Testing for Endogeneity: A Moment-Based Bayesian Approach*

Siddhartha Chib†
Minchul Shin‡
Anna Simoni§

July 5, 2024

Abstract

A standard assumption in the Bayesian estimation of linear regression models is that the regressors are exogenous in the sense that they are uncorrelated with the model error term. In practice, however, this assumption can be invalid. In this paper, under the rubric of the exponentially tilted empirical likelihood, we develop a Bayes factor test for endogeneity that compares a base model that is correctly specified under exogeneity but misspecified under endogeneity against an extended model that is correctly specified in either case. We provide a comprehensive study of the log-marginal exponentially tilted empirical likelihood. We demonstrate that our testing procedure is consistent from a frequentist point of view: as the sample becomes large, it almost surely selects the base model if and only if the regressors are exogenous, and the extended model if and only if the regressors are endogenous. 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 $y \in \mathbb{R}$ is the response, $x \in \mathbb{R}^{d_x}$ is the treatment vector of interest, $z_1 \in \mathbb{R}^{d_{z_1}}$ is a vector of controls and ε is the idiosyncratic noise. A standard assumption in the Bayesian estimation of such models is that the regressors x are exogenous in the sense that they are uncorrelated with the error term ε. 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 $z_2 \in \mathbb{R}^{d_{z_2}}$, 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).

A missing element in the existing Bayesian literature is a test for the exogeneity/endogeneity of the regressors. To fill this gap, in this paper, we derive the first Bayesian test for the endogeneity of regressors. This test is based on the marginal likelihoods of two models: a base model that is defined by 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 $\theta := (\beta, \gamma)$, and an extended model in which the exogeneity condition of the base model is amended to $\mathbf{E}[\varepsilon(\theta)x] = v$. In the extended model, v is an additional parameter that accounts for the covariance between the error and x. Analysis of each model is based on the exponentially tilted empirical likelihood (ETEL), avoiding parametric distributional assumptions.

We use the Bayes factor of the extended model versus the base model to test whether x is endogenous. The Bayes factor is the ratio of the marginal likelihoods of the extended and base models. We show that, with probability approaching one as the sample size increases, the Bayes factor selects the base model if and only if x is exogenous, and the extended model if and only if x is endogenous. These results are proven through a detailed large-sample analysis of the marginal likelihoods of the two models. Starting from the expression of the log-marginal likelihood given by

the Chib (1995) identity - as the sum of the log ETEL, the log prior, and the negative log posterior of the parameters, each evaluated at a particular point in the parameter space - we show that the log-marginal likelihoods of each model are asymptotically proportional to the Kullback-Leibler divergence between the true data distribution and the closest distribution satisfying the moment restrictions, plus a BIC penalty term. We provide a new derivation of this penalty by expressing the posterior ordinate of the parameters at the true or pseudo-true value, by a change of variable, as the posterior ordinate of a local parameter. The log of the Jacobian of this transformation is the penalty, while the ordinate of the posterior density of the local parameter at zero is bounded in probability as $n \to \infty$.

Model consistency of the Bayes factor arises because the first and third terms that characterize the log-marginal likelihood behave differently depending on whether x is endogenous or exogenous. When x is exogenous, the average log-ETEL of the two models is asymptotically the same, but the penalties differ. When x is endogenous, model consistency of the test is driven by the differing behaviors of the respective log-ETEL values, which dominate any difference in the penalties. Thus, the difference in penalties asymptotically plays no role in this case.

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 Bayes factor test we provide is a Bayesian analogue of the Hausman test.

The rest of the paper is organized as follows. In Section 2 we summarize in general terms the Bayesian estimation and comparison of moment condition models using the exponentially tilted empirical likelihood. In Section 3 we present the base and extended models as well as finite-sample results from a simulated data example to illustrate the implementation of our procedure. Section 4 describes our test for endogeneity/exogeneity. Then, it analyses the large-sample behavior of the log-marginal likelihood and establishes consistency of the testing procedure. A simulation exercise is provided. Finally, in Section 5 we consider three examples using real data. Concluding remarks

are given in Section 6. An appendix collects the proofs of the main results. Additional results and their proofs are in the Supplementary Appendix.

2 Preliminaries

In this section we briefly provide the background on Bayesian estimation of moment condition models using the exponentially tilted empirical likelihood (ETEL). Further details can be found in Schennach (2005) and Chib et al. (2018).

Consider a random vector $w \in \mathbb{R}^{d+1}$, a vector of parameters $\theta \in \Theta \subset \mathbb{R}^p$ with $p = d_x + d_{z_1}$, and let $\mathbb{M}$ denote the set of all probability distributions on $\mathbb{R}^d$. For a known function $g(w, \theta) : \mathbb{R}^{d+1} \times \mathbb{R}^p \to \mathbb{R}^d$, let

$$\mathbf{E}^Q[g(w, \theta)] = 0$$

denote a vector of moment conditions where Q is an element of the subset of distributions

$$\mathcal{Q}_\theta = \Big\{ Q \in \mathbb{M} : \mathbf{E}^Q[g(w, \theta)] = 0 \Big\} \tag{2.1}$$

that satisfy the moment conditions for a given $\theta \in \Theta$.

Suppose now that the data $w_{1:n} := (w_1, w_2, \ldots, w_n)$ are independently drawn from the true distribution P (that does not necessarily belong to $\mathcal{Q}_\theta$ for some $\theta \in \Theta$). To find the empirical counterpart of $Q \in \mathcal{Q}_\theta$, $\theta \in \Theta$, consider the discrete distribution $\{\widehat{q}_i(\theta)\}_{i=1}^n$, with support on $\{w_i,\, i = 1, \ldots, n\}$, that is the nearest in the Kullback-Leibler (KL) discrepancy to the empirical distribution that places probability mass $\{\frac{1}{n}\}$ on each observation. Enforcing the requirement that the moment restrictions are satisfied under this discrete distribution, the probability masses emerge as the solution to the optimization program

$$\begin{aligned} \{\widehat{q}_i(\theta)\} :=& \arg \max_{q_1, \ldots, q_n} \sum_{i=1}^n [-q_i \log(n q_i)] \\ \text{subject to } \sum_{i=1}^n q_i =& 1, \qquad \text{and} \qquad \sum_{i=1}^n q_i g(w_i, \theta) = 0, \quad \forall \theta \in \Theta. \end{aligned} \tag{2.2}$$

The ETEL is the likelihood constructed from this discrete distribution. It is defined as

$$\widehat{q}(w_{1:n}|\theta) = \prod_{i=1}^{n} \widehat{q}_i(\theta)$$

and is the the joint density of the observations after integrating out Q with respect to a particular nonparametric prior that imposes the moment restrictions for a given $\theta \in \Theta$ as demonstrated in Schennach (2005). Let $\pi(\theta)$ denote the prior density of the parameters. Then, the ETEL-based posterior distribution is given by the truncated distribution

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

where $I[A]$ denotes the indicator function, and 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 (2.2). The posterior $\pi^n(\theta|w_{1:n})$ is not available in closed form, but it can be summarized by tailored MCMC methods.

A convenient way to compute $\{\widehat{q}_i(\theta)\}$ is from the dual of (2.2). If we let

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

then

$$\widehat{q}_i(\theta) = \frac{e^{\widehat{\lambda}(\theta)' g(w_i, \theta)}}{\sum_{j=1}^n e^{\widehat{\lambda}(\theta)' g(w_j, \theta)}}, \qquad i \leq n. \tag{2.4}$$

The population counterpart of $\{\widehat{q}_i(\theta)\}$ is the distribution $Q^*(\theta) \in \mathcal{Q}_\theta$ that is closest to P in the KL divergence. It is defined as

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

where $\text{KL}(Q||P) := \int \log\left(\frac{dQ}{dP}\right) dQ$ is the KL discrepancy of Q from P if Q is absolutely continuous with respect to P and is equal to $+\infty$, otherwise. The population counterpart of $\widehat{\lambda}(\theta)$

is

$$\lambda_*(\theta) := \arg\min_{\lambda \in \mathbb{R}^d} \mathbf{E}[e^{\lambda' g(w_i,\theta)}].$$

If one or more moment conditions are misspecified, then $Q^*(\theta) \neq P$ for any $\theta \in \Theta$ and the pseudo-true value θ_* is defined as the value of θ that minimizes $\mathrm{KL}(P||Q^*(\theta))$ over Θ. Notice the inversion of the probabilities in the KL discrepancies used to define $Q^*(\theta)$ and θ_*. Under correct specification, $Q^*(\theta_\circ) = P$, for some $\theta_\circ \in \Theta$, and $\theta_* = \theta_\circ$.

When the dual representation of the optimization problem (2.2) holds, the pseudo-true value can also be obtained as

$$\theta_* = \arg\max_{\theta \in \Theta} \mathbf{E} \log\Big(\frac{e^{\lambda_*'(\theta) g(w,\theta)}}{\mathbf{E}[e^{\lambda_*'(\theta) g(W,\theta)}]}\Big), \tag{2.5}$$

where in this case the term within the brackets is $[dQ^*(\theta)/dP](w)$.

3 Models

In this section, we link the general setting presented in Section 2 to the semiparametric linear regression presented in Section 1. Let $w := (y, x, z_1, z_2) \in \mathbb{R}^{d+1}$ follow the unknown probability distribution P, where $d := d_x + d_{z_1} + d_{z_2}$. Let $\mathbf{E}[\cdot] := \mathbf{E}^P[\cdot]$ denote the expectation with respect to P. Now suppose that under P and $\theta_\circ := (\beta_\circ, \gamma_\circ)$, w follows the regression model

$$y = \beta_\circ' x + \gamma_\circ' z_1 + \varepsilon, \qquad \mathbf{E}[\varepsilon_i(\theta_\circ) z_{j,i}] = 0 \quad \text{for } j = 1, 2, \tag{3.1}$$

that is, $z_{j,i}$, $j = 1, 2$ are exogenous vectors, z_1 is a vector of controls and z_2 is a vector of instrumental variables. Suppose that the intercept is contained in z_1. The focus is on the causal effect of x on y, captured by the parameter $\beta_\circ$. Let

$$\begin{aligned} \varepsilon(\theta) :&= y - \beta' x - \gamma' z_1 \\ &\equiv y - \theta' \tilde{w} \ , \end{aligned}$$

where $\tilde{w}_{1,i} := (x, z_{1i})$, $\theta := (\beta, \gamma) \in \Theta \subset \mathbb{R}^p$ ($p := d_x + d_{z_1}$). Under the assumption that $d_{z_2} \geq d_x$, the instruments help to identify $\beta_\circ$ when $\mathbf{E}[\varepsilon_i(\theta_\circ)x_i] \neq 0$.

Base model: The base model is defined by the moment conditions

$$\mathbf{E}^Q[g_b(w, \theta)] = 0$$

where

$$g_b(w, \theta) := \varepsilon(\theta) \begin{pmatrix} x \\ z_1 \\ z_2 \end{pmatrix}$$

and Q is an element of the subset of distributions:

$$\mathcal{Q}_{b,\theta} = \Big\{ Q \in \mathbb{M};\ \mathbf{E}^Q[g_b(w, \theta)] = 0 \Big\} \tag{3.2}$$

that satisfy the moment conditions for a given $\theta \in \Theta$. The expectation is with respect to a distribution $Q \in \mathcal{Q}_{b,\theta}$ (as opposed to P) because, if x is endogenous under P, there is no θ that satisfies the moment conditions under P. That is, $P \notin \mathcal{Q}_{b,\theta}$. In this case, the ETEL function, constructed from the sample $w_{1:n}$, solves the empirical counterpart of the moment conditions:

$$\mathbf{E}^{Q_b^*(\theta)}[g_b(w, \theta)] = 0$$

where, for every θ, $Q_b^*(\theta)$ is the distribution in the set $\mathcal{Q}_b(\theta)$ closest to P in the KL divergence, that is,

$$Q_b^*(\theta) := \mathrm{arginf}_{Q \in \mathcal{Q}_{b,\theta}} \mathrm{KL}(Q \| P).$$

In addition,

$$\theta_* = \arg\max_{\theta \in \Theta} \mathbf{E} \log\Big(\frac{e^{\lambda_*'(\theta) g_b(w,\theta)}}{\mathbf{E}[e^{\lambda_*'(\theta) g_b(W,\theta)}]} \Big) \tag{3.3}$$

denotes the pseudo-true value in the base model. On the other hand, if x is exogenous, then

$Q_b^*(\theta_*) = P$ and $\theta_* = \theta_\circ$, where $\theta_\circ$ denotes the true value of θ.

Extended model: We define the extended model by the moment conditions

$$\mathbf{E}^Q[g_e(w,\psi)] = 0, \quad Q \in \mathcal{Q}_{e,\psi}, \tag{3.4}$$

where

$$g_e(w,\psi) := \varepsilon(\theta)\begin{pmatrix} x \\ z_1 \\ z_2 \end{pmatrix} - \begin{pmatrix} v \\ 0 \\ 0 \end{pmatrix} = g_b(w,\theta) - \begin{pmatrix} v \\ 0 \\ 0 \end{pmatrix},$$

$\psi := (\theta, v) \in \Psi$, $\Psi := \Theta \times \mathcal{V}$ with $\mathcal{V} \subset \mathbb{R}^{d_x}$, and $\mathcal{Q}_{e,\psi} := \left\{Q \in \mathbb{M};\ \mathbf{E}^Q[g_e(w,\psi)] = 0\right\}$. In this model, $v := \mathbf{E}[\varepsilon(\theta)x]$ is the covariance between the error and x.

