Bayesian Estimation and Comparison of Moment Condition Models*

Siddhartha Chib† Minchul Shin‡ Anna Simoni§

Washington University in St. Louis *University of Illinois* *CREST, CNRS*

This version: July 16, 2017
First version: June, 2016

Abstract

In this paper we develop a Bayesian semiparametric analysis of moment condition models by casting the problem within the Exponentially Tilted Empirical Likelihood (ETEL) framework. We use this framework to develop a fully Bayesian analysis of correctly and misspecified moment condition models. We show that even under misspecification, the Bayesian ETEL posterior distribution satisfies the Bernstein - von Mises (BvM) theorem. We also develop a unified approach based on marginal likelihoods and Bayes factors for comparing different moment restricted models and for discarding any misspecified moment restrictions. Computation of the marginal likelihoods is by the method of Chib (1995) as extended to Metropolis-Hastings samplers in Chib and Jeliazkov (2001). We establish the model selection consistency of the marginal likelihood and show that the marginal likelihood favors the model with the minimum number of parameters and the maximum number of valid moment restrictions. When the models are misspecified, the marginal likelihood model selection procedure selects the model that is closer to the (unknown) true data generating process in terms of the Kullback-Leibler divergence. The ideas and results in this paper broaden the theoretical underpinning and value of the Bayesian ETEL framework with many practical applications. The discussion is illuminated through several examples.

Key words: Bernstein-von Mises theorem; Estimating Equations; Exponentially Tilted Empirical Likelihood; Marginal Likelihood; Misspecification; Model selection consistency.

*The authors gratefully thank the Co-Editor, an Associate Editor, and four anonymous referees for their many constructive comments on the previous version of the paper. Anna Simoni gratefully acknowledges financial support from ANR-13-BSH1-0004, and ANR-11-LABEX-0047.

†Olin Business School, Washington University in St. Louis, Campus Box 1133, 1 Brookings Dr. St. Louis, MO 63130, USA, e-mail: chib@wustl.edu

‡Department of Economics, University of Illinois, 214 David Kinley Hall, 1407 W. Gregory, Urbana, IL 61801, e-mail: mincshin@illinois.edu

§CREST - ENSAE - École Polytechnique, 5, avenue Henry Le Chatelier, 91120 Palaiseau , France, e-mail: simoni.anna@gmail.com

1 Introduction

Our goal in this paper is to develop a Bayesian analysis of moment condition models. By moment condition models, we mean models that are specified only through moment restrictions of the type $\mathbf{E}^P[\boldsymbol{g}(\boldsymbol{X}, \boldsymbol{\theta})] = \mathbf{0}$, where $\boldsymbol{g}(\boldsymbol{X}, \boldsymbol{\theta})$ is a known vector-valued function of a random vector $\boldsymbol{X}$ and an unknown parameter vector $\boldsymbol{\theta}$, and P is the unknown data distribution. Models of this type, which arise frequently in statistics and econometrics, see *e.g.* Broniatowski and Keziou (2012), can be attractive since full modeling of P is not invoked and inferences about $\boldsymbol{\theta}$ are based only on the partial information supplied by the set of moment conditions. For instance, in a regression context, letting $\boldsymbol{X} = (y, x)$ and $y = x\beta + \varepsilon$, where y is the scalar response and x is a scalar predictor, one can learn about the regression parameter β from the orthogonality assumption $\mathbf{E}^P[(y - x\beta)x] = 0$ without fully modeling the error distribution or the parameters of the error distribution. More generally, β can be inferred in this setting from the orthogonality conditions $\mathbf{E}^P[(y - x\beta)\boldsymbol{z}] = \mathbf{0}$, given a set of instrumental variables $\boldsymbol{z}$. Examples of such moment condition models abound, but for the most part the analysis of such models from the Bayesian perspective has proved elusive since typical parametric and semiparametric Bayesian methods are reliant on a full probability model of P.

On the frequentist side, the recent developments in empirical likelihood (EL) based methods, see *e.g.* Owen (1988, 1990, 2001), Qin and Lawless (1994), Kitamura and Stutzer (1997), Imbens (1997), Schennach (2007), Chen and Van Keilegom (2009), and references therein, have opened up a promising approach for dealing with moment condition models. There are emerging cogent arguments for using the EL in Bayesian analysis. For example, Lazar (2003) has argued that the EL can be used in a Bayesian framework in place of the data distribution P. In fact, Schennach (2005) shows that it is possible to obtain a nonparametric likelihood closely related to EL, called the exponentially tilted empirical likelihood (ETEL), by marginalizing over P with a nonparametric prior that favors distributions that are close to the empirical distribution function in terms of the Kullback-Leibler (KL) divergence while satisfying the moment restrictions. In addition, Grendar and Judge (2009) show that the EL is the mode of the posterior of P under a general prior on P. Thus, by combining either the EL or the ETEL functions with a prior $\pi(\boldsymbol{\theta})$ on $\boldsymbol{\theta}$, moment condition models can in principle

be subjected to a Bayesian semiparametric analysis. Applications of this idea are given for instance by Lancaster and Jun (2010), Kim and Yang (2011), Yang and He (2012), Xi et al. (2016) to handle moment condition models, by Rao and Wu (2010) in complex survey estimation, and by Chaudhuri and Ghosh (2011), Porter et al. (2015), Chaudhuri et al. (2017) in small area estimation. On the theory side, Yang and He (2012) show the asymptotic normality of the Bayesian EL posterior distribution of the quantile regression parameter, and Fang and Mukerjee (2006) and Chang and Mukerjee (2008) study the higher-order asymptotic and coverage properties of the Bayesian EL/ETEL posterior distribution for the population mean, while Schennach (2005) and Lancaster and Jun (2010) consider the large-sample behavior of the Bayesian ETEL posterior distribution under the assumption that all moment restrictions are valid. Alternative, non-EL/ETEL based approaches for moment condition models, which we do not consider in this paper, have also been examined, for example, Bornn et al. (2015), Florens and Simoni (2016) and Kitamura and Otsu (2011).

The purpose of this paper is to establish a number of new results for the Bayesian analysis of moment condition models, within the ETEL framework, complementing and extending the aforementioned papers in important directions. One goal is the Bayesian analysis of moment condition models that are potentially misspecified. For this reason, our analysis is built on the ETEL function which, as shown by Schennach (2007), leads to frequentist estimators of $\boldsymbol{\theta}$ that have the same orders of bias and variance (as a function of the sample size) as the EL estimators but, importantly, maintain the root n convergence even under model misspecification (see Schennach (2007, Theorem 1)). Within this useful framework, we develop a fully Bayesian treatment of correctly and misspecified moment condition models. We show that even under misspecification, the Bayesian ETEL posterior distribution has desirable properties, and that it satisfies the Bernstein - von Mises (BvM) theorem. Another goal is to develop a Bayesian approach for comparing different moment restricted models and for discarding any misspecified moment restrictions. For an overview on Bayesian model selection in standard models we refer to Robert (2007) and references therein. Our proposal is to select the model with the largest marginal likelihood. Since one aim of this model selection comparison is to discard misspecified moment restrictions, we do not consider the model averaging perspective. In order to operationalize model comparisons in our set-up, in particular when models are defined by different numbers of moment conditions, we show that

it is necessary to linearly transform the moment functions $\boldsymbol{g}(\boldsymbol{X}, \boldsymbol{\theta})$ so that all the transformed moments are included in each model. This linear transformation simply consists of adding an extra parameter different from zero to the components of the vector $\boldsymbol{g}(\boldsymbol{X}, \boldsymbol{\theta})$ that correspond to the restrictions not included in a specific model.

We compute the marginal likelihood by the method of Chib (1995), as extended to Metropolis-Hastings samplers in Chib and Jeliazkov (2001). This method renders computation of the marginal likelihood simple and is a key feature of both our numerical and theoretical analysis. Our asymptotic theory covers the following exhaustive possibilities: the case where the models in the comparison set contain only valid moment restrictions, the case where all the models in the set are misspecified, and finally the case where some of the models contain only valid moment restrictions while the others contain at least one invalid moment restriction. Our analysis shows that the marginal likelihood based selection procedure is consistent in the sense that: *(i)* it discards misspecified moment restrictions, *(ii)* it selects the model that is the "less misspecified" when comparing models that are all misspecified, *(iii)* it selects the model that contains the maximum number of overidentifying valid moment restrictions when comparing correctly specified models, and *(iv)* when some models are correctly specified and some are misspecified, it selects the model that is correctly specified and contains the maximum number of overidentifying moment conditions. These important model selection consistency results are based on the asymptotic behavior of the ETEL function, and the validity of the BvM theorem, both under correct specification and misspecificition. These results, developed within a formal Bayesian setting, can be viewed as complementary to the less Bayesian formulations described in Variyath et al. (2010) and Vexler et al. (2013) where the focus is on quasi-Bayes factors constructed from the EL, and Hong and Preston (2012) where models are compared based on a quasi-marginal likelihood obtained from an approximation to the true P.

The rest of the article is organized as follows. In Section 2 we describe the moment condition model, define the notion of misspecification in this setting, and then discuss the prior-posterior analysis with the ETEL function. We then provide the first pair of major results dealing with the asymptotic behavior of the posterior distribution for both correctly specified and misspecified models. Section 3 introduces our model selection procedure based on marginal likelihoods and the associated large sample results. Throughout the paper,

for expository purposes, we include numerical examples. Then in Section 4 we discuss the problems of variable selection in a count regression model and instrument validity in an instrumental variable regression. Section 5 concludes. Proofs of our results are collected in the Appendix and in the online Appendix.

2 Setting

Suppose that $\boldsymbol{X}$ is an $\mathbb{R}^{d_x}$-valued random vector with (unknown) distribution P. Suppose that the operating assumption is that the distribution P satisfies the d unconditional moment restrictions

$$\mathbf{E}^P[\boldsymbol{g}(\boldsymbol{X}, \boldsymbol{\theta})] = \mathbf{0} \tag{2.1}$$

where $\mathbf{E}^P$ denotes the expectation taken with respect to P, $\boldsymbol{g} : \mathbb{R}^{d_x} \times \Theta \mapsto \mathbb{R}^d$ is a vector of known functions with values in $\mathbb{R}^d$, $\boldsymbol{\theta} := (\theta_1, \ldots, \theta_p)' \in \Theta \subset \mathbb{R}^p$ is the parameter vector of interest, and $\mathbf{0}$ is the $d \times 1$ vector of zeros. We assume that $\mathbf{E}^P[\boldsymbol{g}(\boldsymbol{X}, \boldsymbol{\theta})]$ is bounded for every $\boldsymbol{\theta} \in \Theta$. We also suppose that we are given a random sample $\boldsymbol{x}_{1:n} := (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ on $\boldsymbol{X}$ and that $d \geq p$.

When the number of moment restrictions d exceeds the number of parameters p, the parameter $\boldsymbol{\theta}$ in such a setting is said to be overidentified (over restricted). In such a case, there is a possibility that a subset of the moment conditions may be invalid in the sense that the true data generating process is not contained in the collection of probability measures that satisfy the moment conditions for all $\boldsymbol{\theta} \in \Theta$. That is, there is no parameter $\boldsymbol{\theta}$ in Θ that is consistent with the moment restrictions (2.1) under the true data generating process P. To deal with possibly invalid moment restrictions, we reformulate the moment conditions in terms of an additional nuisance parameter $\boldsymbol{V} \in \mathfrak{V} \subset \mathbb{R}^d$. For example, if the k-th moment condition is not expected to be valid, we subtract $\boldsymbol{V} = (V_1, \ldots, V_d)$ from the moment restrictions where V_k is a free parameter and all other elements of $\boldsymbol{V}$ are zero. To accommodate this situation, we rewrite the above conditions as the following augmented moment conditions

$$\mathbf{E}^P[\boldsymbol{g}^A(\boldsymbol{X}, \boldsymbol{\theta}, \boldsymbol{V})] = \mathbf{0} \tag{2.2}$$

where $\boldsymbol{g}^A(\boldsymbol{X}, \boldsymbol{\theta}, \boldsymbol{V}) := \boldsymbol{g}(\boldsymbol{X}, \boldsymbol{\theta}) - \boldsymbol{V}$. Note that in this formalism, the parameter $\boldsymbol{V}$ indicates

which moment restrictions are active where by 'active moment restrictions' we mean the restrictions for which the corresponding components of $\boldsymbol{V}$ are zero. In order to guarantee identification of $\boldsymbol{\theta}$, at most $(d-p)$ elements of $\boldsymbol{V}$ can be different than zero. If all the elements of $\boldsymbol{V}$ are zero, we recover the restrictions in (2.1).

Let $d_v \leq (d-p)$ be the number of non-zero elements of $\boldsymbol{V}$ and let $\boldsymbol{v} \in \mathcal{V} \subset \mathbb{R}^{d_v}$ be the vector that collects all the non-zero components of $\boldsymbol{V}$. We call $\boldsymbol{v}$ the augmented parameter and $\boldsymbol{\theta}$ the parameter of interest. Therefore, the number of active moment restrictions is $d - d_v$. In the following, we write $\boldsymbol{g}^A(\boldsymbol{X}, \boldsymbol{\theta}, \boldsymbol{v})$ as a shorthand for $\boldsymbol{g}^A(\boldsymbol{X}, \boldsymbol{\theta}, \boldsymbol{V})$, with $\boldsymbol{v}$ the vector obtained from $\boldsymbol{V}$ by collecting only its non-zero components.

The central problem of misspecification of the moment conditions, mentioned in the preceding paragraph, can now be formally defined in terms of the augmented moment conditions.

Definition 2.1 (Misspecified model)**.** *We say that the augmented moment condition model is misspecified if the set of probability measures implied by the moment restrictions does not contain the true data generating process P for every $(\boldsymbol{\theta}, \boldsymbol{v}) \in \Theta \times \mathcal{V}$, that is, $P \notin \mathcal{P}$ where $\mathcal{P} = \bigcup_{(\boldsymbol{\theta},\boldsymbol{v}) \in \Theta \times \mathcal{V}} \mathcal{P}_{(\boldsymbol{\theta},\boldsymbol{v})}$ and $\mathcal{P}_{(\boldsymbol{\theta},\boldsymbol{v})} = \{Q \in \mathbb{M};\, \mathbf{E}^Q[\boldsymbol{g}^A(\boldsymbol{X}, \boldsymbol{\theta}, \boldsymbol{v})] = \mathbf{0}\}$ with $\mathbb{M}$ the set of all probability measures on $\mathbb{R}^{d_x}$.*

In a nutshell, a set of augmented moment conditions is misspecified if there is no pair $(\boldsymbol{\theta}, \boldsymbol{v})$ in $(\Theta \times \mathcal{V})$ that satisfies $\mathbf{E}^P[\boldsymbol{g}^A(\boldsymbol{X}, \boldsymbol{\theta}, \boldsymbol{v})] = \mathbf{0}$ where P is the true data generating process. On the other hand, if such a pair of values $(\boldsymbol{\theta}, \boldsymbol{v})$ exists then the set of augmented moment conditions is correctly specified.

Throughout the paper, we use regression models to understand the various concepts and ideas.

