hence, $\mathrm{KL}(P||Q^*(\theta_*)) = 0$. But this contradicts the assumption that $\mathrm{KL}(P||Q_e^*(\psi_\circ)) < \mathrm{KL}(P||Q^*(\theta_*))$. Hence, there is no θ such that $\mathbf{E}[\varepsilon_i(\theta)x_i] = 0$.

We now prove the reverse implication. Suppose that there is no value θ such that $\mathbf{E}[\varepsilon_i(\theta)x_i] = 0$. Hence, $P \notin \mathcal{Q}_\theta$ for every $\theta \in \Theta$ which implies $P \notin \mathcal{Q}_{\theta_*}$ and $\mathrm{KL}(P||Q^*(\theta_*)) > 0$. On the other hand, there exists a unique $\psi_\circ \in \mathbb{R}^{d_x}$ such that $P \in \mathcal{Q}_{e,\psi_\circ}$ since $\mathcal{P}_{e,\circ}$ is always correctly specified. This implies that $\mathrm{KL}(P||Q_e^*(\psi_\circ)) = 0$ and so $\mathrm{KL}(P||Q_e^*(\psi_\circ)) < \mathrm{KL}(P||Q^*(\theta_*))$.