Note that the extended model is correctly specified under both endogeneity and exogeneity of x. For instance, if $\mathbf{E}[\varepsilon(\theta)x] \neq 0$ for every $\theta \in \Theta$, while $\mathbf{E}[\varepsilon(\theta_\circ)(z_1', z_2')'] = 0$, then v will be equal to $\mathbf{E}[\varepsilon(\theta_\circ)x]$, and (3.4) is satisfied for this $\theta_\circ$. In the following, we use the notation $v_\circ = \mathbf{E}[\varepsilon(\theta_\circ)x]$. Therefore, the minimizer, $Q_e^*(\psi) = \operatorname{arginf}_{Q \in \mathcal{Q}_{b,\theta}} \mathrm{KL}(Q\|P)$, is equal to P, and the population moment conditions in the extended model are

$$\mathbf{E}^P[g_e(w,\psi_\circ)] = 0.$$

Moreover,

$$\psi_\circ = \arg\max_{\psi \in \Psi} \mathbf{E} \log\left(\frac{e^{\lambda_*'(\psi) g_e(w,\psi)}}{\mathbf{E}[e^{\lambda_*'(\psi) g_e(W,\psi)}]}\right), \tag{3.5}$$

where $\lambda_*(\psi) := \arg\min_{\lambda \in \mathbb{R}^d} \mathbf{E}[e^{\lambda' g_e(w,\psi)}]$.

3.1 Numerical illustration

To illustrate the fitting of the base and extended models, consider first the base model under endogeneity. Let the DGP be

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

for $i = 1, \ldots, n$, where $n \in \{250, 500, 1000, 2000\}$. Suppose that the (u_i, v_i, ω_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 ω_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 uncorrelated with ε_i and correlated with x_i but since ω_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 conditions

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

The ETEL posterior is sampled by the tailored one-block M-H algorithm (Chib and Greenberg, 1995) 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. In the extended (correctly specified)

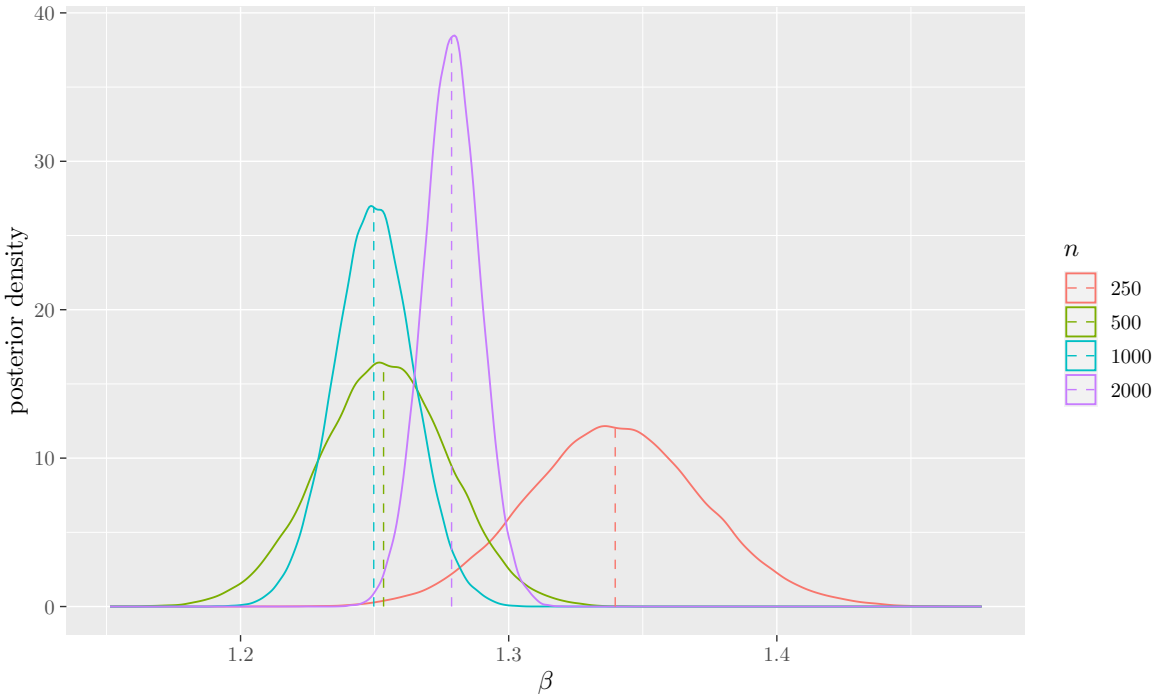

Figure 1: Base model under neglected endogeneity: Marginal posterior densities of β for different sample sizes. Posterior mean is indicated by dashed vertical line.

model we have

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

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 and those of v are in Figure 3. 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. In addition, the posterior of v, the cov(x, ε), tends to concentrate around the true value of 0.6.

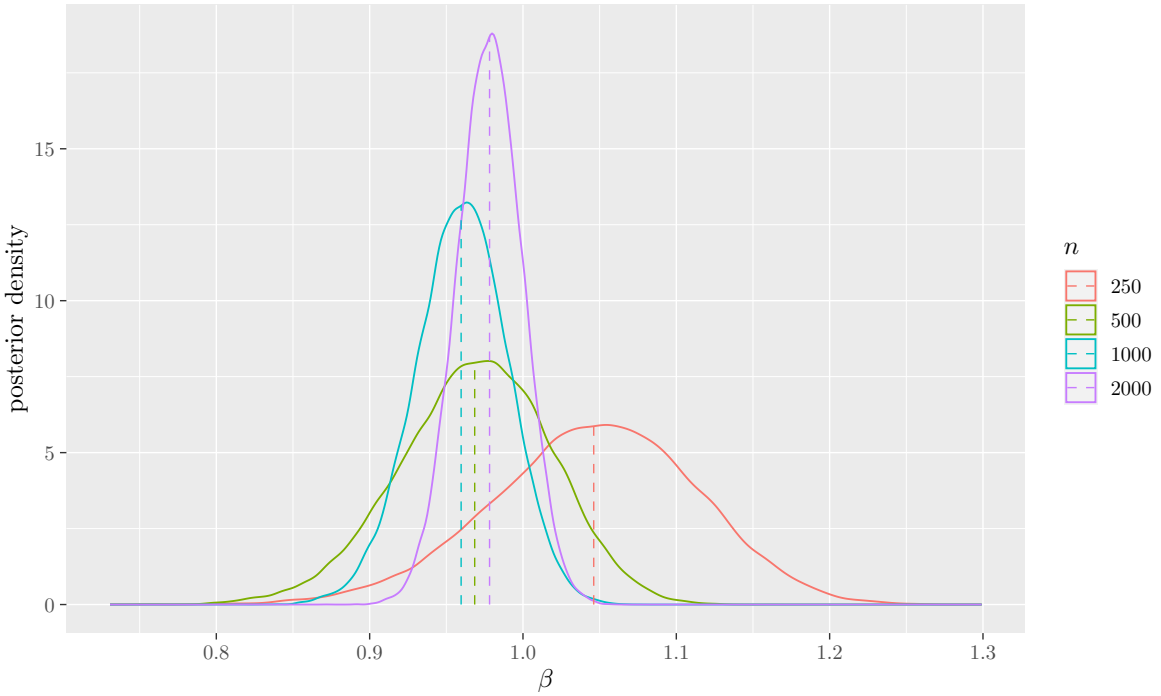

Figure 2: Extended model (x_i moment is inactive): Marginal posterior densities of β for different sample sizes. Posterior mean is indicated by dashed vertical line.

4 Testing procedure

4.1 Bayes factor

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

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

where $m(w_{1:n}|\mathcal{M}_b) := \int \widehat{q}(w_{1:n}|\theta, \mathcal{M}_b)\pi(\theta)d\theta$ and $m(w_{1:n}|\mathcal{M}_e) := \int \widehat{q}(w_{1:n}|\psi, \mathcal{M}_e)\pi(\psi)d\psi$ are the model marginal likelihoods arising from the ETEL functions (also called marginal ETEL functions later on). We compute these by the method of Chib (1995), as extended to general M-H chains in Chib and Jeliazkov (2001). We select $\mathcal{M}_e$ over $\mathcal{M}_b$ if $\log(\text{BF}_{eb}) > 0$, and select $\mathcal{M}_b$ otherwise.

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

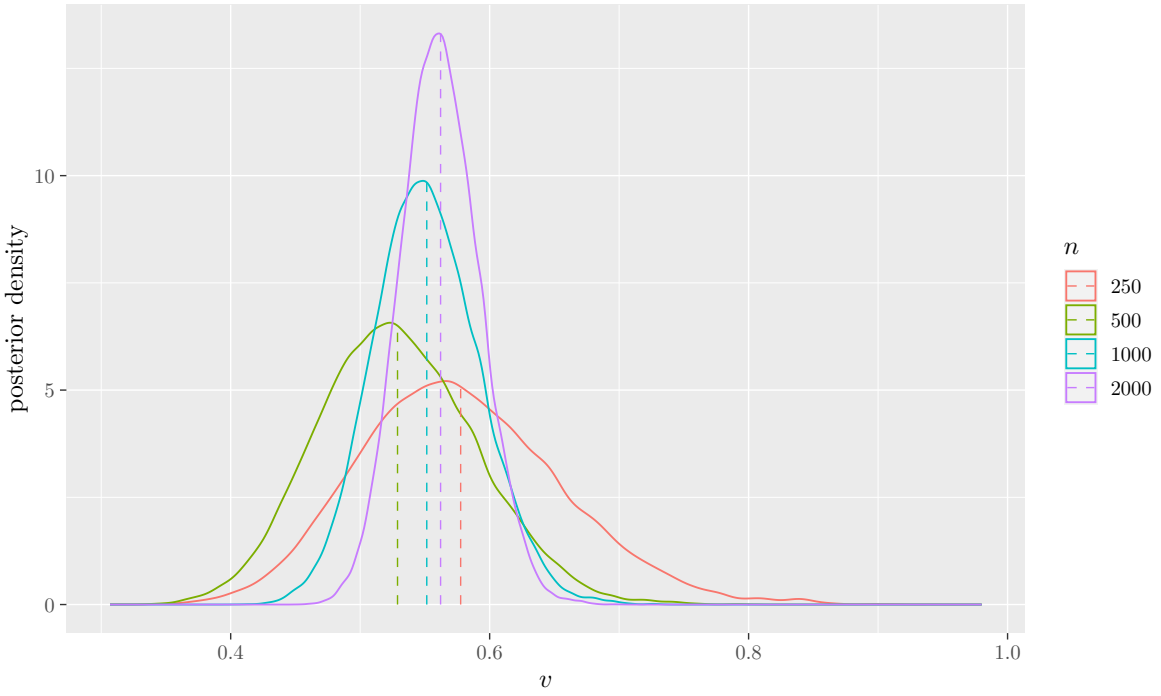

Figure 3: Extended model under neglected endogeneity: Marginal posterior densities of $v = \text{cov}(x, \varepsilon)$ for different sample sizes. Posterior mean is indicated by dashed vertical line.

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.

Intuitively, the Bayes factor picks the correct model because $\mathcal{M}_b$ 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).

4.2 Understanding the marginal likelihood

In this section we explain the rational behind our testing procedure. The hypothesis that we want to test is the following one:

$$H_{miss}: \quad P \text{ is such that } \nexists \theta \in \mathbb{R}^p \text{ such that } \mathbf{E}^P[\varepsilon_i(\theta)x_i] = 0$$

against

$$H_{cs}: \quad P \text{ is such that } \exists \theta \in \mathbb{R}^p \text{ such that } \mathbf{E}^P[\varepsilon_i(\theta)x_i] = 0.$$

Here, the subscripts $miss$ and cs are for misspecification and correct specification, respectively. We notice that the previous hypothesis can equivalently be written as $H'_{miss} : v \neq 0$ and $H'_{cs} : v = 0$. Our approach based on BF_{eb} is equivalent to a Bayes test for H'_{miss} versus H'_{cs} based on a prior on v of the type $\pi_0\delta_0(v) + (1-\pi_0)\pi(v)$, where $\delta_0(\cdot)$ denotes a Dirac mass on zero and $\pi(\cdot)$ is a continuous distribution. The two Bayes factors for these two approaches are numerically the same. The testing procedure works as follows: if $\text{BF}_{eb} \geq 1$ then accept H_{miss}, if $\text{BF}_{eb} < 1$ then accept H_{cs}.

The next theorem shows that H_{miss} and H_{cs} can be expressed in terms of Kullback-Leibler divergences between P and the set $\mathcal{Q}_{b,\theta}$ of distributions that satisfy the moment restriction that we want to test as well as additional moment restrictions that are known to hold for P.

Theorem 4.1 *Suppose that there is a $\theta \in \Theta$ such that $\mathbf{E}^P[\varepsilon_i(\theta)(z'_{1,i}z'_{2,i})'] = 0$. Consider the following statements:*

(i). P *is such that* $\nexists \theta \in \Theta$ *such that* $\mathbf{E}^P[\varepsilon_i(\theta)x_i] = 0$

(ii). $\mathrm{KL}(P||Q_b^*(\theta_*)) > 0$

(iii). P *is such that* $\exists \theta \in \Theta$ *such that* $\mathbf{E}^P[\varepsilon_i(\theta)x_i] = 0$

(iv). $\mathrm{KL}(P||Q_b^*(\theta_*)) = 0$.

Then, (i) is equivalent to (ii) and (iii) is equivalent to (iv).

This theorem makes clear that to test H_{miss} and H_{cs} one can equivalently focus on the Kullback-Leibler divergence $\mathrm{KL}(P||Q_b^*(\theta_*))$. Our Bayes test is based on Bayes factor and comparison of marginal likelihoods. There is a strict link between marginal likelihood and the Kullback-Leibler divergence: log-marginal likelihood of the base model behaves asymptotically as $-n\mathrm{KL}(P||Q_b^*(\theta_*))$ plus a penalty term, where the penalty depends on the number of parameters to estimate, and similarly for the log-marginal likelihood of the extended model. We are going to demonstrate this fact in the rest of this section.