Example 1 (Linear regression model)**.** *Suppose that we are interested in estimating the following linear regression model with an intercept and a predictor:*

$$y_i = \alpha + \beta z_i + e_i, \qquad i = 1, \ldots, n \tag{2.3}$$

where $(z_i, e_i)'$ are independently drawn from some distribution P. Under the assumption that

$\mathbf{E}^P[e_i|z_i] = 0$, *we can use the following moment restrictions to estimate* $\boldsymbol{\theta} := (\alpha, \beta)$*:*

$$\mathbf{E}^P[e_i(\boldsymbol{\theta})] = 0, \quad \mathbf{E}^P[e_i(\boldsymbol{\theta})z_i] = 0, \quad \mathbf{E}^P[(e_i(\boldsymbol{\theta}))^3] = v, \tag{2.4}$$

where $e_i(\boldsymbol{\theta}) := (y_i - \alpha - \beta z_i)$*. The first two moment restrictions are derived from the standard orthogonality condition and identify* $\boldsymbol{\theta}$*. The last restriction potentially serves as additional information. In terms of the notation in* (2.1) *and* (2.2)*,* $\boldsymbol{x}_i := (y_i, z_i)$*,* $\boldsymbol{g}(\boldsymbol{x}_i, \boldsymbol{\theta}) = (e_i(\boldsymbol{\theta}), e_i(\boldsymbol{\theta})z_i, e_i(\boldsymbol{\theta})^3)'$*,* $\boldsymbol{V} = (0, 0, v)'$*,* $d_v = 1$ *and* $\boldsymbol{g}^A(\boldsymbol{x}_i, \boldsymbol{\theta}, \boldsymbol{V}) = \boldsymbol{g}(\boldsymbol{x}_i, \boldsymbol{\theta}) - (0, 0, v)'$*. If one believes that the underlying distribution of* e_i *is indeed symmetric, then one could use this information by setting* v *to zero. Otherwise, it is desirable to treat* v *as an unknown object. If the distribution of* e_i *is skewed and* v *is forced to be zero, then the model becomes misspecified because no* (α, β) *can be consistent with the three moment restrictions jointly under* P*. When the augmented parameter* v *is treated as a free parameter, the model is correctly specified even under asymmetry.*

2.1 Prior-Posterior analysis

Consider now the question of prior-posterior analysis under the ETEL function. Although our setting is similar to that of Schennach (2005), the presence of the augmented parameter $\boldsymbol{v}$ and the possibility of misspecification, lead to a new analysis and new results.

For any $(\boldsymbol{\theta}, \boldsymbol{v})$, define the convex hull of $\bigcup_{i=1}^n \boldsymbol{g}^A(\boldsymbol{x}_i, \boldsymbol{\theta}, \boldsymbol{v})$ as the following convex subset of $\mathbb{R}^d$: $\{\sum_{i=1}^n p_i \boldsymbol{g}^A(\boldsymbol{x}_i, \boldsymbol{\theta}, \boldsymbol{v}); p_i \geq 0, \forall i = 1, \ldots, n, \sum_{i=1}^n p_i = 1\}$. Now suppose that *(i)* $\boldsymbol{g}^A(\boldsymbol{x}, \boldsymbol{\theta}, \boldsymbol{v})$ is continuous in $\boldsymbol{x}$ for every $(\boldsymbol{\theta}, \boldsymbol{v}) \in \Theta \times \mathcal{V}$ (or has a finite number of step discontinuities) and *(ii)* the interior of the convex hull of $\bigcup_{i=1}^n \boldsymbol{g}^A(\boldsymbol{x}_i, \boldsymbol{\theta}, \boldsymbol{v})$ contains the origin. Suppose also that the nonparametric prior on P is the mixture of uniform probability densities described in Schennach (2005), which is capable to approximating any distribution as the number of mixing components increases. Then, adapting the arguments of Schennach (2005), the posterior distribution of $(\boldsymbol{\theta}, \boldsymbol{v})$ after marginalization over P has the form

$$\pi(\boldsymbol{\theta}, \boldsymbol{v}|\boldsymbol{x}_{1:n}) \propto \pi(\boldsymbol{\theta}, \boldsymbol{v}) p(\boldsymbol{x}_{1:n}|\boldsymbol{\theta}, \boldsymbol{v}) \tag{2.5}$$

where $\pi(\boldsymbol{\theta}, \boldsymbol{v})$ is the prior of $(\boldsymbol{\theta}, \boldsymbol{v})$ and $p(\boldsymbol{x}_{1:n}|\boldsymbol{\theta}, \boldsymbol{v})$ is the ETEL function defined as

$$p(\boldsymbol{x}_{1:n}|\boldsymbol{\theta}, \boldsymbol{v}) = \prod_{i=1}^{n} p_i^*(\boldsymbol{\theta}, \boldsymbol{v}) \tag{2.6}$$

and $p_i^*(\boldsymbol{\theta}, \boldsymbol{v})$ are the probabilities that minimize the KL divergence between the probabilities $(p_1, \ldots, p_n)$ assigned to each sample observation and the empirical probabilities $(\frac{1}{n}, \ldots, \frac{1}{n})$, subject to the conditions that the probabilities $(p_1, \ldots, p_n)$ sum to one and that the expectation under these probabilities satisfies the given moment conditions:

$$\max_{p_1,\ldots,p_n} \sum_{i=1}^{n} [-p_i \log(np_i)] \tag{2.7}$$

$$\text{subject to} \qquad \sum_{i=1}^{n} p_i \;=\; 1 \quad \text{and} \quad \sum_{i=1}^{n} p_i \boldsymbol{g}^A(\boldsymbol{x}_i, \boldsymbol{\theta}, \boldsymbol{v}) = \mathbf{0}. \tag{2.8}$$

For numerical and theoretical purposes below, the preceding probabilities are computed more conveniently from the dual (saddlepoint) representation as, for $i = 1, \ldots, n$

$$p_i^*(\boldsymbol{\theta}, \boldsymbol{v}) := \frac{e^{\widehat{\boldsymbol{\lambda}}(\boldsymbol{\theta},\boldsymbol{v})' \boldsymbol{g}^A(\boldsymbol{x}_i,\boldsymbol{\theta},\boldsymbol{v})}}{\sum_{j=1}^{n} e^{\widehat{\boldsymbol{\lambda}}(\boldsymbol{\theta},\boldsymbol{v})' \boldsymbol{g}^A(\boldsymbol{x}_j,\boldsymbol{\theta},\boldsymbol{v})}}, \qquad \text{where } \widehat{\boldsymbol{\lambda}}(\boldsymbol{\theta}, \boldsymbol{v}) = \arg\min_{\boldsymbol{\lambda} \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} \exp\left(\boldsymbol{\lambda}' \boldsymbol{g}^A(\boldsymbol{x}_i, \boldsymbol{\theta}, \boldsymbol{v})\right). \tag{2.9}$$

Therefore, the posterior distribution takes the form

$$\pi(\boldsymbol{\theta}, \boldsymbol{v}|\boldsymbol{x}_{1:n}) \propto \pi(\boldsymbol{\theta}, \boldsymbol{v}) \prod_{i=1}^{n} \frac{e^{\widehat{\boldsymbol{\lambda}}(\boldsymbol{\theta},\boldsymbol{v})' \boldsymbol{g}^A(\boldsymbol{x}_i,\boldsymbol{\theta},\boldsymbol{v})}}{\sum_{j=1}^{n} e^{\widehat{\boldsymbol{\lambda}}(\boldsymbol{\theta},\boldsymbol{v})' \boldsymbol{g}^A(\boldsymbol{x}_j,\boldsymbol{\theta},\boldsymbol{v})}}, \tag{2.10}$$

which may be called the Bayesian Exponentially Tilted Empirical Likelihood (BETEL) posterior distribution. It can be efficiently simulated by Markov chain Monte Carlo (MCMC) methods. For example, the one block tailored Metropolis-Hastings (M-H) algorithm (Chib and Greenberg, 1995) is applied as follows. Let $q(\boldsymbol{\theta}, \boldsymbol{v}|\boldsymbol{x}_{1:n})$ denote a student-t distribution whose location parameter is the mode of the log ETEL function and whose dispersion matrix is the inverse of the negative Hessian matrix of the log ETEL function at the mode. Then, starting from some initial value $(\boldsymbol{\theta}^{(0)}, \boldsymbol{v}^{(0)})$, we get a sample of draws from the BETEL posterior by repeating the following steps for $s = 1, \ldots, S$:

1. Propose $(\boldsymbol{\theta}^\dagger, \boldsymbol{v}^\dagger)$ from $q(\boldsymbol{\theta}, \boldsymbol{v}|\boldsymbol{x}_{1:n})$ and solve for $p_i^*(\boldsymbol{\theta}^\dagger, \boldsymbol{v}^\dagger)$, $1 \leq i \leq n$, from the Expo-

nential Tilting saddlepoint problem (2.9).

2. Calculate the M-H probability of move

$$\alpha\left((\boldsymbol{\theta}^{s-1},\boldsymbol{v}^{s-1}),(\boldsymbol{\theta}^{\dagger},\boldsymbol{v}^{\dagger})\middle|\,\boldsymbol{x}_{1:n}\right)=\min\left\{1,\frac{\pi(\boldsymbol{\theta}^{\dagger},\boldsymbol{v}^{\dagger}|\boldsymbol{x}_{1:n})}{\pi(\boldsymbol{\theta}^{s-1},\boldsymbol{v}^{s-1}|\boldsymbol{x}_{1:n})}\frac{q(\boldsymbol{\theta}^{s-1},\boldsymbol{v}^{s-1}|\boldsymbol{x}_{1:n})}{q(\boldsymbol{\theta}^{\dagger},\boldsymbol{v}^{\dagger}|\boldsymbol{x}_{1:n})}\right\}.$$

3. Set $(\boldsymbol{\theta}^{s},\boldsymbol{v}^{s})=(\boldsymbol{\theta}^{\dagger},\boldsymbol{v}^{\dagger})$ with probability $\alpha((\boldsymbol{\theta}^{s-1},\boldsymbol{v}^{s-1}),(\boldsymbol{\theta}^{\dagger},\boldsymbol{v}^{\dagger})|\boldsymbol{x}_{1:n})$. Otherwise, set $(\boldsymbol{\theta}^{s},\boldsymbol{v}^{s})=(\boldsymbol{\theta}^{s-1},\boldsymbol{v}^{s-1})$. Go to step 1.

Note that when the dimension of $(\boldsymbol{\theta},\boldsymbol{v})$ is large, the Tailored Randomized Block M-H algorithm of Chib and Ramamurthy (2010) can be used instead for improved simulation efficiency.

Prior specification. In our examples, we focus on two prior distributions. Under the first prior, which we call the default prior, each element θ_k and v_l of $\boldsymbol{\theta}$ and $\boldsymbol{v}$, respectively, is given independent student-t distributions with $\nu=2.5$ degrees of freedom, location zero and dispersion equal to 5:

$$\theta_k\sim t_{2.5}(0,5^2)\quad\text{ and }\quad v_l\sim t_{2.5}(0,5^2). \tag{2.11}$$

In the second prior, which we call the training sample prior, an initial portion of the sample (which is not used for subsequent inferences) is used to find the ETEL estimate of the unknown parameters, that is, the maximizer of the ETEL function (2.6) whose definition is recalled in (C.1) in the Appendix. Then, the prior of each element of $(\boldsymbol{\theta}',\boldsymbol{v}')'$ is equal to the default prior except that now the location is set equal to the corresponding ETEL estimate.

To see the different implications of these prior distributions, consider two moment condition models defined by the restrictions:

$$\begin{aligned}M_1:&\quad \mathbf{E}^P[g_1(\boldsymbol{X},\boldsymbol{\theta})]=0,\quad \mathbf{E}^P[g_2(\boldsymbol{X},\boldsymbol{\theta})]=0\\ M_2:&\quad \mathbf{E}^P[g_1(\boldsymbol{X},\boldsymbol{\theta})]=0,\quad \mathbf{E}^P[g_2(\boldsymbol{X},\boldsymbol{\theta})]=v\end{aligned} \tag{2.12}$$

where both moments restrictions are active under M_1 but only the first is active under M_2. Then, under the default prior, a prior mean of 0 on v implies the belief that the second moment restriction is likely to hold. On the other hand, in the training sample prior, the prior location of v is determined by the ETEL estimate of v in the training sample. If this is

substantially different from zero (relative to the prior dispersion) this prior implies the belief that the second moment restriction is, a priori, less likely to be active.

Example 1 (continued)**.** *To illustrate the prior-posterior analysis, we generate* y_i*,* $i = 1, \ldots, n$ *from the regression model in* (2.3) *with the covariate* $z_i \sim \mathcal{N}(0.5, 1)$*, intercept* $\alpha = 0$*, slope* $\beta = 1$ *and* e_i *distributed according to the skewed distribution:*

$$e_i \sim \begin{cases} \mathcal{N}(0.75, 0.75^2) & \textit{with probability } 0.5 \\ \mathcal{N}(-0.75, 1.25^2) & \textit{with probability } 0.5. \end{cases} \tag{2.13}$$

Our analysis is based on the moment restrictions in (2.4)*, that is,*

$$\boldsymbol{g}^A(\boldsymbol{x}_i, \boldsymbol{\theta}, v) = (e_i(\boldsymbol{\theta}), e_i(\boldsymbol{\theta}) z_i, e_i(\boldsymbol{\theta})^3 - v)', \qquad e_i(\boldsymbol{\theta}) = y_i - \alpha - \beta z_i,$$

with $\boldsymbol{\theta} = (\alpha, \beta)$*. These moment conditions are correctly specified because* v *is free. Under the default independent student-t prior in (2.11), the marginal posterior distributions of* α*,* β *and* v *are summarized in Table 1 for two different values of* n*. It can be seen from the .025 and .975 quantiles (called "lower" and "upper", respectively) that the marginal posterior distributions of* α *and* β *are already concentrated around the true values for* $n = 250$ *but concentrate even more closely around the true values for* $n = 2000$*. This example showcases the ease with which such Bayesian inferences are possible.*

	mean	sd	median	lower	upper	ineff
			$n = 250$			
α	-0.03	0.10	-0.03	-0.24	0.16	1.49
β	0.99	0.08	0.98	0.83	1.15	1.70
v	-1.42	0.36	-1.39	-2.20	-0.82	3.21
			$n = 2000$			
α	0.00	0.03	0.00	-0.06	0.06	1.30
β	0.99	0.03	0.99	0.93	1.04	1.21
v	-0.97	0.10	-0.97	-1.18	-0.78	1.34

Table 1: Posterior summary for two simulated sample sizes from Example 1 (a regression model with skewed error distribution). The true value of α is 0 and that of β is 1. The summaries are based on $10,000$ MCMC draws beyond a burn-in of 1000. The M-H acceptance rate is around 90% in both cases. "Lower" and "upper" refer to the .025 and .975 quantiles of the simulated draws, respectively, and "ineff" to the inefficiency factor, the ratio of the numerical variance of the mean to the variance of the mean assuming independent draws: an inefficiency factor close to 1 indicates that the MCMC draws, although serially correlated, are essentially independent.

Notation. In Sections 2.2 and 2.3, and in the online Appendix we use the following notations. For ease of exposition, we denote $\boldsymbol{\psi} := (\boldsymbol{\theta}, \boldsymbol{v})$, $\boldsymbol{\psi} \in \Psi$ with $\Psi := \Theta \times \mathcal{V}$. Moreover, $\|\cdot\|_F$ denotes the Frobenius norm and $\|\cdot\|$ the Euclidean norm. The notation '$\xrightarrow{p}$' is for convergence in probability with respect to the product measure $P^n = \bigotimes_{i=1}^n P$. The log-likelihood function for one observation is denoted by $l_{n,\boldsymbol{\psi}}$:

$$l_{n,\boldsymbol{\psi}}(\boldsymbol{x}) := \log \frac{e^{\widehat{\boldsymbol{\lambda}}(\boldsymbol{\psi})' \boldsymbol{g}^A(\boldsymbol{x},\boldsymbol{\psi})}}{\sum_{j=1}^n e^{\widehat{\boldsymbol{\lambda}}(\boldsymbol{\psi})' \boldsymbol{g}^A(\boldsymbol{x}_j,\boldsymbol{\psi})}} = -\log n + \log \frac{e^{\widehat{\boldsymbol{\lambda}}' \boldsymbol{g}^A(\boldsymbol{x},\boldsymbol{\psi})}}{\frac{1}{n}\sum_{j=1}^n \left[e^{\widehat{\boldsymbol{\lambda}}' \boldsymbol{g}^A(\boldsymbol{x}_j,\boldsymbol{\psi})}\right]}$$

so that the log-ETEL function is $\log p(\boldsymbol{x}_{1:n}|\boldsymbol{\psi}) = \sum_{i=1}^n l_{n,\boldsymbol{\psi}}(\boldsymbol{x}_i)$. For a set $\mathcal{A} \subset \mathbb{R}^m$, we denote by $int(\mathcal{A})$ its interior relative to $\mathbb{R}^m$. Further notations are introduced as required.