From the Chib (1995) identity, we have for the base model: $\forall \theta \in \Theta \subset \mathbb{R}^p$,

$$\log m(w_{1:n}|\mathcal{M}_b) = \log \pi(\theta|\mathcal{M}_b) + \log \widehat{q}(w_{1:n}|\theta, \mathcal{M}_b) - \log \pi^n(\theta|w_{1:n}, \mathcal{M}_b),$$

and similarly for the extended model. Because this identity is true for every $\theta \in \Theta$, it is true for $\theta = \theta_*$: $\log m(w_{1:n}|\mathcal{M}_b) = \log \pi(\theta_*|\mathcal{M}_b) + \log \widehat{q}(w_{1:n}|\theta_*, \mathcal{M}_b) - \log \pi^n(\theta_*|w_{1:n}, \mathcal{M}_b)$. Next, let us introduce the local parameters $h_\theta := \sqrt{n}(\theta - \theta_*)$ and $h_\psi := \sqrt{n}(\psi - \psi_\circ)$, so that by the formula for transformation of random variables: $\pi^n(\theta|w_{1:n}, \mathcal{M}_b) = \pi^n_{h_\theta}(\sqrt{n}(\theta - \theta_*)|w_{1:n}, \mathcal{M}_b)n^{p/2}$ and $\pi^n(\psi|w_{1:n}, \mathcal{M}_e) = \pi^n_{h_\psi}(\sqrt{n}(\psi - \psi_\circ)|w_{1:n}, \mathcal{M}_e)n^{(p+d_x)/2}$, where $\pi^n_{h_\theta}(\cdot|w_{1:n}, \mathcal{M}_b)$ and $\pi^n_{h_\psi}(\cdot|w_{1:n}, \mathcal{M}_e)$ denote the posterior density of h_θ and h_ψ respectively. By replacing this in the expression of the marginal likelihoods we obtain:

$$\begin{aligned}\log m(w_{1:n}|\mathcal{M}_b) &= \log \pi(\theta|\mathcal{M}_b) + \log \widehat{q}(w_{1:n}|\theta, \mathcal{M}_b) - \log \pi^n_{h_\theta}(\sqrt{n}(\theta - \theta_*)|w_{1:n}, \mathcal{M}_b) - \frac{p}{2}\log(n) \\ &= \log \pi(\theta_*|\mathcal{M}_b) + \log \widehat{q}(w_{1:n}|\theta_*, \mathcal{M}_b) - \log \pi^n_{h_\theta}(0|w_{1:n}, \mathcal{M}_b) - \frac{p}{2}\log(n),\end{aligned} \tag{4.1}$$

and

$$\begin{aligned}\log m(w_{1:n}|\mathcal{M}_e) &= \log \pi(\psi|\mathcal{M}_e) + \log \widehat{q}(w_{1:n}|\psi, \mathcal{M}_e) \\ &\quad - \log \pi^n_{h_\psi}(\sqrt{n}(\psi - \psi_\circ)|w_{1:n}, \mathcal{M}_e) - \frac{p + d_x}{2}\log(n) \\ &= \log \pi(\psi_\circ|\mathcal{M}_e) + \log \widehat{q}(w_{1:n}|\psi_\circ, \mathcal{M}_e) - \log \pi^n_{h_\psi}(0|w_{1:n}, \mathcal{M}_e) - \frac{p + d_x}{2}\log(n).\end{aligned} \tag{4.2}$$

The intuition for expressing the posterior of θ in terms of the posterior of the local parameter is that the Jacobian of the transformation makes explicit the role played by the dimension of the model while the local parameter has a posterior distribution that is approximately Gaussian. This is true in both cases *(i)* and *(iii)* of Theorem 4.1.

Therefore, the log-marginal likelihood is equal to the sum of two terms that are bounded in probability as $n \to \infty$ – these are the prior of θ, or ψ, and the posterior of the local parameter–, and two terms that are diverging with n: the log-ETEL and the term involving the dimension of the model. Consequently, asymptotically the marginal likelihood behaves like a penalized log-ETEL criterion where, remarkably, the penalization is coming from the prior distribution and is not ad-hoc.

Of course, to be sure that a testing procedure based on marginal likelihood works, one has to show that $\pi^n_{h_\theta}(\sqrt{n}(\theta - \theta_*)|w_{1:n}, \mathcal{M}_b)$ and $\pi^n_{h_\psi}(\sqrt{n}(\psi - \psi_\circ)|w_{1:n}, \mathcal{M}_e)$ are bounded in probability as $n \to \infty$ on the support of the prior. This can be quite challenging, in particular in non-standard models like the one we are considering here where there is no parametric likelihood and where models can be misspecified. We provide these results in Theorems C.6 and C.7 in the Supplementary Material, which are refinements of Theorems 1 and 2 in Chib et al. (2018). Compared to the latter, we offer a new proof of the stochastic local asymptotic normality (LAN) of the log-ETEL (see Theorems C.1, C.2 and C.3 in the Supplementary Material) that is more direct because it exploits the specific structure of the instrumental variable regression problem. Stochastic LAN is an essential step to prove asymptotic normality of the posterior distribution of the local parameter.

The final step in order to have a clear comprehension of the marginal likelihood is provided by

Theorems 4.2 and 4.3 below which supply an asymptotic expression for the log-ETEL in the base and extended models. For simplicity, we recall the expressions of the log-ETEL function for one observation w_i:

$$\log \widehat{q}_i(\theta|\mathcal{M}_b) = \log \frac{e^{\widehat{\lambda}(\theta)' g_b(w_i,\theta)}}{\sum_{k=1}^n e^{\widehat{\lambda}(\theta)' g_b(w_k,\theta)}}, \qquad \log \widehat{q}_i(\psi|\mathcal{M}_e) = \log \frac{e^{\widehat{\lambda}(\psi)' g_e(w_i,\theta)}}{\sum_{k=1}^n e^{\widehat{\lambda}(\psi)' g_e(w_k,\theta)}}$$

so that the log-ETEL functions are $\log \widehat{q}(w_{1:n}|\theta, \mathcal{M}_b) := \sum_{i=1}^n \log \widehat{q}_i(\theta|\mathcal{M}_b)$ and $\log \widehat{q}(w_{1:n}|\psi, \mathcal{M}_e) := \sum_{i=1}^n \log \widehat{q}_i(\psi|\mathcal{M}_e)$, respectively. The assumptions under which the following results hold are relegated to section 4.4 in order to facilitates readability and because they are standard assumptions.

Theorem 4.2 (Base model.) *Let Assumptions 1 - 4 and 5 (d)-(f) hold. Then,*

$$\begin{aligned}\log \widehat{q}(w_{1:n}|\theta_*, \mathcal{M}_b) = &-n \log n - \mathbb{G}_n \left[\tau_i^\dagger(\lambda_*, \theta_*) g_b(w_i, \theta_*)'\right] \Omega_*^\dagger(\theta_*)^{-1} \mathbb{G}_n \left[g_b(w_i, \theta_*)\right] \\ &+ \sum_{i=1}^n \log \left(\frac{e^{\lambda_*(\theta_*)' g_b(w_i,\theta_*)}}{\mathbf{E}[e^{\lambda_*(\theta_*)' g_b(w_j,\theta_*)}]} \right) + n(\widehat{\lambda}(\theta_*) - \lambda_*(\theta_*))' \mathbf{E}[g_b(w_i, \theta_*)] \\ &+ \frac{1}{2} \mathbb{G}_n \left[\tau_i^\dagger(\lambda_*, \theta_*) g_b(w_i, \theta_*)'\right] \Omega_*^\dagger(\theta_*)^{-1} \mathbb{G}_n \left[\tau_i^\dagger(\lambda_*, \theta_*) g_b(w_i, \theta_*)\right] + o_p(1), \quad (4.3)\end{aligned}$$

where $\mathbb{G}_n \left[\tau_i^\dagger(\lambda_*, \theta_*) g_b(w_i, \theta_*)\right] \xrightarrow{d} \mathcal{N}(0, \Omega_*^\dagger(\theta_*))$, $\tau_i^\dagger(\lambda_*, \theta_*) := [dQ^*(\theta_*)/dP](w_i)$ *and* $\Omega_*^\dagger(\theta_*) := \mathbb{E}^{Q^*(\theta_*)}[\varepsilon_i(\theta_*)^2 \widetilde{w}_i \widetilde{w}_i']$.

For the extended model, we recall that $\psi_\circ = (\theta_\circ', v_\circ')'$ denotes the true value of the parameter in the extended model with $v_\circ = \mathbb{E}[\varepsilon(\theta_\circ)x]$.

Theorem 4.3 (Extended model.) *Let Assumptions 2, 3 with* Θ *replaced by* Ψ*, 4 and 5 (d)-(f) with* θ_* *and* $\lambda_*(\theta_*)$ *replaced with* $\theta_\circ$ *and* 0*, respectively, hold. Then,*

$$\log \widehat{q}(w_{1:n}|\psi_\circ, \mathcal{M}_e) = -n \log n - \frac{1}{2} \mathbb{G}_n \left[g_e(w_i, \psi_\circ)'\right] \Omega_{\psi_\circ}^{-1} \mathbb{G}_n \left[g_e(w_i, \psi_\circ)\right] + o_p(1), \quad (4.4)$$

where $\Omega_{\psi_\circ} := \mathbf{E}[g_e(w_i, \psi_\circ) g_e(w_i, \psi_\circ)']$*, and* $\mathbb{G}_n \left[g_e(w_i, \psi_\circ)'\right] \Omega_{\psi_\circ}^{-1} \mathbb{G}_n \left[g_e(w_i, \psi_\circ)\right] \xrightarrow{d} \chi_d^2$*, where* χ_d^2 *denotes a chi square distribution with* d *degrees of freedom.*

It is clear that if $\mathbf{E}[\varepsilon_i(\theta_\circ)x_i] = 0$, so that the assumptions in Theorem 4.2 hold with θ_* replaced by $\theta_\circ$, then

$$\log \widehat{q}(w_{1:n}|\theta_\circ, \mathcal{M}_b) = -n\log n - \frac{1}{2}\mathbb{G}_n\left[g_b(w_i,\theta_\circ)'\right]\Omega_\circ^{-1}\mathbb{G}_n\left[g_b(w_i,\theta_\circ)\right] + o_p(1), \tag{4.5}$$

where $\Omega_\circ = \mathbb{E}[\varepsilon_i(\theta_\circ)^2\widetilde{w}_i\widetilde{w}_i']$ because $\lambda_*(\theta_*) = \lambda_*(\theta_\circ) = 0$, and moreover $\mathbb{G}_n\left[g(w_i,\theta_\circ)'\right]\Omega_\circ^{-1}\mathbb{G}_n\left[g(w_i,\theta_\circ)\right] \xrightarrow{d} \chi^2_d$, where χ^2_d denotes a chi square distribution with d degrees of freedom. Similarly, if $\mathbf{E}[\varepsilon_i(\theta_\circ)x_i] = 0$, then

$$\log \widehat{q}(w_{1:n}|\psi_\circ, \mathcal{M}_e) = -n\log n - \frac{1}{2}\mathbb{G}_n\left[g_b(w_i,\theta_\circ)'\right]\Omega_\circ^{-1}\mathbb{G}_n\left[g_b(w_i,\theta_\circ)\right] + o_p(1) \tag{4.6}$$

and $\mathbb{G}_n\left[g_b(w_i,\psi_\circ)'\right]\Omega_{\psi_\circ}^{-1}\mathbb{G}_n\left[g_b(w_i,\psi_\circ)\right] \xrightarrow{d} \chi^2_d$. Hence, when x is exogenous, $\log \widehat{q}(w_{1:n}|\theta_\circ, \mathcal{M}_b)$ and $\log \widehat{q}(w_{1:n}|\psi_\circ, \mathcal{M}_e)$ are equal asymptotically and they cancel in the comparison of the marginal likelihoods.

In case of endogeneity, instead, $\log \widehat{q}(w_{1:n}|\theta_\circ, \mathcal{M}_b)$ and $\log \widehat{q}(w_{1:n}|\psi_\circ, \mathcal{M}_e)$ are different and they play a central role in the comparison of marginal likelihoods. In this case, it is important to consider the behaviour of the average log-ETEL function. We then have the following two corollaries which are useful to relate the asymptotic behaviour of the average log-ETEL function to the Kullback-Leibler divergence. We point out that, while this type of results is implicit in the definition of the ETEL, we provide here a formal proof. In our setting, it is important to explicit these results because they allow us to understand the behaviour of the marginal likelihood.

Corollary 4.1 (Base model.) *Suppose Assumptions 1 - 4 and 5 (d)-(f) hold. Then, as* $n \to \infty$*,*

$$\frac{1}{n}\log \widehat{q}(w_{1:n}|\theta_*, \mathcal{M}_b) + \log(n) \xrightarrow{p} \mathbf{E}^P\left[\log(dQ_b^*(\theta_*)/dP)\right], \tag{4.7}$$

where $\mathbf{E}^P\left[\log(dQ_b^*(\theta_*)/dP)\right] = \mathbf{E}^P\left[\log\left(\frac{e^{\lambda_*(\theta_*)'g_b(w,\theta_*)}}{\mathbf{E}[e^{\lambda_*(\theta_*)'g_b(w,\theta_*)}]}\right)\right] = -\mathrm{KL}(P||Q_b^*(\theta_*))$.

Corollary 4.2 (Extended model.) *Suppose Assumptions 2, 3 with* Θ *replaced by* Ψ*, 4 and 5 (d)-(f)*

with θ_* and $\lambda_*(\theta_*)$ replaced with $\theta_\circ$ and 0, respectively, hold. Then, as $n \to \infty$,

$$\frac{1}{n}\log \widehat{q}(w_{1:n}|\psi_\circ, \mathcal{M}_e) + \log(n) \xrightarrow{p} \mathbf{E}^P\left[\log(dQ_e^*(\psi_\circ)/dP)\right], \tag{4.8}$$

where $\mathbf{E}^P\left[\log(dQ_e^*(\psi_\circ)/dP)\right] = \mathbf{E}^P\left[\log\left(\frac{e^{\lambda_*(\psi_\circ)'g(w,\psi_\circ)}}{\mathbf{E}[e^{\lambda_*(\psi_\circ)'g(w,\psi_\circ)}]}\right)\right] = \mathrm{KL}(P||Q_e^*(\psi_\circ))$.

Notice that $\mathbf{E}^P\left[\log(dQ_e^*(\psi_\circ)/dP)\right] = 0$ since the extended model is correctly specified and so $dQ_e^*(\psi_\circ)/dP = 1$.

From Theorems 4.2 and 4.3, and Theorems C.6 and C.7 in the Online Appendix and from (4.1)-(4.2), then there exists an N such that for every $n > N$:

$$\begin{aligned}
\log m(w_{1:n}|\mathcal{M}_b) &= -n\log n + \sum_{i=1}^{n}\log\left(\frac{e^{\lambda_*(\theta_*)'g_b(w_i,\theta_*)}}{\mathbf{E}[e^{\lambda_*(\theta_*)'g_b(w_j,\theta_*)}]}\right) \\
&\quad + n(\widehat{\lambda}(\theta_*) - \lambda_*(\theta_*))'\mathbf{E}[g_b(w_i,\theta_*)] - \frac{p}{2}\log(n), \qquad (4.9)\\
\log m(w_{1:n}|\mathcal{M}_e) &= -n\log(n) - \frac{p+d_x}{2}\log(n). \qquad (4.10)
\end{aligned}$$

From these expressions, one sees that when the models are both correctly specified, that is, $\mathbf{E}[\varepsilon_i(\theta_\circ)x_i] = 0$, then $\lambda_*(\theta_*) = 0$ and $\sum_{i=1}^n \log\left(\frac{e^{\lambda_*(\theta_*)'g_b(w_i,\theta_*)}}{\mathbf{E}[e^{\lambda_*(\theta_*)'g_b(w_j,\theta_*)}]}\right) = 0$ for every $n \in \mathbb{N}$. Therefore, it is clear that asymptotically $\log m(w_{1:n}|\mathcal{M}_b)$ is larger than $\log m(w_{1:n}|\mathcal{M}_e)$.

On the other hand, when there is no $\theta \in \Theta$ such that $\mathbf{E}[\varepsilon_i(\theta)x_i] = 0$, then $\lambda_*(\theta_*) \neq 0$ and $\sum_{i=1}^n \log\left(\frac{e^{\lambda_*(\theta_*)'g_b(w_i,\theta_*)}}{\mathbf{E}[e^{\lambda_*(\theta_*)'g_b(w_j,\theta_*)}]}\right)$ diverges to $-\infty$ faster than the last two terms in (4.9), so that asymptotically $\log m(w_{1:n}|\mathcal{M}_b)$ is smaller than $\log m(w_{1:n}|\mathcal{M}_e)$. This is the main intuition of the consistency results in Theorems 4.4 and 4.5 in the next section. The proof of these theorems is more complicated than this arguments because the theorems provide an if and only if statement which is stronger than consistency.

4.3 Consistency of the testing procedure

We now use the preceding theory to establish consistency of our testing procedure based on the Bayes factor constructed from the marginal ETEL functions. The theorems below establish that, as the sample size increases, BF_{eb} selects $\mathcal{M}_b$ if and only if x is exogenous, and selects $\mathcal{M}_e$ if and only if x is endogenous, with probability approaching one.

Theorem 4.4 *Let Assumptions 1, 2, 4, 6, 7 hold and let Assumptions 3 and 5 hold for θ_* and $\lambda_*(\theta_*)$ and also for θ_* and $\lambda_*(\theta_*)$ replaced with $\theta_\circ$ and 0, respectively. Let the priors on θ and ψ be continuous probability measures that admit densities with respect to the Lebesgue measure and that are positive on a neighborhood of θ_* and $\psi_\circ$, respectively. Let us consider the comparison of models $\mathcal{M}_b$ and $\mathcal{M}_e$. 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, the failure of the necessary and sufficient condition $\mathbf{E}[\varepsilon_i(\theta)x_i] = 0$ for any θ, is equivalent to the inequality $\mathrm{KL}(P||Q_e^*(\psi)) < \mathrm{KL}(P||Q_b^*(\theta))$, where $\mathrm{KL}(P||Q_e^*(\psi_\circ)) = 0$. 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. In the framework of the present paper, this means that the correctly specified model is selected.

Next, we show what happens when the variables x_i are exogenous so that the moment restriction $\mathbf{E}[\varepsilon_i(\theta)x_i] = 0$ holds for a particular value $\theta_\circ$ and the two models under comparison are correctly specified. The next theorem states that in this case the base model is selected. This is understandable through an argument of parsimony: the base model has the smaller number of parameters to estimate and so it is the preferred one when it is correctly specified.

Theorem 4.5 *Let Assumptions 1, 2, 4, 6, 7 hold and let Assumptions 3 and 5 hold for both θ_* and $\lambda_*(\theta_*)$ and also for θ_* and $\lambda_*(\theta_*)$ replaced with $\theta_\circ$ and 0, respectively. Let the priors on θ and ψ be continuous probability measures that admit densities with respect to the Lebesgue measure and that are positive on a neighborhood of θ_* and $\psi_\circ$, respectively. Let us consider the comparison of models $\mathcal{M}_b$ and $\mathcal{M}_e$. Then,*

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

if and only if there is a $\theta_\circ$ such that $\mathbf{E}[\varepsilon_i(\theta_\circ)x_i] = 0$ holds.

Discussion. In this and previous subsection, we demonstrate that our model selection criteria favor a model with a smaller Kullback-Leibler Information Criterion (KLIC). When two models share the same KLIC, our procedure opts for the model with a greater number of restrictions, i.e., a more parsimonious or less flexible model. Interestingly, this aligns with the goal of Sin and White (1996)'s penalized likelihood criteria for a parametric model. Consequently, our proposed model selection procedure in this paper and Chib et al. (2018) can be viewed as a fully Bayesian semi-parametric version of consistent model selection criteria, applied specifically to an endogeneity testing problem. Unlike other frequentist procedures, the 'penalty' term required for consistency is inherently built into our Bayesian calculation.

Andrews (1999), Andrews and Lu (2001), and Hong, Preston and Shum (2003) have proposed and studied model selection criteria for moment condition models, even though a formal likelihood function is not defined. These criteria involve a penalization term that is attached to the Generalized Method of Moments (GMM), and more broadly, the Generalized Empirical Likelihood (GEL) objective function, rather than the likelihood function. Examples of such frequentists model selection approaches based on GMM estimation can be found in Appendix A. However, the relationship between these model selection criteria and the KLIC minimization principle of Sin and White (1996) for potentially misspecified parametric models is not immediately apparent.

It is noteworthy that our procedure exhibits the same asymptotic behavior as Hong and Pre-

ston (2012)'s generalized empirical likelihood Bayes factor. They impose a separate prior on the Lagrangian multiplier that is independent of θ, which does not guarantee the imposition of moment restrictions. In contrast, we introduce an additional parameter v to the 'inactive' moment restriction, ensuring that our prior on θ and v respects the moment restrictions.

4.4 Assumptions

We provide the assumptions that we use to prove the results in the previous sections. The first assumption guarantees that the dual representation of the optimization problem (2.2) holds even when $P \notin \mathcal{Q}_{b,\theta}$ for every $\theta \in \Theta$. In fact, in the latter case it is possible that $Q_b^*(\theta)$ and P do not have a common support for any θ, in which case, the equality in (2.4) does not hold, see Sueishi (2013) for a discussion on this point.

Assumption 1 (Non-emptyness.) *When $\mathbf{E}^P[\varepsilon_i(\theta)x_i] \neq 0$ for every $\theta \in \Theta$, there exists $Q \in \bigcup_{\theta\in\Theta} \mathcal{Q}_{b,\theta}$ such that Q is mutually absolutely continuous with respect to P, where $\mathcal{Q}_{b,\theta}$ is defined in* (3.2).

This assumption implies that there is a θ for which $\mathcal{Q}_{b,\theta}$ is non-empty, that $dQ_b^*(\theta)/dP = \left(\frac{e^{\lambda_*(\theta)'g(w,\theta)}}{\mathbf{E}[e^{\lambda_*(\theta)'g(w,\theta)}]}\right)$ and that θ_* is identified by (3.3). We then assume that θ_* is unique.

Assumption 2 (Identification.) *The maximizer θ_* defined as the minimizer of $\mathrm{KL}(P||Q^*(\theta))$ with respect to $\theta \in \Theta$ is unique and is in the interior of Θ, where the interior is defined with respect to the topology in $\mathbb{R}^p$.*

Since under Assumption 1 θ_* coincides with the minimizer in (3.3), then the previous assumption implies uniqueness also of the latter. The next three assumptions concern the model. Recall the notation $w_i := (y_i, x_i', z_i')'$, $z_i := (z_{1,i}', z_{2,i}')'$ 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 3 *(a) w_i, $i = 1, \ldots, n$ 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 (3.1); (b) $\Theta \subset \mathbb{R}^p$ is compact and connected; (c) for every λ in a neighborhood of $\lambda_*(\theta_*)$, the matrix $\mathbf{E}[e^{\lambda' \widetilde{w}_i \varepsilon_i(\theta_*)} \varepsilon_i(\theta_*)^2 \widetilde{w}_i \widetilde{w}_i']$ has smallest (resp. largest) eigenvalue bounded away from zero (resp. infinity).

Assumption 4 *(a) $\mathbf{E}[\widetilde{w}_i \widetilde{w}_{1,i}'] < \infty$ with rank p.*

Assumptions 3 and 4 are standard in the literature, see *e.g.* Schennach (2007). The following assumption instead, is new and it is used to prove the approximation for the marginal likelihood. 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, where $\|\cdot\|_2$ denotes the Euclidean norm.

Assumption 5 *(a) For any $\delta > 0$ and every w_i, there is a function $\gamma_0(w_i)$ such that $\left\| e^{\lambda' \widetilde{w}_i \varepsilon_i(\theta_*)} \widetilde{w}_i \varepsilon_i(\theta_*) \right\|_2 \leq \gamma_0(w_i)$ 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 $\delta > 0$ and every w_i, there exists a function $\gamma_1(w_i)$ such that $\left| e^{\lambda' \widetilde{w}_i \varepsilon_i(\theta_*)} \right| \leq \gamma_1(w_i)$ 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 $\left| e^{\lambda' \widetilde{w}_i \varepsilon_i(\theta_*)} \varepsilon_i(\theta)^{j-1} \widetilde{w}_{i,k}^\ell (h' \widetilde{w}_{1,i})^{\ell'} \right| \leq \gamma_2(w_i)$ for every $\lambda \in B_\delta(\lambda_*(\theta_*))$, every w_i, 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*

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

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

For the next assumption denote by $\Theta_n := \{\|\theta - \theta_*\| \leq M_n/\sqrt{n}\}$, a ball around θ_* with radius at most $M_n/\sqrt{n}$, where M_n is any sequence of positive constants diverging to $+\infty$. We denote by $\ell_{n,\theta}(w_i)$ the log-likelihood function for one observation w_i: $\ell_{n,\theta}(w_i) := \log \widehat{q}_i(\theta|\mathcal{M}_b)$ and by $\ell_{n,\theta}(w_{1:n}) := \sum_{i=1}^n \ell_{n,\theta}(w_i) = \log \widehat{q}(w_{1:n}|\theta, \mathcal{M}_b)$ the log-ETEL function. The next assumption 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 the θs that are at a certain distance from it.

Assumption 6 (Base model.) *Assume that there exists a constant $C > 0$ such that*

$$P\left(\sup_{\theta \in \Theta_n^c} \frac{1}{n}\sum_{i=1}^n \left(\ell_{n,\theta}(w_i) - \ell_{n,\theta_*}(w_i)\right) \leq -\frac{CM_n^2}{n}\right) \to 1 \ , \ \textit{as } n \to \infty, \tag{4.11}$$

where M_n is the same sequence used to define Θ_n.

A condition similar to Assumption 6 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 (4.11) 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 $n \to 0$,

$$\mathbf{E}[\phi_n] \to 0, \qquad \text{and} \qquad \sup_{\{\theta; \|\theta - \theta_*\| \geq \epsilon\}} \mathbf{E}\left[e^{\ell_{n,\theta}(w_i) - \ell_{n,\theta_*}(w_i)}(1 - \phi_n)\right] \to 0. \tag{4.12}$$

Similarly, for the extended model we denote by $\ell_{n,\psi}(w_i)$ the log-likelihood function for one observation w_i: $\ell_{n,\psi}(w_i) := \log \widehat{q}_i(\psi|\mathcal{M}_e)$ and by $\ell_{n,\psi}(w_{1:n}) := \sum_{i=1}^n \ell_{n,\psi}(w_i) = \log \widehat{q}(w_{1:n}|\psi, \mathcal{M}_e)$ the log-ETEL function. The next assumption has the same interpretation of Assumption 6 but for the extended model.

Assumption 7 (Extended model.) *Assume that there exists a constant $C > 0$ such that as $n \to \infty$,*

$$P\left(\sup_{\|\psi-\psi_\circ\|>M_n/\sqrt{n}} \frac{1}{n}\sum_{i=1}^n \left(\ell_{n,\psi}(w_i) - \ell_{n,\psi_\circ}(w_i)\right) \leq -\frac{CM_n^2}{n}\right) \to 1, \tag{4.13}$$

where M_n is any sequence of positive constants diverging to infinity.

4.5 Experiments

Consider the same generating process as Example 1, and 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 $\mathcal{M}_e$.

ρ	-0.5	-0.4	-0.3	-0.2	-0.1	0.0	0.1	0.2	0.3	0.4	0.5
$n = 250$	99	96	82	48	12	2	18	54	93	100	100
$n = 500$	100	100	98	76	17	1	29	87	99	100	100
$n = 1000$	100	100	100	96	46	1	46	100	100	100	100
$n = 2000$	100	100	100	100	80	1	70	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 $trend_B j$,

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, and ten instruments, while in the augmented variant, the base and extended models have nineteen and twenty parameters, and 53 moment restrictions. We assume that the $n = 2217$ observations on $(y_{ijt}, x_{ijt}, z_{1ijt})$ are a random sample from the population of automobile products across markets and time. Because it is difficult to formulate 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 as the prior standard deviation. The ETEL is constructed from the remaining data and the posterior distribution of each model is sampled by the single block M-H algorithm of Chib and Greenberg (1995). This algorithm is fast and efficient despite the relatively large numbers of parameters and instruments. The results show that the posterior mean of the coefficient on $price$ is -0.14, and the 95% posterior credibility interval is (-.16,-.13). The posterior mean is larger in magnitude than the OLS estimate originally reported by BLP. Note that the posterior distribution of the covariance 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

	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).

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.