2.2 Asymptotic Properties: correct specification

In this section, we first introduce additional notations and assumptions for correctly specified models. Under these assumptions and Assumptions 5-6 in the online Appendix, we establish both the large sample behavior of the BETEL posterior distribution and, in Section 3, the model selection consistency of our marginal likelihood procedure.

Let $\boldsymbol{\theta}_*$ be the true value of the parameter of interest $\boldsymbol{\theta}$ and $\boldsymbol{v}_*$ be the true value of the augmented parameter. So, $\boldsymbol{\psi}_* := (\boldsymbol{\theta}_*, \boldsymbol{v}_*)$. The true value $\boldsymbol{v}_*$ is equal to zero when the non-augmented model (2.1) is correctly specified. Moreover, let $\Delta := \mathbf{E}^P[\boldsymbol{g}^A(\boldsymbol{X}, \boldsymbol{\psi}_*)\boldsymbol{g}^A(\boldsymbol{X}, \boldsymbol{\psi}_*)']$ and $\Gamma := \mathbf{E}^P\left[\frac{\partial}{\partial \boldsymbol{\psi}'}\boldsymbol{g}^A(\boldsymbol{X}, \boldsymbol{\psi}_*)\right]$. Assumption 1 requires that the augmented model is correctly specified in the sense that there is a value of $\boldsymbol{\psi}$ such that (2.2) is satisfied by P, and that this value is unique. A necessary condition for the latter is that $(d - p) \geq d_v \geq 0$.

Assumption 1. *Model* (2.2) *is such that* $\boldsymbol{\psi}_* \in \Psi$ *is the unique solution to* $\mathbf{E}^P[\boldsymbol{g}^A(\boldsymbol{X}, \boldsymbol{\psi})] = \mathbf{0}$.

The next assumption concerns the prior distribution and is a standard assumption to establish asymptotic properties of Bayesian procedures.

Assumption 2. *(a)* π *is a continuous probability measure that admits a density with respect to the Lebesgue measure; (b)* π *is positive on a neighborhood of* $\boldsymbol{\psi}_*$.

For a correctly specified moment conditions model, the asymptotic normality of the BETEL posterior is established in the following theorem where we denote by $\pi(\sqrt{n}(\boldsymbol{\psi} -$

$\boldsymbol{\psi}_*)|\boldsymbol{x}_{1:n})$ the posterior distribution of $\sqrt{n}(\boldsymbol{\psi} - \boldsymbol{\psi}_*)$. The result shows that the BETEL posterior distribution has a Gaussian limiting distribution and that it concentrates on a $n^{-1/2}$-ball centered at the true value of the parameter. An informal discussion of this behavior is given by Schennach (2005) but without the required assumptions. Theorem 2.1 below provides these assumptions. The proof of the result is based on *e.g.* Lehmann and Casella (1998) and Ghosh and Ramamoorthi (2003) and is given in the online Appendix C.

Theorem 2.1 (Bernstein - von Mises – correct specification)**.** *Under Assumptions 1, 2 and Assumptions 5, 6 in the online Appendix and if in addition, for any $\delta > 0$, there exists an $\epsilon > 0$ such that, as $n \to \infty$*

$$P\left(\sup_{\|\boldsymbol{\psi}-\boldsymbol{\psi}_*\|>\delta} \frac{1}{n}\sum_{i=1}^{n}\left(l_{n,\boldsymbol{\psi}}(\boldsymbol{x}_i) - l_{n,\boldsymbol{\psi}_*}(\boldsymbol{x}_i)\right) \leq -\epsilon\right) \to 1, \tag{2.14}$$

then the posteriors converge in total variation towards a normal distribution, that is,

$$\sup_{B}\left|\pi(\sqrt{n}(\boldsymbol{\psi} - \boldsymbol{\psi}_*) \in B|\boldsymbol{x}_{1:n}) - \mathcal{N}_{0,(\Gamma'\Delta^{-1}\Gamma)^{-1}}(B)\right| \xrightarrow{p} 0 \tag{2.15}$$

where $B \subseteq \Psi$ is any Borel set.

According to this result, the posterior distribution $\pi(\boldsymbol{\psi}|\boldsymbol{x}_{1:n})$ of $\boldsymbol{\psi}$ is asymptotically normal, centered on the true value $\boldsymbol{\psi}_*$ and with variance $n^{-1}\left(\Gamma'\Delta^{-1}\Gamma\right)^{-1}$. Thus, the posterior distribution has the same asymptotic variance as the efficient Generalized Method of Moments estimator of Hansen (1982) (see also Chamberlain (1987)). Assumption (2.14) in this theorem is a standard identifiability condition (see *e.g.* Lehmann and Casella (1998, Assumption 6.B.3)) that controls the behavior of the log-ETEL function at a distance from $\boldsymbol{\psi}_*$. Controlling this behavior is important because the posterior involves integration over the whole range of $\boldsymbol{\psi}$. To understand the meaning of this assumption, we remark that asymptotically the log-ETEL function $\boldsymbol{\psi} \mapsto \sum_{i=1}^{n} l_{n,\boldsymbol{\psi}}(\boldsymbol{x}_i)$ is maximized at the true value $\boldsymbol{\psi}_*$ because the model is correctly specified. Hence, Assumption (2.14) means that if the parameter $\boldsymbol{\psi}$ is "far" from the true value $\boldsymbol{\psi}_*$ then the log-ETEL function has to be small, that is, has to be far from the maximum value $\sum_{i=1}^{n} l_{n,\boldsymbol{\psi}_*}(\boldsymbol{x}_i)$.

2.3 Asymptotic Properties: misspecification

In this section, we consider the case where the model is misspecified in the sense of Definition 2.1 and establish that, even in this case, the BETEL posterior distribution has good frequentist asymptotic properties as the sample size n increases. Namely, we show that the BETEL posterior of $\sqrt{n}(\boldsymbol{\psi} - \boldsymbol{\psi}_*)$ is asymptotically normal and the BETEL posterior of $\boldsymbol{\psi}$ concentrates on a $n^{-1/2}$-ball centred at the pseudo-true value of the parameter. To the best of our knowledge, these properties have not been established yet for misspecified moment condition models.

Because in misspecified models there is no value of $\boldsymbol{\psi}$ for which the true data distribution P satisfies the restriction (2.2), we need to define a pseudo-true value for $\boldsymbol{\psi}$. The latter is defined as the value of $\boldsymbol{\psi}$ that minimizes the KL divergence $K(P||Q^*(\boldsymbol{\psi}))$ between the true data distribution P and a distribution $Q^*(\boldsymbol{\psi})$ defined as $Q^*(\boldsymbol{\psi}) := \text{arginf}_{Q \in \mathcal{P}_{\boldsymbol{\psi}}} K(Q||P)$, where $K(Q||P) := \int \log(dQ/dP) dQ$ and $\mathcal{P}_{\boldsymbol{\psi}}$ is defined in Definition 2.1. We remark that these two KL divergences are the population counterparts of the KL divergences used for the definition of the ETEL function in (2.6): the empirical counterpart of $K(Q||P)$ is used to construct the $p_i^*(\boldsymbol{\psi})$ probabilities and is given by (2.7), while the empirical counterpart of $K(P||Q^*(\boldsymbol{\psi}))$ is given by $\log(1/n) - \sum_{i=1}^n l_{n,\boldsymbol{\psi}}(\boldsymbol{x}_i)/n$ where $\sum_{i=1}^n l_{n,\boldsymbol{\psi}}(\boldsymbol{x}_i)$ is the log-ETEL function if the dual theorem holds. Roughly speaking, the pseudo-true value is the value of $\boldsymbol{\psi}$ for which the distribution that satisfies the corresponding restrictions (2.2) is the closest to the true P, in the KL sense. By using the dual representation of the KL minimization problem, the P-density $dQ^*(\boldsymbol{\psi})/dP$ admits a closed-form: $dQ^*(\boldsymbol{\psi})/dP = e^{\boldsymbol{\lambda}_\circ(\boldsymbol{\psi})' \boldsymbol{g}^A(\boldsymbol{X},\boldsymbol{\psi})} / \mathbf{E}^P \left[e^{\boldsymbol{\lambda}_\circ(\boldsymbol{\psi})' \boldsymbol{g}^A(\boldsymbol{X},\boldsymbol{\psi})} \right]$ where $\boldsymbol{\lambda}_\circ(\boldsymbol{\psi})$ is the pseudo-true value of the tilting parameter defined as the solution of $\mathbf{E}^P[\exp\{\boldsymbol{\lambda}' \boldsymbol{g}^A(\boldsymbol{X},\boldsymbol{\psi})\} \boldsymbol{g}^A(\boldsymbol{X},\boldsymbol{\psi})] = \mathbf{0}$ which is unique by the strict convexity of $\mathbf{E}^P[\exp\{\boldsymbol{\lambda}' \boldsymbol{g}^A(\boldsymbol{X},\boldsymbol{\psi})\}]$ in $\boldsymbol{\lambda}$. Therefore,

$$\begin{aligned} \boldsymbol{\lambda}_\circ(\boldsymbol{\psi}) &:= \arg\min_{\boldsymbol{\lambda} \in \mathbb{R}^d} \mathbf{E}^P \left[e^{\boldsymbol{\lambda}' \boldsymbol{g}^A(\boldsymbol{X},\boldsymbol{\psi})} \right], \\ \boldsymbol{\psi}_\circ &:= \arg\max_{\boldsymbol{\psi} \in \Psi} \mathbf{E}^P \log \left[\frac{e^{\boldsymbol{\lambda}_\circ(\boldsymbol{\psi})' \boldsymbol{g}^A(\boldsymbol{X},\boldsymbol{\psi})}}{\mathbf{E}^P \left[e^{\boldsymbol{\lambda}_\circ(\boldsymbol{\psi})' \boldsymbol{g}^A(\boldsymbol{X},\boldsymbol{\psi})} \right]} \right]. \end{aligned} \tag{2.16}$$

However, in a misspecified model, the dual theorem is not guaranteed to hold and so $\boldsymbol{\psi}_\circ$ defined in (2.16) is not necessarily equal to the pseudo-true value defined as the KL-minimizer.

In fact, when the model is misspecified, the probability measures in $\mathcal{P} := \bigcup_{\boldsymbol{\psi}\in\Psi} \mathcal{P}_{\boldsymbol{\psi}}$, which are implied by the model, might not have a common support with the true P, see Sueishi (2013) for a discussion on this point. Following Sueishi (2013, Theorem 3.1), in order to guarantee identification of the pseudo-true value by (2.16) and validity of the dual theorem we introduce the following assumption. This assumption replaces Assumption 1 in misspecified models.

Assumption 3. *For a fixed $\boldsymbol{\psi} \in \Psi$, there exists $Q \in \mathcal{P}_{\boldsymbol{\psi}}$ such that Q is mutually absolutely continuous with respect to P, where $\mathcal{P}_{\boldsymbol{\psi}}$ is defined in Definition 2.1.*

This assumption implies that $\mathcal{P}_{\boldsymbol{\psi}}$ is non-empty. A similar assumption is also made by Kleijn and van der Vaart (2012) to establish the BvM under misspecification. Moreover, because consistency in misspecified models is defined with respect to the pseudo-true value $\boldsymbol{\psi}_\circ$, we need to replace Assumption 2 *(b)* by the following assumption which, together with Assumption 2 *(a)*, requires the prior to put enough mass to balls around $\boldsymbol{\psi}_\circ$.

Assumption 4. *The prior distribution π is positive on a neighborhood of $\boldsymbol{\psi}_\circ$ where $\boldsymbol{\psi}_\circ$ is as defined in (2.16).*

A first step to establish the BvM theorem is to prove that the misspecified model satisfies a stochastic Local Asymptotic Normality (LAN) expansion around the pseudo-true value $\boldsymbol{\psi}_\circ$. Namely, that the log-likelihood ratio $l_{n,\boldsymbol{\psi}} - l_{n,\boldsymbol{\psi}_\circ}$, evaluated at a local parameter around the pseudo-true value, is well approximated by a quadratic form. Such a result is established in Theorem C.1 in the online Appendix C. A second key ingredient for establishing the BvM theorem is the requirement that, as $n \to \infty$, the posterior of $\boldsymbol{\psi}$ concentrates and puts all its mass on $\Psi_n := \{\|\boldsymbol{\psi} - \boldsymbol{\psi}_\circ\| \leq M_n/\sqrt{n}\}$, where M_n is any sequence such that $M_n \to \infty$. We prove this result in Theorem C.2 in the online Appendix C.

Theorem 2.2 states that the limit of the posterior distribution of $\sqrt{n}(\boldsymbol{\psi} - \boldsymbol{\psi}_\circ)$ is a Gaussian distribution with mean and variance defined in terms of the population counterpart of $l_{n,\boldsymbol{\psi}}(\boldsymbol{x})$, which we denote by $\mathfrak{L}_{n,\boldsymbol{\psi}}(\boldsymbol{x}) := \log \frac{\exp(\boldsymbol{\lambda}_\circ(\boldsymbol{\psi})'\boldsymbol{g}^A(\boldsymbol{x},\boldsymbol{\psi}))}{\mathbf{E}^P[\exp(\boldsymbol{\lambda}_\circ(\boldsymbol{\psi})'\boldsymbol{g}^A(\boldsymbol{x},\boldsymbol{\psi}))]} - \log n$ and which involves the pseudo-true value $\boldsymbol{\lambda}_\circ$. With this notation, the variance and mean of the Gaussian limiting distribution are $\boldsymbol{V}_{\boldsymbol{\psi}_\circ}^{-1} := -(\mathbf{E}^P[\ddot{\mathfrak{L}}_{n,\boldsymbol{\psi}_\circ}])^{-1}$ and $\Delta_{n,\boldsymbol{\psi}_\circ} := \frac{1}{\sqrt{n}} \sum_{i=1}^n V_{\boldsymbol{\psi}_\circ}^{-1} \dot{\mathfrak{L}}_{n,\boldsymbol{\psi}_\circ}(\boldsymbol{x}_i)$, respectively, where $\dot{\mathfrak{L}}_{n,\boldsymbol{\psi}_\circ}$ and $\ddot{\mathfrak{L}}_{n,\boldsymbol{\psi}_\circ}$ denote the first and second derivatives of the function $\boldsymbol{\psi} \mapsto \mathfrak{L}_{n,\boldsymbol{\psi}}$ evaluated at $\boldsymbol{\psi}_\circ$. Let $\pi(\sqrt{n}(\boldsymbol{\psi} - \boldsymbol{\psi}_\circ)|\boldsymbol{x}_{1:n})$ denote the posterior distribution of $\sqrt{n}(\boldsymbol{\psi} - \boldsymbol{\psi}_\circ)$.

Theorem 2.2 (Bernstein - von Mises – misspecification)**.** *Assume that the matrix* $\boldsymbol{V}_{\boldsymbol{\psi}_\circ}$ *is nonsingular and that Assumptions 2 (a), 3, 4 and Assumptions 5 (a)-(d), 6 (b), 7, and 8 in the online Appendix hold. If in addition there exists a constant* $C > 0$ *such that for any sequence* $M_n \to \infty$*, as* $n \to \infty$

$$P\left(\sup_{\boldsymbol{\psi}\in\Psi_n^c} \frac{1}{n}\sum_{i=1}^{n}\left(l_{n,\boldsymbol{\psi}}(\boldsymbol{x}_i) - l_{n,\boldsymbol{\psi}_\circ}(\boldsymbol{x}_i)\right) \leq -\frac{CM_n^2}{n}\right) \to 1, \tag{2.17}$$

then the posteriors converge in total variation towards a normal distribution, that is,

$$\sup_B \left|\pi(\sqrt{n}(\boldsymbol{\psi}-\boldsymbol{\psi}_\circ)\in B|\boldsymbol{x}_{1:n}) - \mathcal{N}_{\Delta_{n,\boldsymbol{\psi}_\circ},\boldsymbol{V}_{\boldsymbol{\psi}_\circ}^{-1}}(B)\right| \xrightarrow{p} 0 \tag{2.18}$$

where $B \subseteq \Psi$ *is any Borel set.*

Condition (2.17) involves the log-likelihood ratio $l_{n,\boldsymbol{\psi}}(\boldsymbol{x}) - l_{n,\boldsymbol{\psi}_\circ}(\boldsymbol{x})$ and is an identifiability condition, standard in the literature, and with a similar interpretation as condition (2.14). Theorem 2.2 states that, in misspecified models, the sequence of posterior distributions converges in total variation to a sequence of normal distributions with random mean and fixed covariance matrix $\boldsymbol{V}_{\boldsymbol{\psi}_\circ}^{-1}$. By using the first order condition for $\boldsymbol{\psi}_\circ$ it can be shown that the random mean $\Delta_{n,\boldsymbol{\psi}_\circ}$ has mean zero. We stress that the BvM result of Theorem 2.2 for the BETEL posterior distribution does not directly follow from the assumptions and results in Kleijn and van der Vaart (2012) because the ETEL function contains random quantities. Therefore, we need to strengthen the assumptions in order to establish that a stochastic LAN expansion holds for our case.

As the next lemma shows, the quantity $\Delta_{n,\boldsymbol{\psi}_\circ}$ relates to the Schennach (2007)'s ETEL frequentist estimator $\widehat{\boldsymbol{\psi}}$ (whose definition is recalled in (C.1) in the Appendix for convenience). Because of this connection, it is possible to write the location of the normal limit distribution in a more familiar form in terms of the semi-parametric efficient frequentist estimator $\widehat{\boldsymbol{\psi}}$.

Lemma 2.1. *Assume that the matrix* $\boldsymbol{V}_{\boldsymbol{\psi}_\circ}$ *is nonsingular and that Assumption 3 and Assumptions 5 (a)-(d), 6 (b), 7, and 8 in the online Appendix hold. Then, the ETEL estimator* $\widehat{\boldsymbol{\psi}}$ *satisfies*

$$\sqrt{n}(\widehat{\boldsymbol{\psi}}-\boldsymbol{\psi}_\circ) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\boldsymbol{V}_{\boldsymbol{\psi}_\circ}^{-1}\dot{\boldsymbol{\mathfrak{L}}}_{n,\boldsymbol{\psi}_\circ} + o_p(1). \tag{2.19}$$

Therefore, Lemma 2.1 implies that the BvM Theorem 2.2 can be reformulated with the sequence $\sqrt{n}(\widehat{\boldsymbol{\psi}} - \boldsymbol{\psi}_\circ)$ as the location of the normal limit distribution, that is,

$$\sup_B \left| \pi(\boldsymbol{\psi} \in B | \boldsymbol{x}_{1:n}) - \mathcal{N}_{\widehat{\boldsymbol{\psi}}, n^{-1}\boldsymbol{V}_{\boldsymbol{\psi}_\circ}^{-1}}(B) \right| \xrightarrow{p} 0. \tag{2.20}$$

Two remarks are in order: (I) the limit distribution of $\sqrt{n}(\widehat{\boldsymbol{\psi}} - \boldsymbol{\psi}_\circ)$ is centred on zero because $\mathbf{E}^P[\dot{\mathfrak{L}}_{n,\boldsymbol{\psi}_\circ}] = 0$; (II) the asymptotic covariance matrix of $\sqrt{n}(\widehat{\boldsymbol{\psi}} - \boldsymbol{\psi}_\circ)$ is $\boldsymbol{V}_{\boldsymbol{\psi}_\circ}^{-1}\mathbf{E}^P[\dot{\mathfrak{L}}_{n,\boldsymbol{\psi}_\circ}\dot{\mathfrak{L}}'_{n,\boldsymbol{\psi}_\circ}]\boldsymbol{V}_{\boldsymbol{\psi}_\circ}^{-1}$ (which is also derived in Schennach (2007, Theorem 10)) and, because of misspecification, it does not coincide with the limiting covariance matrix in the BvM theorem. This consequence of misspecification is also discussed in Kleijn and van der Vaart (2012) and implies that, for $\alpha \in (0, 1)$, the central $(1 - \alpha)$ Bayesian credible sets are not in general $(1 - \alpha)$ confidence sets, even asymptotically. In fact, while credible sets are correctly centered, their width/volume need not be correct since the asymptotic variance matrix in the BvM is not the sandwich asymptotic covariance matrix of the frequentist estimator. See Example 2 in the online Appendix A for an illustration of misspecified models and pseudo-true value.

3 Bayesian Model Selection

3.1 Basic idea

Now suppose that there are countable candidate models indexed by ℓ. Suppose that model ℓ is characterized by

$$\mathbf{E}^P[\boldsymbol{g}^\ell(\boldsymbol{X}, \boldsymbol{\theta}^\ell)] = \mathbf{0}, \tag{3.1}$$

with $\boldsymbol{\theta}^\ell \in \Theta^\ell \subset \mathbb{R}^{p_\ell}$, and $\ell = 1, \ldots, J$ for some $J \geq 2$. Different models involve different parameters of interest $\boldsymbol{\theta}^\ell$ and/or different g^ℓ functions. If $\boldsymbol{X}$ contains a dependent variable and covariates it might be that the covariates are not the same for all models, however to lighten the notation we do not explicit this difference across the models.

One or all models may be misspecified. The goal is to compare these models and select the best model. By best model we mean the model that contains the maximum number of over-identifying conditions when all models are correctly specified, and when all models are misspecified, we mean the model that is the closest to the true P. Our purpose in this section is to establish a collection of results on the search for such a best model. We

show that this search can be carried out with the help of the marginal likelihoods (defined as the integral of the sampling density over the parameters with respected to the prior density) of the competing models. The model with the largest marginal likelihood satisfies a model selection consistency property in that the model chosen in this way is the best model asymptotically. This property, which has not been established in this context before, is of enormous practical and theoretical importance.

Before getting to the details, it is crucial to understand that there are some subtleties involved in comparing different moment condition models. The central problem is that the marginal likelihood of models with different sets of moment restrictions and different parameters may not be comparable. In fact, when we have different sets of moment restrictions, we need to be careful about dealing with, and interpreting, unused moment restrictions. This can be best explained by an example.

Example 1 (continued)**.** *Suppose we do not know if e_i is symmetric. In this case, one might be inclined to compare the following two candidate models:*

$$\begin{aligned} &\textit{Model 1}: \mathbf{E}^P[e_i(\boldsymbol{\theta})] = 0, \quad \mathbf{E}^P[e_i(\boldsymbol{\theta})z_i] = 0. \\ &\textit{Model 2}: \mathbf{E}^P[e_i(\boldsymbol{\theta})] = 0, \quad \mathbf{E}^P[e_i(\boldsymbol{\theta})z_i] = 0, \quad \mathbf{E}^P[(e_i(\boldsymbol{\theta}))^3] = 0. \end{aligned} \tag{3.2}$$

where $\boldsymbol{\theta} := (\alpha, \beta)$ is the same parameter in the two models and $e_i(\boldsymbol{\theta}) := (y_i - \alpha - \beta z_i)$. As written, these two models are not comparable because the convex hulls associated with the two models do not have the same dimension. More precisely, let $co_1 := \{\sum_{i=1}^n p_i(e_i(\boldsymbol{\theta}), e_i(\boldsymbol{\theta})z_i)'; p_i \geq 0, \forall i = 1, \ldots, n, \sum_{i=1}^n p_i = 1\}$ be the convex hull associated with Model 1, and $co_2 := \{\sum_{i=1}^n p_i(e_i(\boldsymbol{\theta}), e_i(\boldsymbol{\theta})z_i, e_i(\boldsymbol{\theta})^3)'; p_i \geq 0, \forall i = 1, \ldots, n, \sum_{i=1}^n p_i = 1\}$ be the convex hull associated with Model 2. Because co_1 and co_2 have different dimensions, the $p_i^*(\boldsymbol{\theta})$ in the two ETEL functions are not comparable because they enforce the zero vector constraint (the second constraint in (2.8)) in different spaces ($\mathbb{R}^2$ and $\mathbb{R}^3$).

The foregoing problem can be overcome as follows. We start by defining a grand model that nests all the models that we want to compare. This grand model is constructed such that: (1) it includes all the moment restrictions in the models and, (2) if the same moment restriction is included in two or more models but involves a different parameter in different models, then the grand model includes the moment restriction that involves the parameter

of largest dimension. We write the grand model as $\mathbf{E}^P[\boldsymbol{g}^G(\boldsymbol{X}, \boldsymbol{\theta}^G)] = \mathbf{0}$ where $\boldsymbol{g}^G$ has dimension d, and $\boldsymbol{\theta}^G$ includes the parameters of all models. Next, each original model is obtained from this grand model by first subtracting a vector of nuisance parameters $\boldsymbol{V}$ and then restricting $\boldsymbol{\theta}^G$ and $\boldsymbol{V}$ appropriately. More precisely, an equivalent version of the original model is obtained by: (I) setting equal to zero the components of $\boldsymbol{\theta}^G$ in the shared moment restrictions that are not present in the original model, (II) letting free the components of $\boldsymbol{V}$ that correspond to the over-identifying moment restrictions not present in the original model and, (III) setting equal to zero the components of $\boldsymbol{V}$ that correspond to moment restrictions present in the original model and to moment restrictions that exactly identify the extra parameters arising from the other models.

The set of models that emerge from this strategy are equivalent to the original collection but, crucially, are now comparable. Also importantly for practice, as our illustrations of this strategy show below, the strategy just described is simple to operationalize. With this formulation, model ℓ, denoted by M_ℓ, is then defined from the grand model as

$$\mathbf{E}^P[\boldsymbol{g}^A(\boldsymbol{X}, \boldsymbol{\theta}^\ell, \boldsymbol{v}^\ell)] = 0, \qquad \boldsymbol{\theta}^\ell \in \Theta^\ell \subset \mathbb{R}^{p_\ell} \tag{3.3}$$

where $\boldsymbol{g}^A(\boldsymbol{X}, \boldsymbol{\theta}^\ell, \boldsymbol{v}^\ell) := \boldsymbol{g}^G(\boldsymbol{X}, \boldsymbol{\theta}^\ell) - \boldsymbol{V}^\ell$ with $\boldsymbol{V}^\ell \in \mathfrak{V} \subset \mathbb{R}^d$ and with $\boldsymbol{v}^\ell \in \mathcal{V}^\ell \subset \mathbb{R}^{d_{v_\ell}}$ being the vector that collects all the non-zero components of $\boldsymbol{V}^\ell$. We assume that $0 \leq d_{v_\ell} \leq d - p_\ell$ in order to guarantee identification of $\boldsymbol{\theta}^\ell$. The parameter $\boldsymbol{v}^\ell$ is the augmented parameter and $\boldsymbol{\theta}^\ell$ is the parameter of interest for model ℓ that has been obtained from $\boldsymbol{\theta}^G$ by doing the transformation in (I). Hereafter, we use the notation $\boldsymbol{\psi}^\ell := (\boldsymbol{\theta}^\ell, \boldsymbol{v}^\ell) \in \Psi^\ell$ with $\Psi^\ell := \Theta \times \mathcal{V}^\ell$.

Example 1 (continued)**.** *To be able to compare Model 1 and Model 2 in* (3.2)*, we construct the grand model as* $\mathbf{E}^P[g^G(\boldsymbol{x}_i, \boldsymbol{\theta})] := \mathbf{E}^P[(e_i(\boldsymbol{\theta}), e_i(\boldsymbol{\theta})z_i, e_i(\boldsymbol{\theta})^3)']$*. With respect to this grand model, Model 1 and Model 2 are reformulated as* M_1 *and* M_2*, respectively, by applying (II) and (III) above:*

$$\begin{aligned} M_1 &: \mathbf{E}^P[e_i(\boldsymbol{\theta})] = 0, \quad \mathbf{E}^P[e_i(\boldsymbol{\theta})z_i] = 0, \quad \mathbf{E}^P[(e_i(\boldsymbol{\theta}))^3] = v \\ M_2 &: \mathbf{E}^P[e_i(\boldsymbol{\theta})] = 0, \quad \mathbf{E}^P[e_i(\boldsymbol{\theta})z_i] = 0, \quad \mathbf{E}^P[(e_i(\boldsymbol{\theta}))^3] = 0. \end{aligned} \tag{3.4}$$

So, $\boldsymbol{\psi}^1 = (\boldsymbol{\theta}', \boldsymbol{v})'$ *and* $\boldsymbol{\psi}^2 = \boldsymbol{\theta}$*. The convex hulls of* M_1 *and* M_2 *have both dimension* 3 *and more importantly, if* $co(M_1)$ *and* $co(M_2)$ *denote these two convex hulls, we have* $co(M_1) =$

$co(M_2) - \boldsymbol{V}$ where $\boldsymbol{V} = (0, 0, v)'$ so that the two models are comparable. It is important to note how Model 1 in (3.2) and M_1 deal with uncertainty about the third moment restriction: Model 1 in (3.2) ignores its uncertainty completely while M_1 models the degree of uncertainty through the augmented parameter v. This argument is not limited to comparing two models. When we have multiple models, we need to make sure that the grand model encompasses all candidate models through augmented parameters.

We note that this strategy covers both nested and non-nested models. We say that two models are non-nested, in their original formulation, if neither model can be obtained from the other by eliminating some moment restriction, or by setting to zero some parameter, or both. Points (II) and (III) above are important for the treatment of such non-nested models. In fact, points (II) and (III) imply that, if there are moment restrictions not present in the original model that involve parameters that are not in the original model, then a number of these extra moment restrictions equal to the number of the extra parameters has to be included. This does not alter the original model if these extra moment restrictions exactly identify the extra parameters and so place no restrictions on the data generating process. Moreover, despite the notation, for non-nested models $\boldsymbol{\theta}^\ell$ in (3.3) might be larger than the parameter in the original model ℓ.

In what follows, we show how to compute the marginal likelihood for a model. Then, in Section 3.3 we formally show that, with probability approaching one as the number of observations increases, the marginal likelihood based selection procedure favors the model with the minimum number of parameters of interest and the maximum number of valid moment restrictions. We also consider the situation where all models are misspecified. In this case, our model selection procedure selects the model that is closer to the true data generating process in terms of the KL divergence.

3.2 Marginal Likelihood

For each model M_ℓ, we impose a prior distribution for $\boldsymbol{\psi}^\ell$ on Ψ^ℓ, and obtain the BETEL posterior distribution based on (2.10). Then, we select the model with the largest marginal likelihood, denoted by $m(\boldsymbol{x}_{1:n}; M_\ell)$, which we calculate by the method of Chib (1995) as extended to Metropolis-Hastings samplers in Chib and Jeliazkov (2001). This method makes

computation of the marginal likelihood simple and is a key feature of our procedure. The main advantage of the Chib (1995) method is that it is calculable from the same inputs and outputs that are used in the MCMC sampling of the posterior distribution. The starting point of this method is the following identity of the log-marginal likelihood introduced in Chib (1995):

$$\log m(\boldsymbol{x}_{1:n}; M_\ell) = \log \pi(\tilde{\boldsymbol{\psi}}^\ell | M_\ell) + \log p(\boldsymbol{x}_{1:n} | \tilde{\boldsymbol{\psi}}^\ell, M_\ell) - \log \pi(\tilde{\boldsymbol{\psi}}^\ell | \boldsymbol{x}_{1:n}, M_\ell), \tag{3.5}$$

where $\tilde{\boldsymbol{\psi}}^\ell$ is any point in the support of the posterior (such as the posterior mean) and the dependence on the model M_ℓ has been made explicit. The first two terms on the right-hand side of this decomposition are available directly whereas the third term can be estimated from the output of the MCMC simulation of the BETEL posterior distribution. For example, in the context of the one block MCMC algorithm given in Section 2.1, from Chib and Jeliazkov (2001), we have that

$$\pi(\tilde{\boldsymbol{\psi}}^\ell | \boldsymbol{x}_{1:n}, M_\ell) = \frac{\mathbf{E}_1 \left\{ \alpha \left(\boldsymbol{\psi}^\ell, \tilde{\boldsymbol{\psi}}^\ell | \boldsymbol{x}_{1:n}, M_\ell \right) q(\tilde{\boldsymbol{\psi}}^\ell) \boldsymbol{x}_{1:n}, M_\ell) \right\}}{\mathbf{E}_2 \left\{ \alpha(\tilde{\boldsymbol{\psi}}^\ell, \boldsymbol{\psi}^\ell | \boldsymbol{x}_{1:n}, M_\ell) \right\}}$$

where $\mathbf{E}_1$ is the expectation with respect to $\pi(\boldsymbol{\psi}^\ell | \boldsymbol{x}_{1:n}, M_\ell)$ and $\mathbf{E}_2$ is the expectation with respect to $q(\boldsymbol{\psi}^\ell | \boldsymbol{x}_{1:n}, M_\ell)$. These expectations can be easily approximated by simulations.