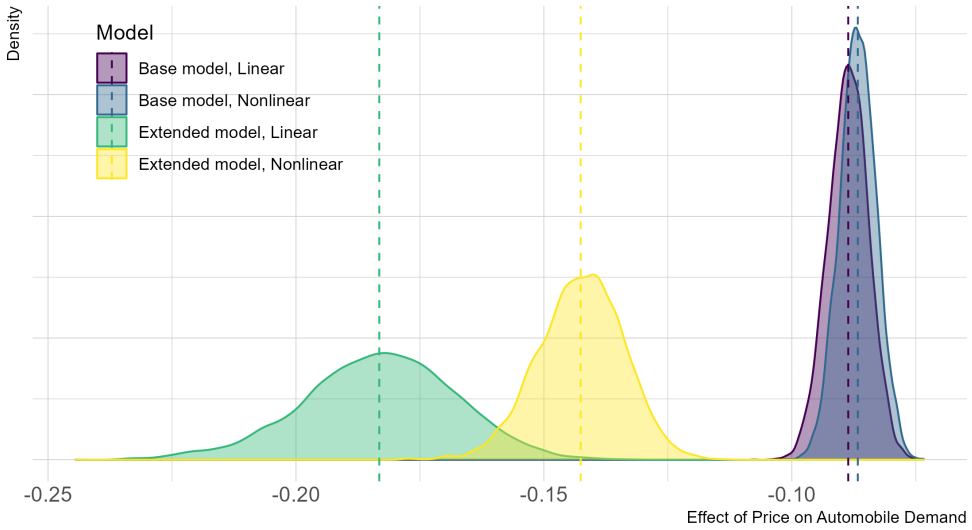

Figure 4: BLP models: Marginal posterior distributions of the coefficient on the price variable, β. Posterior mean and standard deviation of β are -0.089 and 0.004 for the base model with the original BLP (linear) specification while they are -0.087 and 0.004 with the augmented BLP (nonlinear) specification. 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 $y_i = (y_{i1}, \ldots, y_{iT})$ 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 a Bayesian test for exogeneity/endogeneity of the treatment vector of interest in a linear mean regression model. 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 testing procedure for exogeneity/endogeneity is based on Bayes factor where the marginal ETEL of the base and the extended models are compared. The procedure is validated from a frequentist point of view because we establish the large sample consistency of the Bayes factor test. In addition, we provide a comprehensive study of the log-marginal ETEL function and determine which parts of it plays a role in the testing procedure depending on whether the covariates x are exogenous or endogenous.

The real-data examples discussed in the paper showcase the practical relevance of the methods.

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 specification of the prior and 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

Andrews, D. W. K. (1999), 'Consistent Moment Selection Procedures for Generalized Method of Moments Estimation', *Econometrica* **67**(3), 543–563.

Andrews, D. W. and Lu, B. (2001), 'Consistent Model and Moment Selection Procedures for GMM Estimation with Application to Dynamic Panel Data Models', *Journal of Econometrics* **101**(1), 123–164.

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

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. (1995), 'Understanding the Metropolis-Hastings algorithm', *The American Statistician* **49**(4), 327–335.

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.

Hausman, J. A. (1978), 'Specification tests in econometrics', *Econometrica* **46**(6), 1251–1271.

Hong, H. and Preston, B. (2012), 'Bayesian averaging, prediction and nonnested model selection', *Journal of Econometrics* **167**(2), 358–369.

Hong, H., Preston, B. and Shum, M. (2003), 'Generalized Empirical Likelihood-Based Model Selection Criteria For Moment Condition Models', *Econometric Theory* pp. 923–943.

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.

Sin, C. and White, H. (1996), 'Information criteria for selecting possibly misspecified parametric models', *Journal of Econometrics* **71**(1-2), 207–225.

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

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

Online Appendix

A Comparison with GMM-based criteria

There are frequentist approaches to the model (or, moment) selection that can be applied in our context. Andrews (1999) develops a class of moment selection criteria (MSC). Below are some popular criteria that fall into the class:

$$\begin{aligned} \text{GMM-BIC} &= J_n(c) - (|c| - p)\ln n \\ \text{GMM-AIC} &= J_n(c) - 2(|c| - p) \\ \text{GMM-HQIC} &= J_n(c) - 2.01(|c| - p)\ln\ln n, \end{aligned} \tag{A.1}$$

where c is a moment selection vector, $|c|$ is the number of moment conditions selected by c. p is the number of parameters to be estimated, $J_n(c)$ is the J test statistic for over-identifying restrictions constructed using c with the optimal weighting matrix. Similar to the traditional BIC, these criteria penalize model complexity based on the number of parameters and the number of restrictions imposed. The model complexity increases when the number of parameters increases or the number of restrictions decreases. This idea was extended by Hong et al. (2003) to GEL estimation.

We have revisited our simulation exercise, originally presented in the main text (Table 3), and now report results based on other Frequentist methods: GMM-BIC, GMM-AIC, and GMM-HQIC. From the table, several points can be made. First, all methods exhibit model selection consistency, meaning the probability of selecting the true model approaches one as the number of observations increases. Second, our approach has stronger discriminatory power when ρ is close to zero compared to GMM-BIC. Third, GMM-AIC and GMM-HQIC select the right model more often when ρ is not zero (no endogeneity). However, they seem to over-select the model with endogeneity when there is no presence of endogeneity. In summary, under the data generating process considered in this example, our BETEL-based model selection performs better than other alternatives, especially in a finite sample.

Table 3: Table 1 with other Frequentist approaches. 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 $\mathcal{M}_b$. The numbers for BETEL are slightly different from those reported in the main text because they are based on different sets of simulated data, i.e., the random number seed is different.

BETEL	-0.5	-0.4	-0.3	-0.2	-0.1	0	0.1	0.2	0.3	0.4	0.5
250	99	96	82	48	12	2	18	54	93	100	100
500	100	100	98	76	17	1	29	87	99	100	100
1000	100	100	100	96	46	1	46	100	100	100	100
2000	100	100	100	100	80	1	70	100	100	100	100
GMM-BIC	-0.5	-0.4	-0.3	-0.2	-0.1	0	0.1	0.2	0.3	0.4	0.5
250	100	97	77	35	7	3	11	40	84	99	100
500	100	100	96	72	8	1	16	74	99	100	100
1000	100	100	100	92	29	1	25	99	100	100	100
2000	100	100	100	99	63	1	47	100	100	100	100
GMM-AIC	-0.5	-0.4	-0.3	-0.2	-0.1	0	0.1	0.2	0.3	0.4	0.5
250	100	100	96	74	28	15	37	79	98	100	100
500	100	100	100	94	46	11	60	98	100	100	100
1000	100	100	100	99	71	11	76	100	100	100	100
2000	100	100	100	100	95	12	94	100	100	100	100
GMM-HQIC	-0.5	-0.4	-0.3	-0.2	-0.1	0	0.1	0.2	0.3	0.4	0.5
250	100	99	85	54	17	5	20	62	94	99	100
500	100	100	100	82	23	3	31	89	100	100	100
1000	100	100	100	98	54	2	56	100	100	100	100
2000	100	100	100	100	79	1	73	100	100	100	100

B Proofs of the main results

The following notation will be used in the proofs. Further notation will be introduced in the Supplementary Appendix and will be used in the proofs there. 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 $\widetilde{w}_{1,i} := (x_i', z_{1,i}')'$. 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: $\mathbf{E}[\cdot] := \mathbf{E}^P[\cdot]$. We use standard notation in empirical process theory: $\mathbb{P}_n := \mathbf{E}_n[\delta_{w_i}]$ where δ_x is the Dirac measure at x, and $\mathbb{G}_n g := \sqrt{n}(\mathbb{P}_n f - \mathbf{E}f)$ for every function f.

For a function $\lambda(\theta)$ of θ, define $\tau_i^\dagger(\lambda,\theta) := \frac{e^{\lambda(\theta)' g_i(\theta)}}{\mathbf{E}[e^{\lambda(\theta)' g_j(\theta)}]}$, so that $\tau_i(\widehat{\lambda},\theta) = n\widehat{p}_i(\theta)$ and $\tau_i^\dagger(\lambda_*,\theta) = dQ^*(\theta)/dP$. We also use the notation: $\Omega_*^\diamond(\lambda,\theta) := \mathbf{E}[\tau_i^\diamond(\lambda,\theta)\varepsilon_i(\theta)\widetilde{w}_i\widetilde{w}_i']$, $\Omega_*^\diamond(\theta) := \Omega_*^\diamond(\lambda_*,\theta)$ and $\Omega_*^\dagger(\theta) := \mathbf{E}[\tau_i^\dagger(\lambda_*,\theta)\varepsilon_i(\theta)\widetilde{w}_i\widetilde{w}_i'] = \mathbf{E}^{Q^*(\theta)}[\varepsilon_i(\theta)\widetilde{w}_i\widetilde{w}_i']$. Moreover, $\Omega_* \equiv \Omega_*^\diamond(\theta_\circ)$.

B.1 Proof of Theorem 4.1

We first show that *(i)* is equivalent to *(ii)*. Suppose *(i)* is true. Then, $P \notin \mathcal{Q}_{b,\theta}$ for every $\theta \in \Theta$ and the I-projection of P on the set $\mathcal{Q}_{b,\theta}$ is different from P, $Q_b^*(\theta) \neq P$, for every $\theta \in \Theta$. It follows that also the reverse Kullback-Leibler divergence (where we have inverted the role played by the two probabilities) is strictly positive: $\mathrm{KL}(P||Q_b^*(\theta)) > 0$, for every $\theta \in \Theta$. Since this is true for every $\theta \in \Theta$, it is also true for θ_*. Hence *(ii)* holds.

Now, suppose that *(ii)* is true. Because $\mathrm{KL}(P||Q_b^*(\theta_*)) > 0$, then $P \neq Q_b^*(\theta_*)$ and $P \notin \mathcal{Q}_{b,\theta_*}$. Since θ_* minimizes $\mathrm{KL}(P||Q_b^*(\theta))$, then we also have that $P \notin \mathcal{Q}_{b,\theta}$ for every $\theta \in \Theta$. Hence *(i)* holds.

Next, we show that *(iii)* is equivalent to *(iv)*. Suppose *(iii)* holds. Then, there is a $\theta \in \Theta$, say θ_*, for which $P \in \mathcal{Q}_{b,\theta_*}$. Hence, $Q_b^*(\theta_*) = P$ and $\mathrm{KL}(P||Q_b^*(\theta)) = 0$. Hence *(iv)* holds.

Now, suppose that *(iv)* holds. By the properties of the Kullback-Leibler divergence, $\mathrm{KL}(P||Q_b^*(\theta_*))$ $= 0$ if and only if $P = Q_b^*(\theta)$. It follows that $P \in \mathcal{Q}_{b,\theta_*}$ because $Q_b^*(\theta_*) \in \mathcal{Q}_{b,\theta_*}$ and therefore P

satisfies the moment restriction $\mathbf{E}^P[\varepsilon_i(\theta_*)x_i] = 0$. Hence *(iii)* holds.

□

B.2 Proof of Theorem 4.2

Let us consider the expression for the likelihood evaluated at θ_*:

$$\begin{aligned}\log \widehat{q}(w_{1:n}|\theta_*, \mathcal{M}_b) &= -n\log n + \sum_{i=1}^n \widehat{\lambda}(\theta_*)' g_b(w_i,\theta_*) - n\log\frac{1}{n}\sum_{j=1}^n e^{\widehat{\lambda}(\theta_*)' g_b(w_j,\theta_*)} \\ &= -n\log n + \sum_{i=1}^n \widehat{\lambda}(\theta_*)' \widetilde{g}_b(w_i,\theta_*) - n\log\frac{1}{n}\sum_{j=1}^n e^{\widehat{\lambda}(\theta_*)' \widetilde{g}_b(w_j,\theta_*)}, \\ &= -n\log n + \sum_{i=1}^n \widehat{\lambda}(\theta_*)' \widetilde{g}_b(w_i,\theta_*) + n\widehat{\lambda}(\theta_*)'\mathbf{E}[g_b(w_i,\theta_*)] - n\log\frac{1}{n}\sum_{j=1}^n e^{\widehat{\lambda}(\theta_*)' g_b(w_j,\theta_*)}, \quad \text{(B.1)}\end{aligned}$$

where $\widetilde{g}_b(w_i,\theta_*) := g_b(w_i,\theta_*) - \mathbf{E}[g_b(w_i,\theta_*)]$. We first deal with the second term on the right hand side of (B.1). By using the result of Lemma C.2:

$$\begin{aligned}\sum_{i=1}^n \widehat{\lambda}(\theta_*)' \widetilde{g}_b(w_i,\theta_*) &= \sqrt{n}(\widehat{\lambda}(\theta_*) - \lambda_*(\theta_*))' \frac{1}{\sqrt{n}}\sum_{i=1}^n \widetilde{g}_b(w_i,\theta_*) + \lambda_*(\theta_*)' \sum_{i=1}^n \widetilde{g}_b(w_i,\theta_*) \\ &= -\mathbb{G}_n\left[\tau_i^\dagger(\lambda_*,\theta_*) g_b(w_i,\theta_*)'\right] \Omega_*^\dagger(\theta_*)^{-1} \mathbb{G}_n\left[g_b(w_i,\theta_*)\right] + \lambda_*(\theta_*)' \sum_{i=1}^n \widetilde{g}_b(w_i,\theta_*). \quad \text{(B.2)}\end{aligned}$$

Let $\tilde{\lambda}$ be on the line joining $\lambda_*(\theta_*)$ and $\widehat{\lambda}(\theta_*)$, then a second order Taylor expansion of $\lambda \mapsto \frac{1}{n}\sum_{j=1}^n e^{\lambda' g_b(w_j,\theta_*)}$ around $\lambda_*(\theta_*)$ gives

$$\begin{aligned}\frac{1}{n}\sum_{j=1}^n e^{\widehat{\lambda}(\theta_*)' g_b(w_j,\theta_*)} &= \frac{1}{n}\sum_{j=1}^n e^{\lambda_*(\theta_*)' g_b(w_j,\theta_*)} + (\widehat{\lambda}(\theta_*)' - \lambda_*(\theta_*)')\frac{1}{n}\sum_{i=1}^n e^{\lambda_*(\theta_*)' g_b(w_j,\theta_*)} g_b(w_i,\theta_*) \\ &\quad + \frac{1}{2}(\widehat{\lambda}(\theta_*)' - \lambda_*(\theta_*)')\frac{1}{n}\sum_{j=1}^n e^{\tilde{\lambda}' g_b(w_j,\theta_*)} g_b(w_j,\theta_*) g_b(w_j,\theta_*)'(\widehat{\lambda}(\theta_*) - \lambda_*(\theta_*)). \quad \text{(B.3)}\end{aligned}$$