3.3 Model selection consistency results

In this section we establish the consistency of our marginal likelihood based selection procedure for the following exhaustive cases: the case where the models in the comparison set contain only valid moment restrictions, the case where all the models in the set are misspecified, and finally the case where some of the models contain only valid moment restrictions while the others contain at least one invalid moment restriction. Our proofs of consistency are based on: (I) the results of the BvM theorems for correctly and misspecified models stated in Sections 2.2 and 2.3, and (II) the analysis of the asymptotic behavior of the ETEL function under correct and misspecification which we develop in the online Appendix (see Lemmas D.1 and D.3).

The first theorem states that, if the active moment restrictions are all valid, then the

marginal likelihood selects the model that contains the maximum number of overidentifying conditions, that is, the model with the maximum number of active moment restrictions and the smallest number of parameters of interest. This means that the marginal likelihood-based selection procedure enforces parsimony.

For a model M_ℓ, the dimension of the parameter of interest $\boldsymbol{\theta}^\ell$ to be estimated is p_ℓ while the number of active moment restrictions (included in the model for the estimation of $\boldsymbol{\theta}^\ell$) is $(d - d_{v_\ell})$. Consider two generic models M_1 and M_2. Then, $d_{v_2} < d_{v_1}$ means that model M_2 contains more active restrictions than model M_1, and $p_2 < p_1$ means that model M_1 contains more parameters of interest to be estimated than M_2.

Theorem 3.1. *Let Assumption 2, Assumptions 5, 6 in the online Appendix and (2.14) hold, and consider $J < \infty$ different models M_ℓ, $\ell = 1, \ldots, J$, that satisfy Assumption 1, that is, they are all correctly specified. Then,*

$$\lim_{n\to\infty} P\left(\max_{\ell\neq j} \log m(\boldsymbol{x}_{1:n}; M_\ell) < \log m(\boldsymbol{x}_{1:n}; M_j)\right) = 1$$

if and only if $p_j + d_{v_j} < p_\ell + d_{v_\ell}$, $\forall \ell \neq j$.

The result of the theorem implies that, with probability approaching 1, the Bayes factor $B_{j\ell} := m(\boldsymbol{x}_{1:n}; M_j)/m(\boldsymbol{x}_{1:n}; M_\ell)$ is larger than 1 for every $\ell \neq j$. The result in the theorem is an equivalence result saying that, if we compare models that contain only valid moment restrictions, then the marginal likelihood selects a model M_j if and only if M_j contains the maximum number of overidentifying conditions among all the compared models. An illustration of this theorem is provided in Example 3 in the online Appendix.

Next, we consider the case where all models are wrong in the sense of Definition 2.1 and establish a major result of enormous practical significance. The result states that if we compare J misspecified models, then the marginal likelihood-based selection procedure selects the model with the smallest KL divergence $K(P||Q^*(\boldsymbol{\psi}^\ell))$ between P and $Q^*(\boldsymbol{\psi}^\ell)$, where $Q^*(\boldsymbol{\psi}^\ell)$ is such that $K(Q^*(\boldsymbol{\psi}^\ell)||P) = \inf_{Q\in\mathcal{P}_{\boldsymbol{\psi}^\ell}} K(Q||P)$ and $dQ^*(\boldsymbol{\psi}^\ell)/dP = e^{\boldsymbol{\lambda}_\circ(\boldsymbol{\psi})'\boldsymbol{g}^A(\boldsymbol{X},\boldsymbol{\psi})}/\mathbf{E}^P\left[e^{\boldsymbol{\lambda}_\circ(\boldsymbol{\psi})'\boldsymbol{g}^A(\boldsymbol{X},\boldsymbol{\psi})}\right]$ by the dual theorem, as defined in Section 2.3. Because the I-projection $Q^*(\boldsymbol{\psi}^\ell)$ on $\mathcal{P}_{\boldsymbol{\psi}^\ell}$ is unique (Csiszar (1975)), which $Q^*(\boldsymbol{\psi}^\ell)$ is closer to P (in terms of $K(P||Q^*(\boldsymbol{\psi}^\ell))$) depends only on the "amount of misspecification" contained in each model $\mathcal{P}_{\boldsymbol{\psi}^\ell}$.

Theorem 3.2. *Let Assumptions 2 – 4, Assumptions 5 – 8 in the online Appendix and (2.17) be satisfied. Let us consider the comparison of $J < \infty$ models M_j, $j = 1, \ldots, J$ that all use misspecified moments, that is, M_j does not satisfy Assumption 1, $\forall j$. Then,*

$$\lim_{n\to\infty} P\left(\log m(\boldsymbol{x}_{1:n}; M_j) > \max_{\ell\neq j} \log m(\boldsymbol{x}_{1:n}; M_\ell)\right) = 1$$

if and only if $K(P||Q^*(\boldsymbol{\psi}^j)) < \min_{\ell\neq j} K(P||Q^*(\boldsymbol{\psi}^\ell))$, where $K(P||Q) := \int \log(dP/dQ)dP$.

Similarly as in Theorem 3.1, Theorem 3.2 establishes the equivalence result that, if we compare models that all use misspecified moments, then the marginal likelihood selects a model M_j if and only if M_j has the smallest Kullback-Leibler divergence $K(P||Q^*(\boldsymbol{\psi}^j))$ between the true data distribution P and $Q^*(\boldsymbol{\psi}^j)$. Remark that the condition $K(P||Q^*(\boldsymbol{\psi}^j)) < K(P||Q^*(\boldsymbol{\psi}^\ell))$, $\forall \ell \neq j$, given in the theorem does not depend on a particular value of $\boldsymbol{\psi}^j$ and $\boldsymbol{\psi}^\ell$. Indeed, the result of the theorem hinges on the fact that the marginal likelihood selects the model with the $Q^*(\boldsymbol{\psi}^j)$ closer to P, that is, the model that contains the "less misspecified" moment restrictions for every value of $\boldsymbol{\psi}^j$.

The result of the theorem also applies to the case where we compare a correctly specified model M_1 to misspecified models. Indeed, if model M_1 is correctly specified then $K(P||Q^*(\boldsymbol{\psi}^1)) = 0$ while if model M_j is misspecified then $K(P||Q^*(\boldsymbol{\psi}^j)) > 0$.

Example 4 (Model selection when all models are misspecified)**.** *For $i = 1, \ldots, n$, let $y_i = \alpha + \beta z_i + e_i$. Here, we generate $z_i \sim \mathcal{N}(0.5, 1)$ and e_i from the skewed distribution in (2.13) with mean zero and variance 1.625, independently of z_i. Let $\boldsymbol{\theta} := (\alpha, \beta)'$, $e_i(\boldsymbol{\theta}) := (y_i - \alpha - \beta z_i)$ and the true value of $\boldsymbol{\theta}$ be $(0, 1)'$. We compare the following models. Model 4: $\mathbf{E}^P[(e_i(\boldsymbol{\theta}), e_i(\boldsymbol{\theta})z_i, e_i(\boldsymbol{\theta})^3, e_i(\boldsymbol{\theta})^2 - 2)'] = 0$, Model 5: $\mathbf{E}^P[(e_i(\boldsymbol{\theta}), e_i(\boldsymbol{\theta})z_i, e_i(\boldsymbol{\theta})^2 - 2)'] = 0$ and Model 6: $\mathbf{E}^P[e_i(\boldsymbol{\theta}), e_i(\boldsymbol{\theta})^2 - 2] = 0$ which, written in terms of an encompassing grand model, become respectively:*

$$\begin{aligned}
&M_4 : \mathbf{E}^P[e_i(\boldsymbol{\theta})] = 0, \quad \mathbf{E}^P[e_i(\boldsymbol{\theta})z_i] = 0, \quad \mathbf{E}^P[(e_i(\boldsymbol{\theta}))^3] = 0, \quad \mathbf{E}^P[(e_i(\boldsymbol{\theta}))^2 - 2] = 0 \\
&M_5 : \mathbf{E}^P[e_i(\boldsymbol{\theta})] = 0, \quad \mathbf{E}^P[e_i(\boldsymbol{\theta})z_i] = 0, \quad \mathbf{E}^P[(e_i(\boldsymbol{\theta}))^3] = v_1, \quad \mathbf{E}^P[(e_i(\boldsymbol{\theta}))^2 - 2] = 0 \\
&M_6 : \mathbf{E}^P[e_i(\boldsymbol{\theta})] = 0, \quad \mathbf{E}^P[e_i(\boldsymbol{\theta})z_i] = v_2, \quad \mathbf{E}^P[(e_i(\boldsymbol{\theta}))^3] = v_1, \quad \mathbf{E}^P[(e_i(\boldsymbol{\theta}))^2 - 2] = 0
\end{aligned} \tag{3.6}$$

with $\boldsymbol{\psi}^4 = \boldsymbol{\theta}$, $\boldsymbol{\psi}^5 = (\boldsymbol{\theta}, v_1)'$ and $\boldsymbol{\psi}^6 = (\boldsymbol{\theta}, v_1, v_2)'$. Thus, compared to Example 3 in the online Appendix, here we change the moment restriction that involves the variance of e_i. When

the underlying distribution has variance different from 2, all models M_4, M_5, and M_6 are misspecified due to the new moment restriction: $\mathbf{E}^P[(e_i(\boldsymbol{\theta}))^2 - 2] = 0$. *In Table 2, we report the percentage of times the marginal likelihood selects each model out of 500 trials, by sample size, under the default and training sample prior (based on 50 prior observations).*

Because we know the true data generating process, we can compute, for each model, the KL divergence between the true model P and $Q^*(\boldsymbol{\psi}_\circ^j)$ at the pseudo-true parameter $\boldsymbol{\psi}_\circ^j$ for model M_j based on (2.16). Using $10,000,000$ *simulated draws from P, our calculations show that $K(P||Q^*(\boldsymbol{\psi}_\circ^j))$ is equal to* 0.0283 *for M_4,* 0.0096 *for M_5, and* 1.4901×10^{-13} *for M_6. Intuitively, M_6 is the closest to the true model since it imposes fewer restrictions (only two moment restrictions are active). This means that the set of probability distributions that satisfy M_6 is larger than (and contains) the sets of probabilities that conform with M_4 and M_5. This flexibility ensures that the divergence between the set of probabilities that satisfy M_6 and P (as measured by the KL) will be at least as small as for M_4 and M_5. As the empirical results show, under each prior, the best model M_6 picked out by our marginal likelihood ranking is also the model that is the closest to the true model, consistent with the prediction of our theory.*

	Default prior			Training sample prior		
Model	M_4	M_5	M_6	M_4	M_5	M_6
$n = 250$	2.6	56.8	40.5	3.0	62.0	35.0
$n = 500$	0.2	28.1	71.6	1.0	31.0	68.0
$n = 1000$	0.0	4.2	95.8	0.0	4.0	96.0
$n = 2000$	0.0	0.0	100.0	0.0	0.0	100.0

Table 2: Model selection when all models are misspecified. Frequency (%) of times each of the three models in Example 4 are selected by the marginal likelihood criterion in 500 trials, by sample size, for two different prior distributions.

Finally, suppose that some of the models that we consider are correctly specified and others are misspecified in the sense of Definition 2.1. This means that, for the latter, one or more of the active moment restrictions are invalid, or in other words, that one or more components of $\boldsymbol{V}$ are incorrectly set equal to zero. Indeed, all the models for which the active moment restrictions are valid are not misspecified even if some invalid moment restrictions are included among the inactive moment restrictions. This is because there always exists a

value $\boldsymbol{v} \in \mathbb{R}^{d_{v_\ell}}$ that equates the invalid moment restriction. In this case, the true $\boldsymbol{v}_*$ for this model will be different from the zero vector: $\boldsymbol{v}_* \neq \mathbf{0}$ and the true value of the corresponding tilting parameter $\boldsymbol{\lambda}$ will be zero.

For this situation, Theorems 3.1 and 3.2 together imply an interesting corollary: the marginal likelihood selects the correctly specified model that contains the maximum number of overidentifying moment conditions. Without loss of generality, denote this model by M_1. Then we have the following result.

Corollary 3.1. *Let Assumptions 2 – 4, and Assumptions 5 – 8 in the online Appendix hold, and let either (2.14) or (2.17) be satisfied, depending on the model. Let us consider the comparison of J different models M_j, $j = 1, 2, \ldots, J$ where M_1 satisfies Assumption 1 whereas M_j, $j \neq 1$ can either satisfy Assumption 1 or not. Then,*

$$\lim_{n\to\infty} P\left(\log m(\boldsymbol{x}_{1:n}; M_1) > \max_{j\neq 1} \log m(\boldsymbol{x}_{1:n}; M_j)\right) = 1$$

if and only if $(p_1 + d_{v_1}) < (p_j + d_{v_j})$, $\forall j \neq 1$ such that M_j satisfies Assumption 1.

This corollary says that, if we compare a set of models, some of them are correctly specified and the others are misspecified, then the marginal likelihood selects model M_1 if and only if M_1 is correctly specified and contains the maximum number of overidentifying moment conditions among the correctly specified models.

4 Applications

The techniques discussed in the previous sections have wide-ranging applications to various statistical settings, such as generalized linear models, and to many different fields, such as biostatistics and economics. In fact, the methods discussed above can be applied to virtually any problem that, in the frequentist setting, would be approached by generalized method of moments or estimating equation techniques. To illustrate some of the possibilities, we consider in this section two important problems: one in the context of count regression, and the second in the setting of instrumental variable (IV) regression.

4.1 Count regression: variable selection

Suppose that y_i, $i = 1, \ldots, n$ arise from the negative binomial (NB) regression model

$$\begin{aligned} y_i|\boldsymbol{\beta}, \boldsymbol{x}_i &\sim NB\left(\frac{p}{1-p}\mu_i,\ p\right), \qquad \mu_i > 0, \quad p \in (0,1) \\ \log(\mu_i) &= \boldsymbol{x}_i'\boldsymbol{\beta} \end{aligned} \tag{4.1}$$

where μ_i is the size parameter, $\boldsymbol{x}_i = (x_{1,i}, x_{2,i}, x_{3,i})'$, and $\boldsymbol{\beta} = (\beta_1 = 1, \beta_2 = 1, \beta_3 = 0)$. Thus, $x_{3,i}$ is a redundant regressor. Each explanatory variable $x_{j,i}$ is generated *i.i.d.* from a $\mathcal{N}(.4, 1/9)$ distribution and p is set equal to $1/2$. In this setting, suppose we wish to learn about $\boldsymbol{\beta}$ under the moment conditions

$$\begin{aligned} \mathbf{E}^P\left[\left(y_i - \exp(\beta_1 x_{1,i} + \beta_2 x_{2,i} + \beta_3 x_{3,i})\right)\boldsymbol{x}_i\right] &= \mathbf{0} \\ \mathbf{E}^P\left[\left(\frac{y_i - \exp(\boldsymbol{x}_i'\boldsymbol{\beta})}{\sqrt{\exp(\boldsymbol{x}_i'\beta)}}\right)^2 - 1\right] &= v. \end{aligned} \tag{4.2}$$

The first type of moment restriction (one for each $x_{j,i}$ for $j = 1, 2, 3$) is derived from the fact that the conditional expectation of y_i is $\exp(\boldsymbol{x}_i'\boldsymbol{\beta})$ and this identifies $\boldsymbol{\beta}$. The second type of restriction is suggested by a Poisson model (which is misspecified when the data arise from a NB). More specifically, if $v = 0$ that moment condition asserts that the conditional variance of y_i is equal to the conditional mean.