Under Assumption 3 and because $\|\tilde{\lambda} - \lambda_*(\theta_*)\|_2 = \mathcal{O}_p(n^{-1/2})$ (since by Lemma C.1 $\|\widehat{\lambda}(\theta_*) - \lambda_*(\theta_*)\|_2 = \mathcal{O}_p(n^{-1/2})$ and $\tilde{\lambda} = \tau(\widehat{\lambda}(\theta_*) - \lambda_*(\theta_*)) + \lambda_*(\theta_*)$) we can apply the same argument of

the proof of Lemma C.5 to get:

$$\frac{1}{n}\sum_{j=1}^{n} e^{\tilde{\lambda}' g_b(w_j,\theta_*)} g_b(w_j,\theta_*) g_b(w_j,\theta_*)' \xrightarrow{p} \Omega_*^{\diamond}(\theta_*) := \mathbf{E}[e^{\lambda_*(\theta_*)' g_b(w_j,\theta_*)} \varepsilon_i(\theta_*)^2 \widetilde{w}_i \widetilde{w}_i']. \tag{B.4}$$

By replacing this in (B.3) and by using Lemma C.1 to get the rate of the $o_p(1/n)$ term, we obtain:

$$\begin{aligned}
\frac{1}{n}\sum_{j=1}^{n} e^{\widehat{\lambda}(\theta_*)' g_b(w_j,\theta_*)} &= \frac{1}{n}\sum_{j=1}^{n} e^{\lambda_*(\theta_*)' g_b(w_j,\theta_*)} + (\widehat{\lambda}(\theta_*)' - \lambda_*(\theta_*)')\frac{1}{n}\sum_{i=1}^{n} e^{\lambda_*(\theta_*)' g_b(w_j,\theta_*)} g_b(w_i,\theta_*) \\
&\quad + \frac{1}{2}(\widehat{\lambda}(\theta_*)' - \lambda_*(\theta_*)')\Omega_*^{\diamond}(\theta_*)(\widehat{\lambda}(\theta_*) - \lambda_*(\theta_*)) + o_p\left(\frac{1}{n}\right) \\
&= \mathbf{E}\left[e^{\lambda_*(\theta_*)' g_b(w_j,\theta_*)}\right] + (\widehat{\lambda}(\theta_*)' - \lambda_*(\theta_*)')\frac{1}{n}\sum_{i=1}^{n} e^{\lambda_*(\theta_*)' g_b(w_j,\theta_*)} g_b(w_i,\theta_*) \\
&\quad + \frac{1}{2}(\widehat{\lambda}(\theta_*)' - \lambda_*(\theta_*)')\Omega_*^{\diamond}(\theta_*)(\widehat{\lambda}(\theta_*) - \lambda_*(\theta_*)) + \mathcal{O}_p\left(\frac{1}{\sqrt{n}}\right),
\end{aligned} \tag{B.5}$$

where we have used the fact that $|\frac{1}{n}\sum_{j=1}^{n} e^{\lambda_*(\theta_*)' g_b(w_j,\theta_*)} - \mathbf{E}\left[e^{\lambda_*(\theta_*)' g_b(w_j,\theta_*)}\right]| = \mathcal{O}_p(1/\sqrt{n})$ by the Markov's inequality and under Assumption 5 *(b)*. We now use the first order Taylor expansion of the function $u \mapsto \log(u)$ around v: $\log(u) = \log(v) + \frac{u-v}{v} + o(|u-v|)$, and plug (B.5) in it to obtain:

$$\begin{aligned}
&\log\left(\frac{1}{n}\sum_{i=1}^{n} e^{\widehat{\lambda}(\theta_*)' g_b(w_i,\theta_*)}\right) = \log\left(\mathbf{E}\left[e^{\lambda_*(\theta_*)' g_b(w_i,\theta_*)}\right]\right) + (\widehat{\lambda}(\theta_*)' - \lambda_*(\theta_*)')\frac{1}{n}\sum_{i=1}^{n} \tau_i^{\dagger}(\lambda_*,\theta_*) g_b(w_i,\theta_*) \\
&+ \frac{1}{2}(\widehat{\lambda}(\theta_*)' - \lambda_*(\theta_*)')\Omega_*^{\dagger}(\theta_*)(\widehat{\lambda}(\theta_*) - \lambda_*(\theta_*)) + \mathcal{O}_p\left(\frac{1}{\sqrt{n}}\right) + o\left(\left|\frac{1}{n}\sum_{i=1}^{n} e^{\widehat{\lambda}(\theta_*)' g_b(w_i,\theta_*)} - \mathbf{E}\left[e^{\lambda_*(\theta_*)' g_b(w_i,\theta_*)}\right]\right|\right).
\end{aligned} \tag{B.6}$$

Since $\mathbf{E}[\tau_i^{\dagger}(\lambda_*,\theta_*) g_b(w_i,\theta_*)] = 0$ and by using Lemma C.2, then

$$\begin{aligned}
&n\left((\widehat{\lambda}(\theta_*)' - \lambda_*(\theta_*)')\frac{1}{n}\sum_{i=1}^{n} \tau_i^{\dagger}(\lambda_*,\theta_*) g_b(w_i,\theta_*) + \frac{1}{2}(\widehat{\lambda}(\theta_*)' - \lambda_*(\theta_*)')\Omega_*^{\dagger}(\theta_*)(\widehat{\lambda}(\theta_*) - \lambda_*(\theta_*))\right) \\
&= -\mathbb{G}_n\left[\tau_i^{\dagger}(\lambda_*,\theta_*) g_b(w_i,\theta_*)'\right] \Omega_*^{\dagger}(\theta_*)^{-1} \mathbb{G}_n\left[\tau_i^{\dagger}(\lambda_*,\theta_*) g_b(w_i,\theta_*)\right] \\
&+ \frac{1}{2}\mathbb{G}_n\left[\tau_i^{\dagger}(\lambda_*,\theta_*) g_b(w_i,\theta_*)'\right] \Omega_*^{\dagger}(\theta_*)^{-1} \mathbb{G}_n\left[\tau_i^{\dagger}(\lambda_*,\theta_*) g_b(w_i,\theta_*)\right] + o_p(1)
\end{aligned}$$

$$= -\frac{1}{2}\mathbb{G}_n\left[\tau_i^\dagger(\lambda_*,\theta_*)g_b(w_i,\theta_*)'\right]\Omega_*^\dagger(\theta_*)^{-1}\mathbb{G}_n\left[\tau_i^\dagger(\lambda_*,\theta_*)g_b(w_i,\theta_*)\right] + o_p(1) \quad \text{(B.7)}$$

Finally, we have to deal with $\left|\frac{1}{n}\sum_{i=1}^n e^{\widehat{\lambda}(\theta_*)'g_b(w_i,\theta_*)} - \mathbf{E}\left[e^{\lambda_*(\theta_*)'g_b(w_i,\theta_*)}\right]\right|$. By using (B.5), Lemma C.2 and the fact that

$$\frac{1}{\sqrt{n}}\sum_{i=1}^n e^{\lambda_*(\theta_*)'g_b(w_j,\theta_*)}g_b(w_i,\theta_*) = \mathbb{G}_n\left[\frac{1}{n}\sum_{i=1}^n e^{\lambda_*(\theta_*)'g_b(w_j,\theta_*)}g_b(w_i,\theta_*)\right]$$

we get that $\left|\frac{1}{n}\sum_{i=1}^n e^{\widehat{\lambda}(\theta_*)'g_b(w_i,\theta_*)} - \mathbf{E}\left[e^{\lambda_*(\theta_*)'g_b(w_i,\theta_*)}\right]\right| = \mathcal{O}_p(1/n)$.

By replacing this result, (B.2), (B.6), and (B.7) in (B.1) we get:

$$\begin{aligned}\log\widehat{q}(w_{1:n}|\theta_*,M_b) = &-n\log n - \mathbb{G}_n\left[\tau_i^\dagger(\lambda_*,\theta_*)g_b(w_i,\theta_*)'\right]\Omega_*^\dagger(\theta_*)^{-1}\mathbb{G}_n\left[g_b(w_i,\theta_*)\right]\\ &+ \lambda_*(\theta_*)'\sum_{i=1}^n\widetilde{g}_b(w_i,\theta_*) + n\widehat{\lambda}(\theta_*)'\mathbf{E}[g_b(w_i,\theta_*)] - n\log\left(\mathbf{E}\left[e^{\lambda_*(\theta_*)'g_b(w_i,\theta_*)}\right]\right)\\ &+ \frac{1}{2}\mathbb{G}_n\left[\tau_i^\dagger(\lambda_*,\theta_*)g_b(w_i,\theta_*)'\right]\Omega_*^\dagger(\theta_*)^{-1}\mathbb{G}_n\left[\tau_i^\dagger(\lambda_*,\theta_*)g_b(w_i,\theta_*)\right] + o_p(1), \quad \text{(B.8)}\end{aligned}$$

where $\mathbb{G}_n\left[\tau_i^\dagger(\lambda_*,\theta_*)g_b(w_i,\theta_*)\right] \xrightarrow{d} \mathcal{N}(0,\Omega_*^\dagger(\theta_*))$. By noticing that $\lambda_*(\theta_*)'\sum_{i=1}^n\widetilde{g}_b(w_i,\theta_*) - n\log\left(\mathbf{E}\left[e^{\lambda_*(\theta_*)'g_b(w_i,\theta_*)}\right]\right) = \sum_{i=1}^n\log\left(\tau_i^\dagger(\lambda_*,\theta_*)\right) - n\lambda_*(\theta_*)'\mathbf{E}[g_b(w_i,\theta_*)]$, we prove (4.3).

□

B.3 Proof of Theorem 4.3

Since we are in the extended model, then there exists a $\psi_\circ := (\theta_\circ', v_\circ')'$ such that $\mathbf{E}[\varepsilon_i(\theta_\circ)\widetilde{w}_i] = (v_\circ', 0')'$ and $\lambda_*(\psi_\circ) = 0$. Let us consider the expression for the likelihood evaluated at $\psi_\circ$:

$$\begin{aligned}\log\widehat{q}(w_{1:n}|\psi_\circ,\mathcal{M}_e) &= -n\log n + \sum_{i=1}^n\widehat{\lambda}(\psi_\circ)'g_b(w_i,\theta_\circ) - n\log\frac{1}{n}\sum_{j=1}^n e^{\widehat{\lambda}(\psi_\circ)'g_b(w_j,\theta_\circ)}\\ &= -n\log n + \sum_{i=1}^n\widehat{\lambda}(\psi_\circ)'g_e(w_i,\psi_\circ) - n\log\frac{1}{n}\sum_{j=1}^n e^{\widehat{\lambda}(\psi_\circ)'g_e(w_j,\psi_\circ)}. \quad \text{(B.9)}\end{aligned}$$

We start with dealing with the second term on the right hand side of (B.9). By using the result of Lemma C.14:

$$\sum_{i=1}^{n} \widehat{\lambda}(\psi_\circ)' g_e(w_i, \psi_\circ) = \sqrt{n}\widehat{\lambda}(\psi_\circ)' \frac{1}{\sqrt{n}} \sum_{i=1}^{n} g_e(w_i, \psi_\circ)$$
$$= -\mathbb{G}_n \left[g_e(w_i, \psi_\circ)' \right] \Omega_{\psi_\circ}^{-1} \mathbb{G}_n \left[g_e(w_i, \psi_\circ) \right] + o_p(1). \quad \text{(B.10)}$$

Let $\tilde{\lambda}$ be on the line joining 0 and $\widehat{\lambda}(\psi_\circ)$, then a second order Taylor expansion of the function $\lambda \mapsto \frac{1}{n}\sum_{j=1}^{n} e^{\lambda' g_e(w_j, \theta_\circ)}$ around 0 gives

$$\frac{1}{n}\sum_{j=1}^{n} e^{\widehat{\lambda}(\psi_\circ)' g_e(w_j, \psi_\circ)} = 1 + \frac{1}{n}\sum_{i=1}^{n} \widehat{\lambda}(\psi_\circ)' g_e(w_i, \psi_\circ)$$
$$+ \frac{1}{2}\widehat{\lambda}(\psi_\circ)' \frac{1}{n}\sum_{j=1}^{n} e^{\tilde{\lambda}' g_e(w_j, \psi_\circ)} g_e(w_j, \psi_\circ) g_e(w_j, \psi_\circ)' \widehat{\lambda}(\psi_\circ). \quad \text{(B.11)}$$

Under Assumption 3 and because $\tilde{\lambda} = \mathcal{O}_p(n^{-1/2})$ (since by Lemma C.15 $\widehat{\lambda}(\psi_\circ) = \mathcal{O}_p(n^{-1/2})$ and $\tilde{\lambda}$ is between 0 and $\widehat{\lambda}(\psi_\circ)$) we can apply the same argument of the proof of Lemma C.5 to get:

$$\frac{1}{n}\sum_{j=1}^{n} e^{\tilde{\lambda}' g_e(w_j, \theta_\circ)} g_e(w_j, \theta_\circ) g_e(w_j, \theta_\circ)' \xrightarrow{p} \Omega_{\psi_\circ} := \mathbf{E}[g_e(w_i, \psi_\circ) g_e(w_i, \psi_\circ)']. \quad \text{(B.12)}$$

By replacing this in (B.11) and by using Lemma C.15 to get the rate of the $o_p(1/n)$ term, we obtain:

$$\frac{1}{n}\sum_{j=1}^{n} e^{\widehat{\lambda}(\psi_\circ)' g_e(w_j, \psi_\circ)} = 1 + \frac{1}{n}\sum_{i=1}^{n} \widehat{\lambda}(\psi_\circ)' g_e(w_i, \psi_\circ) + \frac{1}{2}\widehat{\lambda}(\psi_\circ)' \Omega_{\psi_\circ} \widehat{\lambda}(\psi_\circ) + o_p(1/n). \quad \text{(B.13)}$$