Suppose that we are interested in determining if x_3 is a redundant regressor and if the conditional mean and variance are equal. To solve this problem, we can create the following four models based on the grand model (4.2) with the following restrictions:

$$\begin{aligned} &M_1 : \beta_1 \text{ and } \beta_2 \text{ are free parameters, } \beta_3 = 0 \text{ and } v = 0. \\ &M_2 : \beta_1,\ \beta_2,\ \beta_3 \text{ are free parameters and } v = 0. \\ &M_3 : \beta_1,\ \beta_2 \text{ and } v \text{ are free parameters, and } \beta_3 = 0. \\ &M_4 : \beta_1\ \beta_2,\ \beta_3 \text{ and } v \text{ are free parameters.} \end{aligned} \tag{4.3}$$

As required, each model has the same moment restrictions. The different models arise from the different restrictions on β_3 and v. In this set-up, models M_3 and M_4 are the correctly specified models but M_3 has more overidentifying moment restrictions than M_4. We conduct

our MCMC analysis and compute the marginal likelihoods of the four models by the Chib (1995) method under the default student-t prior distribution on $\boldsymbol{\beta}$ given in (2.11). The results are given in Table 3. The results show that the frequency of selecting M_3 is 94% for $n = 250$ and this percentage increases with sample size, in accordance with our theory. In addition, neither model M_1 nor M_2 (which state equality of the conditional mean and variance) is picked for any sample size.

Model	M_1	M_2	M_3	M_4
$n = 250$	0.00	0.00	0.94	0.06
$n = 500$	0.00	0.00	0.95	0.05
$n = 1000$	0.00	0.00	0.96	0.04

Table 3: Frequency (%) of times each of the four models in (4.3) are selected by the marginal likelihood criterion in 500 trials. Model choice with Negative Binomial DGP. Model M_3, defined by $\beta_3 = 0$ and v free, is the true model. The other models are defined in the text.

In the online Appendix A we report a similar analysis where the data are generated from a Poisson model. We emphasize that our analysis of these data, and the comparison across models, was light in terms of assumptions. The Poisson and negative binomial distributions are used to simply obtain a sample. These distributional forms are not featured in the estimation or the model comparison. A reader of this paper wondered how a parametric Poisson model would have performed for these data. Since the data was generated under either a Poisson model or a model close to a Poisson model, the marginal likelihood of the Poisson model (correctly) is higher than that of the moment model. But this performance suffers dramatically if the data are generated from a count process that is quite different from the Poisson. For instance, suppose that the data are generated under the assumption that the first three moment conditions hold. We have developed a way of generating such a sample which works as follows. We first generate a large population of count data from an arbitrary count process (say $y_i = \lfloor \exp\{\beta_1 x_{1,i} + \beta_2 x_{2,i} + 20\mathcal{N}(0,1)\} \rfloor$, setting any negative observations to zero and where $\lfloor a \rfloor$ denotes the largest integer less than or equal to a). We then find the ETEL probabilities p_i^* consistent with the given moment conditions. Finally, we sample the population of observations according to these probabilities. The resulting sample satisfies the moment conditions but has no connection to the Poisson or negative binomial distributional forms. For such a design, in 500 replications, in the parametric Poisson models (one with $\beta_3 = 0$ and one with β_3 free), the Poisson model with $\beta_3 = 0$ is selected 42% when

$n = 250$, 46% when $n = 500$, and 45% when $n = 1000$. Thus, the Poisson assumption is not capable of selecting the correct case. In addition, when these two Poisson models are compared along with the 4 moment models in (4.3), for which the marginal likelihoods are computed with our method, M_3 is decisively preferred over the Poisson models in terms of marginal likelihood and the frequency of times it is selected is similar to that reported above in Table 3.

4.2 IV Regression

Consider now the commonly occurring situation with observational data where one is interested in learning about the causal effect parameter β in the model

$$y = \alpha + x\beta + w\delta + \varepsilon, \qquad \mathbf{E}^P[\varepsilon] = 0$$

but the covariate x is correlated with the error, due to say unmeasured or uncontrolled factors, apart from w, that are correlated with x and that reside in ε. Also suppose that one has two valid instrumental variables z_1 and z_2 that (by definition) are correlated with x but uncorrelated with ε. In this setting we can learn about $\boldsymbol{\theta} := (\alpha, \beta, \delta)$ from the overidentified moment restrictions

$$\mathbf{E}^P\left[(y_i - \alpha - x_i\beta - w_i\delta)\right] = 0 \tag{4.4}$$

$$\mathbf{E}^P\left[(y_i - \alpha - x_i\beta - w_i\delta)\, z_{1i}\right] = 0 \tag{4.5}$$

$$\mathbf{E}^P\left[(y_i - \alpha - x_i\beta - w_i\delta)\, z_{2i}\right] = 0 \tag{4.6}$$

$$\mathbf{E}^P\left[(y_i - \alpha - x_i\beta - w_i\delta)\, w_i\right] = 0, \qquad (i \leq n) \tag{4.7}$$

without having to model the distribution of ε or the model connecting z to x.

In order to demonstrate the performance of our Bayesian prior-posterior analysis in this setting, we generate data on $(y_i, x_i, z_{1i}, z_{2i})$ from a design that incorporates a skewed marginal distribution of ε and substantial correlation between x and ε. In our DGP we assume that $y = 1 + .5x + .7w + \varepsilon$, $x = z_1 + z_2 + w + u$, and generate z_j from $\mathcal{N}(.5, 1)$ and w from $Uniform(0, 1)$. The errors (ε, u) are generated from a Gaussian copula whose covariance matrix has 1 on the diagonal, and .7 on the off-diagonal, such that the ε marginal distribution is the skewed

bivariate mixture $0.5\mathcal{N}(.5, .5^2)+0.5\mathcal{N}(-.5, 1.118^2)$ and the u marginal distribution is $\mathcal{N}(0, 1)$. We generate $n = 250$ and $n = 2000$ observations from this design and use moment conditions (4.4)-(4.7) and our default student-t prior given in (2.11) to learn about $\boldsymbol{\theta}$. The results shown in Table 4 and Figure 1 demonstrate clearly the ability of our method to concentrate on the true values of the parameters, under minimal assumptions.

	mean	sd	median	lower	upper	ineff
			$n = 250$			
α	1.26	0.11	1.26	1.04	1.48	1.26
β	0.55	0.03	0.55	0.48	0.61	1.41
δ	0.28	0.17	0.28	-0.06	0.60	1.27
			$n = 2000$			
α	1.01	0.05	1.01	0.92	1.10	1.30
β	0.51	0.02	0.51	0.48	0.54	1.33
δ	0.66	0.07	0.66	0.52	0.81	1.38

Table 4: Posterior summary for two simulated sample sizes from IV regression model with skewed error. The true value of α is 1, of β is .5 and of δ is .7. The summaries are based on $10,000$ MCMC draws beyond a burn-in of 1000. The M-H acceptance rate is around 90% in both cases.

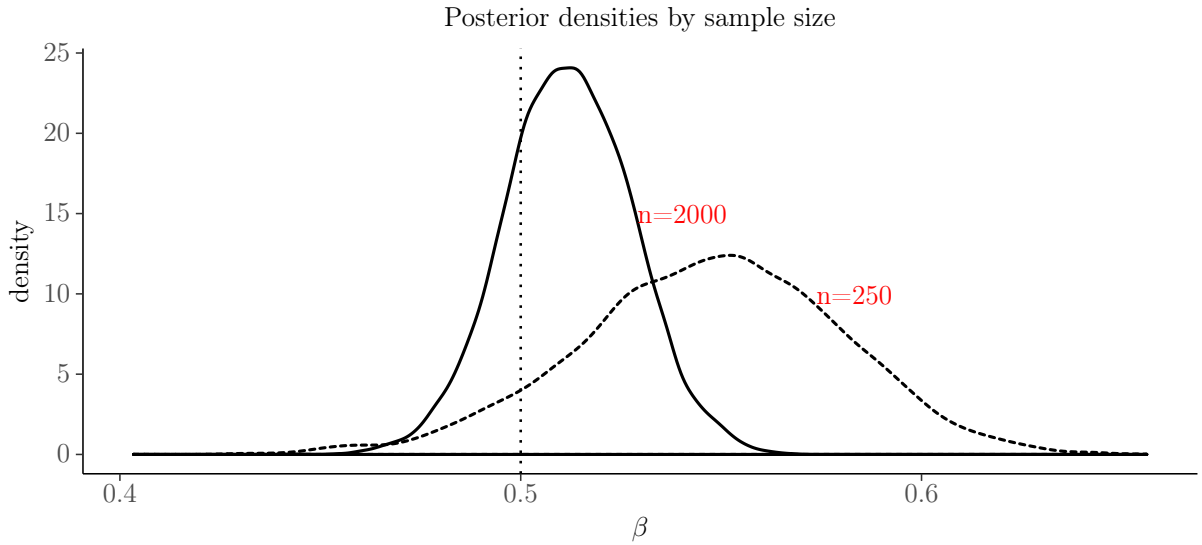

Figure 1: Posterior densities of β in the IV regression with skewed error. Posterior densities are based on 10,000 draws beyond a burn-in of 1000. The M-H acceptance rate is about 90% for each sample size.

Now suppose that we are unsure that z_2 is an appropriate instrument. We can address this concern by estimating a new model M_2 in which the moment condition (4.6) is not

active. The marginal likelihood of this model can be compared with the marginal likelihood of the previous model M_1. The results show that for $n = 250$, the log-marginal likelihood of M_1 is -1395.807 and that of M_2 is -1398.092, while for $n = 2000$, the corresponding log-marginal likelihoods are -15217.78 and -15222.65, respectively, thus correctly indicating for both sample sizes that z_2 is an appropriate instrument.

5 Conclusion

In this paper we have developed a fully Bayesian framework for estimation and model comparisons in statistical models that are defined by moment restrictions. The Bayesian analysis of such models has always been viewed as a challenge because traditional Bayesian semiparametric methods, such as those based on Dirichlet process mixtures and variants thereof, are not suitable for such models. What we have shown in this paper is that the Exponentially Tilted Empirical Likelihood setting is an immensely useful organizing framework within which a fully Bayesian treatment of such models can be developed. We have established a number of new, powerful results surrounding the Bayesian ETEL framework including the treatment of models that are possibly misspecified. We show how the moment conditions can be reexpressed in terms of additional nuisance parameters and that the Bayesian ETEL posterior distribution satisfies a Bernstein-von Mises theorem. We have also developed a framework for comparing moment condition models based on marginal likelihoods and Bayes factors and provided a suitable large sample theory for model selection consistency. Our results show that the marginal likelihood favors the model with the minimum number of parameters and the maximum number of valid moment restrictions. When the models are misspecified, the marginal likelihood-based selection procedure selects the model that is closer to the (unknown) true data generating process in terms of the Kullback-Leibler divergence. The ideas and results illumined in this paper now provide the means for analyzing a whole array of models from the Bayesian viewpoint. This broadening of the scope of Bayesian techniques to previously intractable problems is likely to have far-reaching practical consequences.

Supplementary Material: the supplementary material in the online Appendix contains further examples, assumptions and the technical proofs of the results in the paper.

Appendix

In this Appendix we provide the proof of Theorems 3.1 and 3.2. The proofs of all the other results are in the online Appendix.

Notation: Let $\widehat{\boldsymbol{\lambda}}(\boldsymbol{\psi}) := \arg\min_{\boldsymbol{\lambda}\in\mathbb{R}^d} \frac{1}{n}\sum_{i=1}^n \left[\exp\{\boldsymbol{\lambda}'\boldsymbol{g}^A(\boldsymbol{x}_i,\boldsymbol{\psi})\}\right]$. We recall the Schennach (2007) ETEL estimator of $\boldsymbol{\psi}$, denoted by $\widehat{\boldsymbol{\psi}} := (\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{v}})$:

$$\widehat{\boldsymbol{\psi}} := \arg\max_{\boldsymbol{\psi}\in\Psi} \frac{1}{n}\sum_{i=1}^n \left[\widehat{\boldsymbol{\lambda}}(\boldsymbol{\psi})'\boldsymbol{g}^A(\boldsymbol{x}_i,\boldsymbol{\psi}) - \log\frac{1}{n}\sum_{j=1}^n \exp\{\widehat{\boldsymbol{\lambda}}(\boldsymbol{\psi})'\boldsymbol{g}^A(\boldsymbol{x}_j,\boldsymbol{\psi})\}\right]. \tag{C.1}$$

The following notation is used hereafter. The ETEL estimator of $\boldsymbol{\psi}^\ell$ in model M_ℓ is:

$$\widehat{\boldsymbol{\psi}}^\ell := \arg\max_{\boldsymbol{\psi}^\ell\in\Psi^\ell} \frac{1}{n}\sum_{i=1}^n \left[\widehat{\boldsymbol{\lambda}}(\boldsymbol{\psi}^\ell)'\boldsymbol{g}^A(\boldsymbol{x}_i,\boldsymbol{\psi}^\ell) - \log\frac{1}{n}\sum_{j=1}^n \exp\{\widehat{\boldsymbol{\lambda}}(\boldsymbol{\psi}^\ell)'\boldsymbol{g}^A(\boldsymbol{x}_j,\boldsymbol{\psi}^\ell)\}\right] \tag{D.1}$$

where $\widehat{\boldsymbol{\lambda}}(\boldsymbol{\psi}^\ell) = \arg\min_{\boldsymbol{\lambda}\in\mathbb{R}^d} \frac{1}{n}\sum_{i=1}^n \left[\exp\{\boldsymbol{\lambda}'\boldsymbol{g}^A(\boldsymbol{x}_i,\boldsymbol{\psi}^\ell)\}\right]$. Denote $\widehat{\boldsymbol{g}}^A(\boldsymbol{\psi}^\ell) := \frac{1}{n}\sum_{i=1}^n \boldsymbol{g}^A(\boldsymbol{x}_i,\boldsymbol{\psi}^\ell)$, $\widehat{\boldsymbol{g}}_\ell^A := \widehat{\boldsymbol{g}}^A(\boldsymbol{\psi}^\ell)$, $\widehat{L}(\boldsymbol{\psi}^\ell) := \exp\{\widehat{\boldsymbol{\lambda}}(\boldsymbol{\psi}^\ell)'\widehat{\boldsymbol{g}}^A(\boldsymbol{\psi}^\ell)\}\left[n^{-1}\sum_{i=1}^n \exp\{\widehat{\boldsymbol{\lambda}}(\boldsymbol{\psi}^\ell)'\boldsymbol{g}^A(\boldsymbol{x}_i,\boldsymbol{\psi}^\ell)\}\right]^{-1}$ and $L(\boldsymbol{\psi}^\ell) = \exp\{\boldsymbol{\lambda}_\circ(\boldsymbol{\psi}^\ell)'\mathbb{E}^P[\boldsymbol{g}^A(\boldsymbol{x},\boldsymbol{\psi}^\ell)]\}\left(\mathbb{E}^P\left[\exp\{\boldsymbol{\lambda}_\circ(\boldsymbol{\psi}^\ell)'\boldsymbol{g}^A(\boldsymbol{x},\boldsymbol{\psi}^\ell)\}\right]\right)^{-1}$. Moreover, we use the notation $\Sigma_\ell = \left(\Gamma_\ell'\Delta_\ell^{-1}\Gamma_\ell\right)^{-1}$ where $\Gamma_\ell := \mathbb{E}^P\left[\frac{\partial}{\partial\boldsymbol{\psi}^{\ell\prime}}\boldsymbol{g}^A(\boldsymbol{X},\boldsymbol{\psi}_*^\ell)\right]$ and $\Delta_\ell := \mathbb{E}^P[\boldsymbol{g}^A(\boldsymbol{X},\boldsymbol{\psi}_*^\ell)\boldsymbol{g}^A(\boldsymbol{X},\boldsymbol{\psi}_*^\ell)']$. In the proofs, we omit measurability issues which can be dealt with in the usual manner by replacing probabilities with outer probabilities.

Proof of Theorem 3.1: By (3.5) and Lemmas D.1 and D.2 in the online Appendix we obtain

$$\begin{aligned} P\left(\max_{\ell\neq j}\log m(\boldsymbol{x}_{1:n};M_\ell) < \log m(\boldsymbol{x}_{1:n};M_j)\right) &= P\Big(\max_{\ell\neq j}\Big[-\frac{n}{2}\widehat{\boldsymbol{g}}_\ell^{A\prime}\Delta^{-1}\widehat{\boldsymbol{g}}_\ell^A + \log\pi(\widehat{\boldsymbol{\psi}}^\ell|M_\ell) \\ &\quad - \frac{(p_\ell+d_{v_\ell})}{2}(\log n - \log(2\pi)) + \frac{1}{2}\log|\Sigma_\ell|\Big] + \frac{n}{2}\widehat{\boldsymbol{g}}_j^{A\prime}\Delta^{-1}\widehat{\boldsymbol{g}}_j^A + o_p(1) \\ &\quad < \log\pi(\widehat{\boldsymbol{\psi}}^j|M_j) - \frac{(p_j+d_{v_j})}{2}(\log n - \log(2\pi)) + \frac{1}{2}\log|\Sigma_j|\Big). \end{aligned} \tag{D.2}$$

Remark that $n\widehat{\boldsymbol{g}}_j^{A\prime}\Delta^{-1}\widehat{\boldsymbol{g}}_j^A \xrightarrow{d} \chi^2_{d-(p_j+d_{v_j})}$, $\forall j$, so that $n\widehat{\boldsymbol{g}}_j^{A\prime}\Delta^{-1}\widehat{\boldsymbol{g}}_j^A = O_p(1)$. Suppose first that

$(p_\ell + d_{v_\ell} > p_j + d_{v_j})$, $\forall \ell \neq j$. Since $-n\widehat{\boldsymbol{g}}_\ell^{A'}\Delta^{-1}\widehat{\boldsymbol{g}}_\ell^{A} < 0$ for every ℓ, we lower bound (D.2) as

$$
\begin{aligned}
P\left(\max_{\ell\neq j}\log m(\boldsymbol{x}_{1:n};M_\ell) < \log m(\boldsymbol{x}_{1:n};M_j)\right) &\geq P\Big(\frac{n}{2}\widehat{\boldsymbol{g}}_j^{A'}\Delta^{-1}\widehat{\boldsymbol{g}}_j^{A} + o_p(1) \\
&< \log n\Big[\frac{\min_{\ell\neq j}(p_\ell + d_{v_\ell}) - p_j - d_{v_j}}{2} - \frac{\min_{\ell\neq j}(p_\ell + d_{v_\ell}) - p_j - d_{v_j})}{2\log n}\log(2\pi) \\
&\quad - \frac{\log[\max_{\ell\neq j}\pi(\widehat{\boldsymbol{\psi}}^{\ell}|M_\ell)/\pi(\widehat{\boldsymbol{\psi}}^{j}|M_j)]}{\log n} - \frac{1}{2\log n}\left(\max_{\ell\neq j}\log|\Sigma_\ell| - \log|\Sigma_j|\right)\Big]\Big) \\
&= P\Big(\underbrace{\frac{n}{2}\widehat{\boldsymbol{g}}_j^{A'}\Delta^{-1}\widehat{\boldsymbol{g}}_j^{A} + o_p(1)}_{=:\mathcal{I}_n} < \underbrace{\log n\Big[\frac{\min_{\ell\neq j}(p_\ell + d_{v_\ell}) - p_j - d_{v_j}}{2} + \mathcal{O}_p((\log n)^{-1})\Big]}_{=:\mathcal{II}_n}\Big). \quad \text{(D.3)}
\end{aligned}
$$

Because $\mathcal{I}_n = \mathcal{O}_p(1)$ (and is asymptotically positive) and $\mathcal{II}_n$ is strictly positive as $n \to \infty$ (since $(p_\ell + d_{v_\ell}) > (p_j + d_{v_j})$, $\forall \ell \neq j$) and converges to $+\infty$, then the probability converges to 1. This proves one direction of the statement.

To prove the second direction of the statement, suppose that $\lim_{n\to\infty} P(\max_{\ell\neq j}\log m(\boldsymbol{x}_{1:n};M_\ell) < \log m(\boldsymbol{x}_{1:n};M_j)) = 1$ and consider the following upper bound (which follows from (D.2) and the fact that $n\widehat{\boldsymbol{g}}_j^{A'}\Delta^{-1}\widehat{\boldsymbol{g}}_j^{A} > 0$, $\forall n$):

$$
\begin{aligned}
&P\left(\max_{\ell\neq j}\log m(\boldsymbol{x}_{1:n};M_\ell) < \log m(\boldsymbol{x}_{1:n};M_j)\right) \\
&\qquad\qquad \leq P\left(\log m(\boldsymbol{x}_{1:n};M_\ell) < \log m(\boldsymbol{x}_{1:n};M_j)\right), \qquad \forall \ell \neq j \\
&\leq P\Big(-\frac{n}{2}\widehat{\boldsymbol{g}}_\ell^{A'}\Delta^{-1}\widehat{\boldsymbol{g}}_\ell^{A} + o_p(1) + \log n\Big[\frac{(p_j + d_{v_j}) - (p_\ell + d_{v_\ell})}{2} + \mathcal{O}_p\left(\frac{1}{\log n}\right)\Big] < 0\Big), \quad \forall \ell \neq j. \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{(D.4)}
\end{aligned}
$$

Because the probability in the first line of (D.4) converges to 1 as $n \to \infty$ then, necessarily, the probability in the last line of (D.4) converges to 1 which is possible only if $(p_j + d_{v_j}) < (p_\ell + d_{v_\ell})$ because $\log n\left[\frac{(p_j+d_{v_j})-(p_\ell+d_{v_\ell})}{2}\right]$ is the dominating term since $-\frac{n}{2}\widehat{\boldsymbol{g}}_\ell^{A'}\Delta^{-1}\widehat{\boldsymbol{g}}_\ell^{A} < 0$ and it remains bounded as $n \to \infty$. Since the first inequality in (D.4) holds $\forall \ell \neq j$ then convergence to 1 of the probability in the last line of (D.4) is possible only if $(p_j + d_{v_j}) < (p_\ell + d_{v_\ell})$, $\forall \ell \neq j$.

□

Proof of Theorem 3.2: We can write $\log p(\boldsymbol{x}_{1:n}|\boldsymbol{\psi}^\ell; M_\ell) = -n\log n + n\log\widehat{L}(\boldsymbol{\psi}^\ell)$.

Then, we have:

$$P\left(\log m(\boldsymbol{x}_{1:n};M_j) > \max_{\ell\neq j}\log m(\boldsymbol{x}_{1:n};M_\ell)\right) = P\Big(n\log\widehat{L}(\boldsymbol{\psi}_\circ^j)+\log\pi(\boldsymbol{\psi}_\circ^j|M_j)-\log\pi(\boldsymbol{\psi}_\circ^j|\boldsymbol{x}_{1:n},M_j) > \max_{\ell\neq j}[n\log\widehat{L}(\boldsymbol{\psi}_\circ^\ell)+\log\pi(\boldsymbol{\psi}_\circ^\ell|M_\ell)-\log\pi(\boldsymbol{\psi}_\circ^\ell|\boldsymbol{x}_{1:n},M_\ell)]\Big)$$
$$= P\Big(n\log L(\boldsymbol{\psi}_\circ^j)+n\log\frac{\widehat{L}(\boldsymbol{\psi}_\circ^j)}{L(\boldsymbol{\psi}_\circ^j)}+\mathcal{B}_j > \max_{\ell\neq j}\Big[n\log L(\boldsymbol{\psi}_\circ^\ell)+\mathcal{B}_\ell+n\log\frac{\widehat{L}(\boldsymbol{\psi}_\circ^\ell)}{L(\boldsymbol{\psi}_\circ^\ell)}\Big]\Big) \quad \text{(D.5)}$$

where $\forall\ell$, $\mathcal{B}_\ell := \log\pi(\boldsymbol{\psi}_\circ^\ell|M_\ell) - \log\pi(\boldsymbol{\psi}_\circ^\ell|\boldsymbol{x}_{1:n},M_\ell)$ and $\mathcal{B}_\ell = O_p(1)$ under the assumptions of Theorem 2.2. By definition of $dQ^*(\boldsymbol{\psi})$ in Section 2.3 we have that: $\log L(\boldsymbol{\psi}_\circ^\ell) = \mathbb{E}^P[\log dQ^*(\boldsymbol{\psi}_\circ^\ell)/dP] = -\mathbb{E}^P[\log dP/dQ^*(\boldsymbol{\psi}_\circ^\ell)] = -K(P||Q^*(\boldsymbol{\psi}_\circ^\ell))$. Remark that $\mathbb{E}^P[\log(dP/dQ^*(\boldsymbol{\psi}_\circ^2))] > \mathbb{E}^P[\log(dP/dQ^*(\boldsymbol{\psi}_\circ^1))]$ means that the KL divergence between P and $Q^*(\boldsymbol{\psi}_\circ^\ell)$, is smaller for model M_1 than for model M_2, where $Q^*(\boldsymbol{\psi}_\circ^\ell)$ minimizes the KL divergence between $Q\in\mathcal{P}_{\boldsymbol{\psi}_\circ^\ell}$ and P for $\ell\in\{1,2\}$ (notice the inversion of the two probabilities).
First, suppose that $\min_{\ell\neq j}\mathbb{E}^P\left[\log\left(dP/dQ^*(\boldsymbol{\psi}_\circ^\ell)\right)\right] > \mathbb{E}^P\left[\log\left(dP/dQ^*(\boldsymbol{\psi}_\circ^j)\right)\right]$. By (D.5):

$$P\left(\log m(\boldsymbol{x}_{1:n};M_j) > \max_{\ell\neq j}\log m(\boldsymbol{x}_{1:n};M_\ell)\right) \geq$$
$$P\left(\log\frac{\widehat{L}(\boldsymbol{\psi}_\circ^j)}{L(\boldsymbol{\psi}_\circ^j)} - \max_{\ell\neq j}\log\frac{\widehat{L}(\boldsymbol{\psi}_\circ^\ell)}{L(\boldsymbol{\psi}_\circ^\ell)} + \frac{1}{n}(\mathcal{B}_j - \max_{\ell\neq j}\mathcal{B}_\ell) > \underbrace{\max_{\ell\neq j}\log L(\boldsymbol{\psi}_\circ^\ell) - \log L(\boldsymbol{\psi}_\circ^j)}_{=:\mathcal{I}_n}\right). \quad \text{(D.6)}$$

This probability converges to 1 because $\mathcal{I}_n = K(P||Q^*(\boldsymbol{\psi}_\circ^j)) - \min_{\ell\neq j}K(P||Q^*(\boldsymbol{\psi}_\circ^\ell)) < 0$ by assumption, and $\left[\log\widehat{L}(\boldsymbol{\psi}^\ell) - \log L(\boldsymbol{\psi}^\ell)\right] \xrightarrow{p} 0$, for every $\boldsymbol{\psi}^\ell\in\Psi^\ell$ and every $\ell\in\{1,2\}$ by Lemma D.3 in the online Appendix.
To prove the second direction of the statement, suppose that $\lim_{n\to\infty}P(\log m(\boldsymbol{x}_{1:n};M_j) > \max_{\ell\neq j}\log m(\boldsymbol{x}_{1:n};M_\ell)) = 1$. By (D.5) it holds, $\forall\ell\neq j$

$$P\left(\log m(\boldsymbol{x}_{1:n};M_j) > \max_{\ell\neq j}\log m(\boldsymbol{x}_{1:n};M_\ell)\right) \leq$$
$$P\Big(\log\frac{\widehat{L}(\boldsymbol{\psi}_\circ^j)}{L(\boldsymbol{\psi}_\circ^j)} - \log\frac{\widehat{L}(\boldsymbol{\psi}_\circ^\ell)}{L(\boldsymbol{\psi}_\circ^\ell)} + \frac{1}{n}(\mathcal{B}_j-\mathcal{B}_\ell) > \log\frac{L(\boldsymbol{\psi}_\circ^\ell)}{L(\boldsymbol{\psi}_\circ^j)}\Big). \quad \text{(D.7)}$$

Convergence to 1 of the left hand side implies convergence to 1 of the right hand side which is possible only if $\log L(\boldsymbol{\psi}_\circ^\ell) - \log L(\boldsymbol{\psi}_\circ^j) < 0$. Since this is true for every model ℓ, then this

implies that $K(P||Q^*(\boldsymbol{\psi}_\circ^j)) < \min_{\ell \neq j} K(P||Q^*(\boldsymbol{\psi}_\circ^\ell))$ which concludes the proof.

□

References

Bornn, L., Shephard, N. and Solgi, R. (2015), Moment Conditions and Bayesian Nonparametrics, Technical report, arXiv:1507.08645.

Broniatowski, M. and Keziou, A. (2012), 'Divergences and Duality for Estimation and Test Under Moment Condition Models', *Journal of Statistical Planning and Inference* **142**(9), 2554–2573.

Chamberlain, G. (1987), 'Asymptotic Efficiency in Estimation with Conditional Moment Restrictions', *Journal of Econometrics* **34**(3), 305–334.

Chang, I. H. and Mukerjee, R. (2008), 'Bayesian and Frequentist Confidence Intervals Arising from Empirical-type Likelihoods', *Biometrika* **95**(1), 139–147.

Chaudhuri, S. and Ghosh, M. (2011), 'Empirical Likelihood for Small Area Estimation', *Biometrika* **98**(2), 473–480.

Chaudhuri, S., Mondal, D. and Yin, T. (2017), 'Hamiltonian Monte Carlo Sampling in Bayesian Empirical Likelihood Computation', *Journal of the Royal Statistical Society, Series B* **79**(1), 293–320.

Chen, S. X. and Van Keilegom, I. (2009), 'A Review on Empirical Likelihood Methods for Regression', *TEST* **18**(3), 415–447.

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 Jeliazkov, I. (2001), 'Marginal Likelihood From the Metropolis-Hastings Output', *Journal of the American Statistical Association* **96**(453), 270–281.

Chib, S. and Ramamurthy, S. (2010), 'Tailored Randomized Block MCMC methods with Application to DSGE Models', *Journal of Econometrics* **155**(1), 19–38.

Csiszar, I. (1975), 'I-Divergence Geometry of Probability Distributions and Minimization Problems', *Annals of Probability* **3**(1), 146–158.

Fang, K.-T. and Mukerjee, R. (2006), 'Empirical-Type Likelihoods Allowing Posterior Credible Sets with Frequentist Validity: Higher-Order Asymptotics', *Biometrika* **93**(3), 723–733.

Florens, J.-P. and Simoni, A. (2016), Gaussian Processes and Bayesian Moment Estimation, Technical report, arXiv:1607.07343.

Ghosh, J. and Ramamoorthi, R. (2003), *Bayesian Nonparametrics*, Springer-Verlag New York.

Grendar, M. and Judge, G. (2009), 'Asymptotic Equivalence of Empirical Likelihood and Bayesian MAP', *The Annals of Statistics* **37**(5A), 2445–2457.

Hansen, P. (1982), 'Large Sample Properties of Generalized Method of Moments Estimators', *Econometrica* **50**, 1029–1054.

Hong, H. and Preston, B. (2012), 'Bayesian Averaging, Prediction and Nonnested Model Selection', *Journal of Econometrics* **167**(2), 358–369.

Imbens, G. W. (1997), 'One-Step Estimators for Over-Identified Generalized Method of Moments Models', *The Review of Economic Studies* **64**(3), 359–383.

Kim, M.-O. and Yang, Y. (2011), 'Semiparametric Approach to a Random Effects Quantile Regression Model', *Journal of the American Statistical Association* **106**(496), 1405–1417.

Kitamura, Y. and Otsu, T. (2011), Bayesian Analysis of Moment Condition Models Using Nonparametric Priors, Technical report.