We now use the first order Taylor expansion of the function $\log(u)$ around $u = 1$: $\log(u) = u - 1 + o(|u-1|)$, and plug (B.13) in it to obtain:

$$\log\left(\frac{1}{n}\sum_{i=1}^{n} e^{\widehat{\lambda}(\psi_\circ)' g_e(w_i, \psi_\circ)}\right) = \frac{1}{n}\sum_{i=1}^{n} \widehat{\lambda}(\psi_\circ)' g_e(w_i, \psi_\circ)$$
$$+ \frac{1}{2}\widehat{\lambda}(\psi_\circ)' \Omega_{\psi_\circ} \widehat{\lambda}(\psi_\circ) + o_p\left(1/n\right) + o\left(\left|\frac{1}{n}\sum_{i=1}^{n} e^{\widehat{\lambda}(\psi_\circ)' g_e(w_i, \psi_\circ)} - 1\right|\right). \quad \text{(B.14)}$$

By using the result of Lemma C.14, then

$$n\left(\frac{1}{n}\sum_{i=1}^{n}\widehat{\lambda}(\psi_\circ)'g_e(w_i,\psi_\circ)+\frac{1}{2}\widehat{\lambda}(\psi_\circ)'\Omega_{\psi_\circ}\widehat{\lambda}(\psi_\circ)\right)=\sqrt{n}\widehat{\lambda}(\psi_\circ)'\mathbb{G}_n\left[g_e(w_i,\psi_\circ)\right]+\frac{1}{2}\sqrt{n}\widehat{\lambda}(\psi_\circ)'\Omega_{\psi_\circ}\sqrt{n}\widehat{\lambda}(\psi_\circ)$$
$$=-\mathbb{G}_n\left[g_e(w_i,\psi_\circ)'\right]\Omega_{\psi_\circ}^{-1}\mathbb{G}_n\left[g_e(w_i,\psi_\circ)\right]+\frac{1}{2}\mathbb{G}_n\left[g_e(w_i,\psi_\circ)'\right]\Omega_{\psi_\circ}^{-1}\mathbb{G}_n\left[g_e(w_i,\psi_\circ)\right]+o_p(1)$$
$$=-\frac{1}{2}\mathbb{G}_n\left[g_e(w_i,\psi_\circ)'\right]\Omega_{*}^{-1}\mathbb{G}_n\left[g_e(w_i,\psi_\circ)\right]+o_p(1). \quad \text{(B.15)}$$

Finally, we have to deal with $\left|\frac{1}{n}\sum_{i=1}^{n}e^{\widehat{\lambda}(\psi_\circ)'g_e(w_i,\psi_\circ)}-1\right|$. By using exactly the same argument we have used in the proof of Lemma C.7 with $\lambda_*(\theta_*)$ replaced by $\lambda_*(\psi_\circ)=0$ we get that $\left|\frac{1}{n}\sum_{i=1}^{n}e^{\widehat{\lambda}(\psi_\circ)'g_e(w_i,\psi_\circ)}-1\right|=o_p(1)$.

By replacing this result and (B.15) in (B.14), and then by plugging (B.10) and (B.14) in (B.9) we get:

$$\log\widehat{q}(w_{1:n}|\psi_\circ,M_b)=-n\log n-\frac{1}{2}\mathbb{G}_n\left[g_e(w_i,\psi_\circ)'\right]\Omega_{\psi_\circ}^{-1}\mathbb{G}_n\left[g_e(w_i,\psi_\circ)\right]+o_p(1)$$
$$=-n\log n-\frac{1}{2}\mathbb{G}_n\left[g_b(w_i,\theta_\circ)'\right]\Omega_{\psi_\circ}^{-1}\mathbb{G}_n\left[g_b(w_i,\theta_\circ)\right]-2\sqrt{n}\mathbb{G}_n\left[g_b(w_i,\theta_\circ)'\right]\Omega_{\psi_\circ}^{-1}\widetilde{v}_\circ+n\widetilde{v}_\circ'\Omega_{\psi_\circ}^{-1}\widetilde{v}_\circ+o_p(1). \quad \text{(B.16)}$$

Moreover, by the central limit theorem,

$$\mathbb{G}_n\left[g_e(w_i,\psi_\circ)\right]\xrightarrow{d}\mathcal{N}(0,\Omega_{\psi_\circ})$$

and

$$\mathbb{G}_n\left[g_e(w_i,\psi_\circ)'\right]\Omega_{*}^{-1}\mathbb{G}_n\left[g_e(w_i,\psi_\circ)\right]\xrightarrow{d}\chi^2_d.$$

□

B.4 Proof of Corollary 4.1

By result (4.3) in Theorem 4.2 we have that

$$\frac{1}{n}\log \widehat{q}(w_{1:n}|\theta_*, \mathcal{M}_b) + \log(n) = \frac{1}{n}\sum_{i=1}^{n}\log\left(\frac{e^{\lambda_*(\theta_*)'g_b(w_i,\theta_*)}}{\mathbf{E}[e^{\lambda_*(\theta_*)'g_b(w_j,\theta_*)}]}\right) + (\widehat{\lambda}(\theta_*) - \lambda_*(\theta_*))'\mathbf{E}[g_b(w_i,\theta_*)] + \mathcal{O}_p(1/n). \quad \text{(B.17)}$$

By Lemma C.1 in the Online Appendix, $\|\widehat{\lambda}(\theta_*) - \lambda_*(\theta_*)\|_2 \xrightarrow{p} 0$. By the Law of Large Numbers

$$\frac{1}{n}\sum_{i=1}^{n}\log\left(\frac{e^{\lambda_*(\theta_*)'g_b(w_i,\theta_*)}}{\mathbf{E}[e^{\lambda_*(\theta_*)'g_b(w_j,\theta_*)}]}\right) \xrightarrow{p} \mathbf{E}\left[\log\left(\frac{e^{\lambda_*(\theta_*)'g_b(w_i,\theta_*)}}{\mathbf{E}[e^{\lambda_*(\theta_*)'g_b(w_j,\theta_*)}]}\right)\right].$$

This concludes the proof.

□

B.5 Proof of Corollary 4.2

By result (4.4) in Theorem 4.3 we have that

$$\frac{1}{n}\log \widehat{q}(w_{1:n}|\psi_\circ, \mathcal{M}_e) + \log(n) = \mathcal{O}_p(1/n).$$

Since $\lambda_*(\psi_\circ) = 0$ then, $\frac{e^{\lambda_*(\psi_\circ)'\sum_{i=1}^{n} g_b(w_i,\psi_\circ)}}{\mathbf{E}[e^{\lambda_*(\psi_\circ)'\sum_{i=1}^{n} g_b(w_j,\psi_\circ)}]} = 1$ and so we can equivalently write:

$$\frac{1}{n}\log \widehat{q}(w_{1:n}|\psi_\circ, \mathcal{M}_e) + \log(n) = \frac{1}{n}\sum_{i=1}^{n}\log\left(\frac{e^{\lambda_*(\psi_\circ)'g_b(w_i,\psi_\circ)}}{\mathbf{E}[e^{\lambda_*(\psi_\circ)'g_b(w_j,\psi_\circ)}]}\right) + \mathcal{O}_p(1/n).$$

By the Law of Large Numbers

$$\frac{1}{n}\sum_{i=1}^{n}\log\left(\frac{e^{\lambda_*(\psi_\circ)'g_b(w_i,\psi_\circ)}}{\mathbf{E}[e^{\lambda_*(\psi_\circ)'g_b(w_j,\psi_\circ)}]}\right) \xrightarrow{p} \mathbf{E}\left[\log\left(\frac{e^{\lambda_*(\psi_\circ)'g_b(w_i,\psi_\circ)}}{\mathbf{E}[e^{\lambda_*(\psi_\circ)'g_b(w_j,\psi_\circ)}]}\right)\right].$$

This concludes the proof.

□

B.6 Proof of Theorem 4.4

The proof is organised in two parts. In the first part we show that $\mathrm{KL}(P||Q_e^*(\psi_\circ)) < \mathrm{KL}(P||Q_b^*(\theta_*))$ if and only if there is no θ such that $\mathbf{E}[\varepsilon_i(\theta)x_i] = 0$. In the second part we show that

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

if and only if $\mathrm{KL}(P||Q_e^*(\psi_\circ)) < \mathrm{KL}(P||Q_b^*(\theta_*))$.

First part. We start by proving that $\mathrm{KL}(P||Q_e^*(\psi_\circ)) < \mathrm{KL}(P||Q_b^*(\theta_*))$ if and only if there is no θ such that $\mathbf{E}[\varepsilon_i(\theta)x_i] = 0$. Notice that $\mathrm{KL}(P||Q_e^*(\psi_\circ)) = 0$. Suppose that $\mathrm{KL}(P||Q_e^*(\psi_\circ)) < \mathrm{KL}(P||Q_b^*(\theta_*))$ and suppose that there exists a θ such that $\mathbf{E}[\varepsilon_i(\theta)x_i] = 0$ so that $P \in \mathcal{Q}_{b,\theta}$. By Assumption 2 with θ_* replaced by $\theta_\circ$ then this θ must be equal to $\theta_\circ$ which in turn equals θ_*. It follows that $P \in \mathcal{Q}_{b,\theta_*}$ and by definition of $Q_b^*(\theta_*)$: $Q_b^*(\theta_*) = P$ since $Q_b^*(\theta_*)$ is the closest to P, in the KL sense, among all the distributions in $\mathcal{Q}_{b,\theta_*}$. Hence, $\mathrm{KL}(P||Q_b^*(\theta_*)) = 0$. But this contradicts the assumption that $\mathrm{KL}(P||Q_e^*(\psi_\circ)) < \mathrm{KL}(P||Q_b^*(\theta_*))$. 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, $P \notin \mathcal{Q}_{b,\theta}$ for every $\theta \in \Theta$ which implies $P \notin \mathcal{Q}_{b,\theta_*}$ and $\mathrm{KL}(P||Q_b^*(\theta_*)) > 0$. 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 $\mathrm{KL}(P||Q_e^*(\psi_\circ)) < \mathrm{KL}(P||Q_b^*(\theta_*))$.

Second part. We show that $P\left(\log m(w_{1:n}|\mathcal{M}_e) > \log m(w_{1:n}|\mathcal{M}_b)\right) \to 1$ if and only if $\mathrm{KL}(P||Q_e^*(\psi_\circ)) < \mathrm{KL}(P||Q_b^*(\theta_*))$. By Theorems 4.2 and 4.3, and Theorems C.6 and C.7 in the Online Appendix

and by (4.1)-(4.2), then (4.9)-(4.10) hold. By (4.9), the $\log m(w_{1:n}|\mathcal{M}_b)$ is equal to

$$-n\log n + \sum_{i=1}^{n} \log\left(\frac{e^{\lambda_*(\theta_*)'g_b(w_i,\theta_*)}}{\mathbf{E}[e^{\lambda_*(\theta_*)'g_b(w_j,\theta_*)}]}\right) + n(\widehat{\lambda}(\theta_*) - \lambda_*(\theta_*))'\mathbf{E}[g_b(w_i,\theta_*)] - \frac{p}{2}\log(n) + \mathcal{O}_p(1)$$

and by (4.10), $\log m(w_{1:n}|\mathcal{M}_e) = -n\log(n) - \frac{p+d_x}{2}\log(n) + \mathcal{O}_p(1)$. Hence, since from the Law of Large Numbers:

$$\frac{1}{n}\sum_{i=1}^{n} \log\left(\frac{e^{\lambda_*(\theta_*)'\sum_{i=1}^{n} g_b(w_i,\theta_*)}}{\mathbf{E}[e^{\lambda_*(\theta_*)'\sum_{i=1}^{n} g_b(w_j,\theta_*)}]}\right) \xrightarrow{p} \mathbf{E}\left[\log\left(\frac{e^{\lambda_*(\theta_*)'\sum_{i=1}^{n} g_b(w_i,\theta_*)}}{\mathbf{E}[e^{\lambda_*(\theta_*)'\sum_{i=1}^{n} g_b(w_j,\theta_*)}]}\right)\right] = -\mathrm{KL}(P||Q_b^*(\theta_*)),$$

it follows that

$$\begin{aligned} P\left(\log m(w_{1:n}|\mathcal{M}_e) > \log m(w_{1:n}|\mathcal{M}_b)\right) &= P\left(\frac{1}{n}\log m(w_{1:n}|\mathcal{M}_e) > \frac{1}{n}\log m(w_{1:n}|\mathcal{M}_b)\right) \\ &= P\left(0 > -\mathrm{KL}(P||Q_b^*(\theta_*)) + \mathcal{O}_p(1/\sqrt{n})\right), \end{aligned}$$

where we have used Lemma C.1 in the Online Appendix to control $\sqrt{n}(\widehat{\lambda}(\theta_*) - \lambda_*(\theta_*))$. Suppose that $\mathrm{KL}(P||Q_b^*(\theta_*)) > 0$ then the previous probability converges to 1. On the other hand, suppose that $P\left(0 > -\mathrm{KL}(P||Q_b^*(\theta_*)) + \mathcal{O}_p(1/\sqrt{n})\right) \to 1$ as $n \to \infty$. This is possible only if $\mathrm{KL}(P||Q_b^*(\theta_*)) > 0$. By the first part of the proof $\mathrm{KL}(P||Q_b^*(\theta_*)) > 0$ if and only if there is no θ such that $\mathbf{E}[\varepsilon_i(\theta)x_i] = 0$.

We now prove the last assertion of the theorem. In the case where there is a $\theta_\circ$ such that $\mathbf{E}[\varepsilon_i(\theta_\circ)x_i] = 0$, then $\mathrm{KL}(P||Q_b^*(\theta_\circ)) = 0$ and the probability $P\left(0 > -\mathrm{KL}(P||Q_b^*(\theta_*)) + \mathcal{O}_p(1/\sqrt{n})\right)$ is equal to zero as $n \to \infty$. This concludes the proof.

□

B.7 Proof of Theorem 4.5

We start by supposing that $\mathbf{E}[\varepsilon_i(\theta_\circ)x_i] = 0$. In this case, $\theta_* = \theta_\circ$, $\lambda_*(\theta_*) = \lambda_*(\theta_\circ) = 0$ and by Theorems 4.2 and 4.3:

$$\begin{aligned}\log \widehat{q}(w_{1:n}|\theta_\circ, \mathcal{M}_b) - \log \widehat{q}(w_{1:n}|\psi_\circ, \mathcal{M}_e) \\ = -n\log n - \frac{1}{2}\mathbb{G}_n\left[g_b(w_i,\theta_\circ)'\right]\Omega_\circ^{-1}\mathbb{G}_n\left[g_b(w_i,\theta_\circ)\right] + n\log n \\ + \frac{1}{2}\mathbb{G}_n\left[g_b(w_i,\theta_\circ)'\right]\Omega_\circ^{-1}\mathbb{G}_n\left[g_b(w_i,\theta_\circ)\right] + o_p(1) = o_p(1). \quad \text{(B.18)}\end{aligned}$$

Let $\pi^n_{h_\theta}(\cdot|w_{1:n}, \mathcal{M}_b)$ and $\pi^n_{h_\psi}(\cdot|w_{1:n}, \mathcal{M}_e)$ denote the posterior density of h_θ and h_ψ, respectively. By Corollary C.1 in the Online Appendix (which is valid if $\mathbf{E}[\varepsilon_i(\theta_\circ)x_i] = 0$ holds)

$$\begin{aligned}\log \pi^n_{h_\theta}(\sqrt{n}(\theta-\theta_\circ)|w_{1:n}, \mathcal{M}_b)\Big|_{\theta=\theta_\circ} = -\frac{p}{2}\log(2\pi) + \frac{1}{2}\log|V_{\theta_\circ}^{-1}| \\ - \frac{1}{2}\mathbb{G}_n\left[\varepsilon_i(\theta_\circ)\widetilde{w}_i'\right]\Omega_\circ^{-1}\mathbf{E}[\widetilde{w}_i\widetilde{w}_{1,i}']V_{\theta_\circ}\mathbf{E}[\widetilde{w}_{1,i}\widetilde{w}_i']\Omega_\circ^{-1}\mathbb{G}_n\left[\varepsilon_i(\theta_\circ)\widetilde{w}_i\right] + o_p(1) \quad \text{(B.19)}\end{aligned}$$

and by Theorem C.7 in the Online Appendix

$$\begin{aligned}\log \pi^n_{h_\psi}(\sqrt{n}(\psi-\psi_\circ)|w_{1:n}, \mathcal{M}_e)\Big|_{\psi=\psi_\circ} = -\frac{(p+d_x)}{2}\log(2\pi) + \frac{1}{2}\log|V_{\psi_\circ}^{-1}| \\ - \frac{1}{2}\mathbb{G}_n\left[\varepsilon_i(\theta_\circ)\widetilde{w}_i'\right]\Omega_\circ^{-1}\left[\frac{dg_e(w_i,\psi_\circ)'}{d\psi}\right]V_{\psi_\circ}\left[\frac{dg_e(w_i,\psi_\circ)}{d\psi'}\right]\Omega_\circ^{-1}\mathbb{G}_n\left[\varepsilon_i(\theta_\circ)\widetilde{w}_i\right] + o_p(1), \quad \text{(B.20)}\end{aligned}$$

where $V_{\theta_\circ}$ and $V_{\psi_\circ}$ are defined in Corollary C.1 and Theorem C.7 in the Online Appendix. Hence, by replacing (B.20), (B.19) and (B.18) in $\log m(w_{1:n}|\mathcal{M}_b) - \log m(w_{1:n}|\mathcal{M}_e)$ by using the expressions for the log-marginal likelihoods given in (4.1)-(4.2) with θ_* replaced by $\theta_\circ$, we obtain:

$$\begin{aligned}P\left(\log m(w_{1:n}|\mathcal{M}_b) > \log m(w_{1:n}|\mathcal{M}_e)\right) = P\Big(\log\pi(\theta_\circ|\mathcal{M}_b) + \frac{p}{2}\log(2\pi) - \frac{1}{2}\log|V_{\theta_\circ}^{-1}| \\ + \frac{1}{2}\mathbb{G}_n\left[\varepsilon_i(\theta_\circ)\widetilde{w}_i'\right]\Omega_\circ^{-1}\mathbf{E}[\widetilde{w}_i\widetilde{w}_{1,i}']V_{\theta_\circ}\mathbf{E}[\widetilde{w}_{1,i}\widetilde{w}_i']\Omega_\circ^{-1}\mathbb{G}_n\left[\varepsilon_i(\theta_\circ)\widetilde{w}_i\right] - \frac{p}{2}\log(n) \\ > \log\pi(\psi_\circ|\mathcal{M}_e) + \frac{(p+d_x)}{2}\log(2\pi) - \frac{1}{2}\log|V_{\psi_\circ}^{-1}|\end{aligned}$$

$$+\frac{1}{2}\mathbb{G}_n\left[\varepsilon_i(\theta_\circ)\widetilde{w}_i'\right]\Omega_\circ^{-1}\left[\frac{dg(w_i,\psi_\circ)'}{d\psi}\right]V_{\psi_\circ}\left[\frac{dg(w_i,\psi_\circ)}{d\psi'}\right]\Omega_\circ^{-1}\mathbb{G}_n\left[\varepsilon_i(\theta_\circ)\widetilde{w}_i\right]-\frac{p+d_x}{2}\log(n)+o_p(1)\Big). \tag{B.21}$$

Because $\mathbb{G}_n\left[\varepsilon_i(\theta_\circ)\widetilde{w}_i\right]=\mathcal{O}_p(1)$, $|V_{\theta_\circ}^{-1}|=\mathcal{O}(1)$ and $|V_{\psi_\circ}^{-1}|=\mathcal{O}(1)$ (since $V_{\theta_\circ}$ and $V_{\psi_\circ}$ are positive definite under Assumption 4), then we can factorize $\log(n)$ in (B.21) and get:

$$\begin{aligned}
&P\left(\log m(w_{1:n}|\mathcal{M}_b)>\log m(w_{1:n}|\mathcal{M}_e)\right)=\\
&\quad=P\Big(0>\log(n)\Big[\frac{1}{\log(n)}\log\frac{\pi(\psi_\circ|\mathcal{M}_e)}{\pi(\theta_\circ|\mathcal{M}_b)}+\frac{d_x\log(2\pi)}{2\log(n)}-\frac{1}{2\log(n)}\log\frac{|V_{\theta_\circ}|}{|V_{\psi_\circ}|}-\frac{d_x}{2}\\
&+\frac{1}{2\log(n)}\mathbb{G}_n\left[\varepsilon_i(\theta_\circ)\widetilde{w}_i'\right]\Omega_\circ^{-1}\left(\left[\frac{dg(w_i,\psi_\circ)'}{d\psi}\right]V_{\psi_\circ}\left[\frac{dg(w_i,\psi_\circ)}{d\psi'}\right]-\mathbf{E}[\widetilde{w}_i\widetilde{w}_{1,i}']V_{\theta_\circ}\mathbf{E}[\widetilde{w}_{1,i}\widetilde{w}_i']\right)\\
&\quad\times\Omega_\circ^{-1}\mathbb{G}_n\left[\varepsilon_i(\theta_\circ)\widetilde{w}_i\right]\Big]+o_p(1)\Big)=P\left(0>\log(n)\left[o_p(1)-\frac{d_x}{2}\right]+o_p(1)\right)\to 1
\end{aligned} \tag{B.22}$$

as $n\to\infty$. This proves the first implication.

We now prove the reverse implication. Suppose that $P\left(\log m(w_{1:n}|\mathcal{M}_b)>\log m(w_{1:n}|\mathcal{M}_e)\right)\to 1$. By (4.1)-(4.2):

$$\begin{aligned}
P\left(\log m(w_{1:n}|\mathcal{M}_b)>\log m(w_{1:n}|\mathcal{M}_e)\right)=P\Big(&\log\pi(\theta_*|\mathcal{M}_b)+\log\widehat{q}(w_{1:n}|\theta_*,\mathcal{M}_b)\\
&-\log\pi_{h_\theta}^n(0|w_{1:n},\mathcal{M}_b)-\frac{p}{2}\log(n)>\log\pi(\psi_\circ|\mathcal{M}_e)+\log\widehat{q}(w_{1:n}|\psi_\circ,\mathcal{M}_e)\\
&-\log\pi_{h_\psi}^n(0|w_{1:n},\mathcal{M}_e)-\frac{p+d_x}{2}\log(n)\Big)
\end{aligned} \tag{B.23}$$

By using Theorems 4.2 and 4.3, we get:

$$\begin{aligned}
\log\widehat{q}(w_{1:n}|\theta_*,\mathcal{M}_b)-\log\widehat{q}(w_{1:n}|\psi_\circ,\mathcal{M}_e)=&-\mathcal{A}_n'\Omega_*^\dagger(\theta_*)^{-1}\mathcal{B}_n\\
&+\sum_{i=1}^n\log\left(\frac{e^{\lambda_*(\theta_*)'g_b(w_i,\theta_*)}}{\mathbf{E}[e^{\lambda_*(\theta_*)'g_b(w_j,\theta_*)}]}\right)+n(\widehat{\lambda}(\theta_*)-\lambda_*(\theta_*))'\mathbf{E}[g_b(w_i,\theta_*)]\\
&+\frac{1}{2}\mathbb{G}_n\left[\tau_i^\dagger(\lambda_*,\theta_*)g_b(w_i,\theta_*)'\right]\Omega_*^\dagger(\theta_*)^{-1}\mathbb{G}_n\left[\tau_i^\dagger(\lambda_*,\theta_*)g_b(w_i,\theta_*)\right]\\
&+\frac{1}{2}\mathbb{G}_n\left[g_e(w_i,\psi_\circ)'\right]\Omega_{\psi_\circ}^{-1}\mathbb{G}_n\left[g_e(w_i,\psi_\circ)\right]+o_p(1),
\end{aligned} \tag{B.24}$$

where $\mathcal{A}_n := \mathbb{G}_n\left[\tau_i^\dagger(\lambda_*,\theta_*)g_b(w_i,\theta_*)'\right] \xrightarrow{d} \mathcal{N}(0,\Omega_*^\dagger(\theta_*))$, $\mathcal{B}_n := \mathbb{G}_n\left[g_b(w_i,\theta_*)'\right] \xrightarrow{d} \mathcal{N}(0,\mathbf{E}[\varepsilon_i(\theta_\circ)]\widetilde{w}_i\widetilde{w}_i')$ and $\mathbb{G}_n\left[g_e(w_i,\psi_\circ)'\right]\Omega_{\psi_\circ}^{-1}\mathbb{G}_n\left[g_e(w_i,\psi_\circ)\right] \xrightarrow{d} \chi^2_d$ and so they are bounded in probability. Moreover, by the Law of Large Numbers:

$$\left|\frac{1}{n}\sum_{i=1}^n \log\left(\frac{e^{\lambda_*(\theta_*)'g_b(w_i,\theta_*)}}{\mathbf{E}[e^{\lambda_*(\theta_*)'g_b(w_j,\theta_*)}]}\right) - \mathbf{E}^P\left[\log(dQ_b^*(\theta_*)/dP)\right]\right| = \mathcal{O}_p(1/\sqrt{n}),$$

where $\mathbf{E}^P\left[\log(dQ_b^*(\theta_*)/dP)\right] = -\mathrm{KL}(P||Q_b^*(\theta_*))$, and by Lemma C.2 in the Online Appendix,

$$\sqrt{n}(\widehat{\lambda}(\theta_*) - \lambda(\theta_*))'\mathbf{E}[g_b(w_i,\theta_*)] = -\mathbb{G}_n[\tau_i^\dagger(\lambda_*,\theta_*)\varepsilon_i(\theta_*)\widetilde{w}_i']\Omega_*^\dagger(\theta_*)^{-1}\mathbf{E}[g_b(w_i,\theta_*)] + o_p(1).$$

Therefore,

$$\begin{aligned}&\log\widehat{q}(w_{1:n}|\theta_*,\mathcal{M}_b) - \log\widehat{q}(w_{1:n}|\psi_\circ,\mathcal{M}_e)\\ &\qquad = \mathcal{O}_p(1) + n\left(\mathcal{O}_p(1/\sqrt{n}) - \mathrm{KL}(P||Q_b^*(\theta_*))\right) + \sqrt{n}\mathcal{O}_p(1) \quad \text{(B.25)}\end{aligned}$$

By replacing (B.25) in (B.23), and by using Theorems C.6 and C.7 in the Online Appendix to show that $\log\pi_{h_\theta}^n(0|w_{1:n},\mathcal{M}_b) = \mathcal{O}_p(1)$ and $\log\pi_{h_\psi}^n(0|w_{1:n},\mathcal{M}_e) = \mathcal{O}_p(1)$, the expression in (B.23) is equal to:

$$\begin{aligned}&P\left(\log m(w_{1:n}|\mathcal{M}_b) > \log m(w_{1:n}|\mathcal{M}_e)\right)\\ &\qquad = P\left(\mathcal{O}_p(1)\sqrt{n} - n\mathrm{KL}(P||Q_b^*(\theta_*)) > \mathcal{O}_p(1) - \frac{d_x}{2}\log(n)\right) \quad \text{(B.26)}\end{aligned}$$

where $\mathcal{O}_p(1)\sqrt{n} - n\mathrm{KL}(P||Q_b^*(\theta_*))$ in the left hand side converges to $-\infty$ if $\mathrm{KL}(P||Q_b^*(\theta_*)) > 0$ (since the term in n is diverging faster than the term in $\sqrt{n}$) while the term on the right hand side also converges towards $-\infty$. The inequality is then satisfied with probability approaching 1 only if $\mathrm{KL}(P||Q_b^*(\theta_*)) = 0$. This is equivalent to have $\mathbf{E}[\varepsilon_i(\theta_\circ)x_i] = 0$ (by the first part of the proof of Theorem 4.4) and we have proved the second part of the 'if and only if' statement.

□