Kitamura, Y. and Stutzer, M. (1997), 'An Information-Theoretic Alternative to Generalized Method of Moments Estimation', *Econometrica* **65**(4), pp. 861–874.

Kleijn, B. and van der Vaart, A. (2012), 'The Bernstein-Von-Mises Theorem Under Misspecification', *Electronic Journal Statistics* **6**, 354–381.

Lancaster, T. and Jun, S. J. (2010), 'Bayesian Quantile Regression Methods', *Journal of Applied Econometrics* **25**(2), 287–307.

Lazar, N. A. (2003), 'Bayesian Empirical Likelihood', *Biometrika* **90**(2), 319–326.

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

Owen, A. (1990), 'Empirical Likelihood Ratio Confidence Regions', *Annals of Statistics*

18(1), 90–120.

Owen, A. B. (1988), 'Empirical Likelihood Ratio Confidence Intervals for a Single Functional', *Biometrika* **75**(2), 237–249.

Owen, A. B. (2001), *Empirical Likelihood*, Chapman & Hall/CRC Monographs on Statistics & Applied Probability.

Porter, A. T., Holan, S. H. and Wikle, C. K. (2015), 'Bayesian Semiparametric Hierarchical Empirical Likelihood Spatial Models', *Journal of Statistical Planning and Inference* **165**, 78–90.

Qin, J. and Lawless, J. (1994), 'Empirical Likelihood and General Estimating Equations', *Annals of Statistics* **22**(1), 300–325.

Rao, J. and Wu, C. (2010), 'Bayesian Pseudo-Empirical-Likelihood Intervals for Complex Surveys', *Journal of the Royal Statistical Society Series B* **72**(4), 533–544.

Robert, C. (2007), *The Bayesian Choice: From Decision-Theoretic Foundations to Computational Implementation*, Springer New York.

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.

Sueishi, N. (2013), 'Identification Problem of the Exponential Tilting Estimator under Misspecification', *Economics Letters* **118**(3), 509 – 511.

Variyath, A. M., Chen, J. and Abraham, B. (2010), 'Empirical Likelihood Based Variable Selection', *Journal of Statistical Planning and Inference* **140**(4), 971–981.

Vexler, A., Deng, W. and Wilding, G. E. (2013), 'Nonparametric Bayes Factors Based on Empirical Likelihood Ratios', *Journal of Statistical Planning and Inference* **143**(3), 611–620.

Xi, R., Li, Y. and Hu, Y. (2016), 'Bayesian Quantile Regression Based on the Empirical Likelihood with Spike and Slab Priors', *Bayesian Analysis* **11**(3), 821–855.

Yang, Y. and He, X. (2012), 'Bayesian Empirical Likelihood for Quantile Regression', *The Annals of Statistics* **40**(2), 1102–1131.

Online appendix to:

Bayesian Estimation and Comparison of Moment Condition Models

SIDDHARTHA CHIB* MINCHUL SHIN† ANNA SIMONI‡

Olin Business School *University of Illinois* *CREST, CNRS*

This version: July 16, 2017

A Examples

Example 2 (Misspecified model and pseudo-true value)**.** *Let us consider the model* $y_i = \alpha + e_i$, $i = 1, \ldots, n$, *with* e_i *independently drawn from the skewed distribution* P *given in* (2.13)*. We consider the following two moment conditions* $\mathbf{E}^P[y_i - \alpha] = 0$ *and* $\mathbf{E}^P[(y_i - \alpha)^3] = 0$*. This situation is different from the one illustrated in Example 1 in the paper because there are no covariates and the augmented parameter* v *is (incorrectly) forced to be zero. In turn,* α *has to satisfy both the moment restrictions, which is impossible under* P*. Instead, for each* α *the ETEL function is defined by the probability measure* $Q^*(\alpha)$ *which is the closest to the true generating process* P *in terms of KL divergence among the probability measures that are consistent with the given moment restrictions for a given* α*. In Figure 2 (left panel), we present* $\mathbf{E}^P[\log(dQ^*(\alpha)/dP)]$ *which is equal to* $-K(P||Q^*(\alpha))$*. The value that maximizes this function is different from the true value* ($\alpha = 0$) *and it is peaked around* -0.056*. This value is the pseudo-true value. In the right panel of Figure 2, we present the BETEL posterior distribution of* α *for five different sample sizes. The BETEL posterior distribution shrinks*

*Olin Business School, Washington University in St. Louis, Campus Box 1133, 1 Brookings Dr. St. Louis, MO 63130, USA, e-mail: chib@wustl.edu

†Department of Economics, University of Illinois, 214 David Kinley Hall, 1407 W. Gregory, Urbana, IL 61801, e-mail: mincshin@illinois.edu

‡CREST - ENSAE - École Polytechnique, 5, avenue Henry Le Chatelier, 91120 Palaiseau France, e-mail: simoni.anna@gmail.com

and moves toward the pseudo-true value, in conformity with our theoretical result.

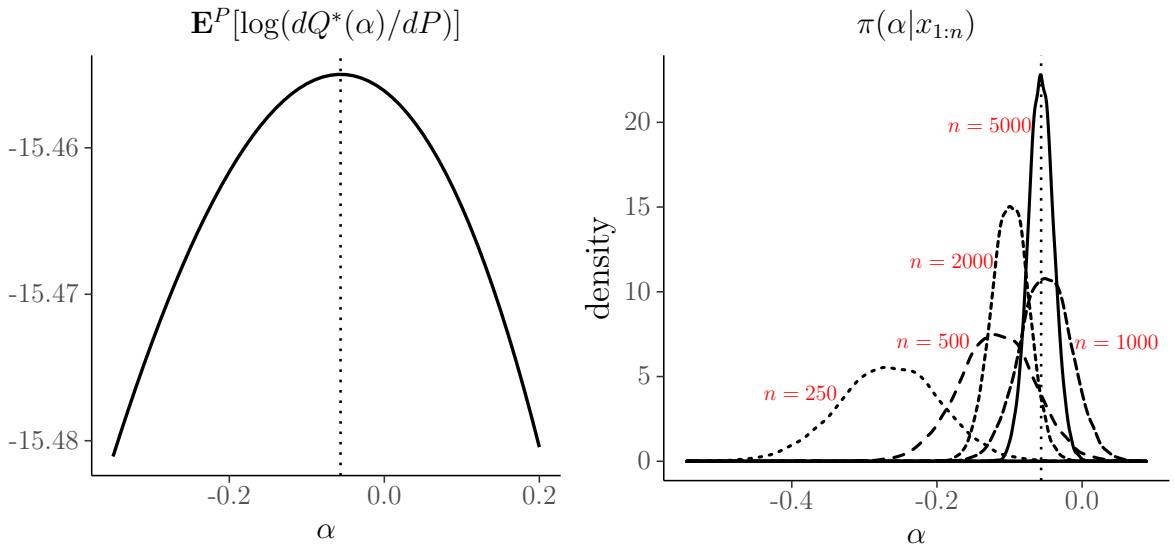

Figure 2: Posterior distributions in Example 2 under misspecification. Left panel presents the function $\alpha \mapsto \mathbf{E}^P[\log(dQ^*(\alpha)/dP)]$ where $Q^*(\alpha) := \text{arginf}_{Q \in \mathcal{P}_\psi} K(Q||P)$ with $\psi := (\alpha, 0)$. For each α, we approximate this function based on the dual representation in (2.16) – which is valid under Assumption 3 – using five million simulation draws from P. In the right panel, we present the BETEL posterior distribution of the location parameter α for $n = 250, 500, 1000, 2000, 5000$ where n is the number of observations. The prior distribution for α is student-t with mean 0 and dispersion 5. Vertical dashed lines are at the pseudo-true parameter value, approximately equal to -0.056. Posterior densities are based on 10,000 draws beyond a burn-in of 1000. The M-H acceptance rate is about 90% for each sample size.

Example 3 (Model selection when all models are correctly specified)**.** *We suppose that for every* $i = 1, \ldots, n$, $y_i = \alpha + \beta z_i + e_i$, *where* $z_i \sim \mathcal{N}(0.5, 1)$ *and* $e_i \sim \mathcal{N}(0, 1)$ *independently of* z_i. *Let* $\theta := (\alpha, \beta)$, $e_i(\theta) := (y_i - \alpha - \beta z_i)$ *and the true value of* θ *be* $(0, 1)$. *We compare the following models. Model 1:* $\mathbf{E}^P[(e_i(\theta), e_i(\theta)z_i, e_i(\theta)^3, e_i(\theta)^2 - 1)'] = 0$, *Model 2:* $\mathbf{E}^P[(e_i(\theta), e_i(\theta)z_i, e_i(\theta)^2 - 1)'] = 0$ *and Model 3:* $\mathbf{E}^P[(e_i(\theta), e_i(\theta)^2 - 1)'] = 0$ *which, reformulated in terms of an encompassing grand model, become respectively:*

$$
\begin{aligned}
M_1 &: \mathbf{E}^P[e_i(\theta)] = 0, \quad \mathbf{E}^P[e_i(\theta)z_i] = 0, \quad \mathbf{E}^P[e_i(\theta)^3] = 0, \quad \mathbf{E}^P[e_i(\theta)^2 - 1] = 0 \\
M_2 &: \mathbf{E}^P[e_i(\theta)] = 0, \quad \mathbf{E}^P[e_i(\theta)z_i] = 0, \quad \mathbf{E}^P[e_i(\theta)^3] = v_1, \quad \mathbf{E}^P[e_i(\theta)^2 - 1] = 0, \\
M_3 &: \mathbf{E}^P[e_i(\theta)] = 0, \quad \mathbf{E}^P[e_i(\theta)z_i] = v_2, \quad \mathbf{E}^P[e_i(\theta)^3] = v_1, \quad \mathbf{E}^P[e_i(\theta)^2 - 1] = 0.
\end{aligned} \tag{A.1}
$$

with $\psi^1 = \theta$, $\psi^2 = (\theta, v_1)$ *and* $\psi^3 = (\theta, v_1, v_2)$. *Note that the last two moment restrictions (which concern the third and second moments) serve as extra information to infer the parameter* θ, *when they are active. Under the standard normal error distribution, all the three*

models are correctly specified: M_1 has four active moment restrictions while M_2 and M_3 have three and two active moment restrictions, respectively.

In Table 5, we report the percentage of times the marginal likelihood selects each of these models in 500 trials, for different sample sizes. Model M_1, the model with the larger number of valid restrictions, is selected 99% of times by sample size of $n = 500$. The results are virtually indistinguishable for the training sample prior (based on 50 prior samples). Under both priors the proportion of correct selection tends to one.

	Default prior			Training sample prior		
Model	M_1	M_2	M_3	M_1	M_2	M_3
$n = 250$	97.8	1.6	0.6	98.0	1.6	0.4
$n = 500$	99.0	0.8	0.2	99.0	0.8	0.2
$n = 1000$	99.2	0.6	0.2	99.2	0.6	0.0
$n = 2000$	99.2	0.8	0.0	99.2	0.8	0.0

Table 5: Model selection when all models are correctly specified. Frequency (%) of times each of the three models in Example 3 are selected by the marginal likelihood criterion in 500 trials, by sample size, for two different prior distributions.

Example (Count regression: variable selection (continued))**.** *Consider the case where the data are drawn under the Poisson assumption (this information is, of course, not used in the estimation). Specifically, suppose we generate n realizations of $\{y_i, x_i\}$ from the Poisson model*

$$\begin{aligned} &y_i|\beta, x_i \sim Poisson(\mu_i), \qquad i = 1, \ldots, n \\ &\log(\mu_i) = x_i'\beta. \end{aligned} \tag{A.2}$$

where $\beta = (\beta_1, \beta_2, \beta_3)'$ and $x_i = (x_{1,i}, x_{2,i}, x_{3,i})'$, with $\beta_1 = 1$, $\beta_2 = 1$, $\beta_3 = 0$. Thus, $x_{3,i}$ is a redundant regressor. Each explanatory variable $x_{j,i}$ is generated i.i.d. from normal distributions with mean .4 and standard deviation 1/3. Given these data, our goal is to evaluate the finite-sample performance of our marginal likelihood criterion in picking out the correct model. We conduct our MCMC analysis and compute the marginal likelihoods of the four models by the Chib (1995) method under the default student-t prior distribution on β given in (2.11). The results, in Table 6, give the percentage of times in 500 replications that the marginal likelihood criterion picks each model for three different sample sizes. As can be seen, the model with the largest number of overidentifying moment restrictions M_1 is selected by the marginal likelihood criterion with frequency close to one even when $n = 250$.

Model	M_1	M_2	M_3	M_4
$n = 250$	0.99	0.01	0.01	0.00
$n = 500$	0.99	0.00	0.01	0.00
$n = 1000$	0.99	0.00	0.00	0.00

Table 6: Frequency (%) of times each of the four models in (4.3) are selected by the marginal likelihood criterion in 500 trials. The DGP is the Poisson model given in (A.2).

B Assumptions

In this section we state the assumptions that are used to prove the theorems in the paper. For completeness we also report Assumptions 1 – 4 that were already stated in the paper. We start by stating the assumptions that are used in Theorem 2.1 and then the other assumptions relevant for misspecification. As a consequence the numbering is not in the order.

Assumption 1. *Model (2.2) is such that $\psi_* \in \Psi$ is the unique solution to $\mathbf{E}^P[g^A(X, \psi)] = 0$.*

Assumption 2. *(a) π is a continuous probability measure that admits a density with respect to the Lebesgue measure; (b) π is positive on a neighborhood of ψ_*.*

The following two assumptions relate to the smoothness of the function $g^A(x, \psi)$, its moments, and the parameter space.

Assumption 5. *(a) X_i, $i = 1, \ldots, n$ are i.i.d. random variables that take values in $(\mathcal{X}, \mathfrak{B}_\mathcal{X})$ with probability distribution P, where $\mathcal{X} \subseteq \mathbb{R}^{d_x}$; (b) for every $0 \leq d_v \leq d - p$, $\psi \in \Psi \subset \mathbb{R}^p \times \mathbb{R}^{d_v}$ where Θ and $\mathcal{V}$ are compact and connected and $\Psi := \Theta \times \mathcal{V}$; (c) $g(x, \theta)$ is continuous at each $\theta \in \Theta$ with probability one; (d) $\mathbf{E}^P[\sup_{\psi \in \Psi} \|g^A(X, \psi)\|^\alpha] < \infty$ for some $\alpha > 2$; (e) Δ is nonsingular.*

Assumption 6. *(a) $\psi_* \in int(\Psi)$; (b) $g^A(x, \psi)$ is continuously differentiable in a neighborhood $\mathfrak{U}$ of ψ_* and $\mathbf{E}^P[\sup_{\psi \in \mathfrak{U}} \|\partial g^A(X, \psi)/\partial \psi'\|_F] < \infty$; (c) $rank(\Gamma) = p$.*

Assumptions 5 and 6 are the same as the assumptions of Newey and Smith (2004, Theorem 3.2) and Schennach (2007, Theorem 3).

We now consider misspecified models.

Assumption 3. *For a fixed $\psi \in \Psi$, there exists $Q \in \mathcal{P}_\psi$ such that Q is mutually absolutely continuous with respect to P, where $\mathcal{P}_\psi$ is defined in Definition 2.1.*

Assumption 4. *The prior distribution π is positive on a neighborhood of $\psi_\circ$ where $\psi_\circ$ is as defined in (2.16).*

In addition to these assumptions, to prove Theorem 2.2 we also use Assumptions 5 *(a)-(d)* and 6 *(b)* in the previous section. Finally, in order to guarantee $n^{-1/2}$-convergence of $\widehat{\lambda}$ towards $\lambda_\circ$ and $n^{-1/2}$-contraction of the posterior distribution of ψ around $\psi_\circ$, we introduce Assumptions 7 and 8. These assumptions require the pseudo-true values $\lambda_\circ$ and $\psi_\circ$ to be in the interior of a compact parameter space, and the function $g^A(x,\psi)$ to be sufficiently smooth and uniformly bounded as a function of ψ. These assumptions are not new in the literature and are also required by Schennach (2007, Theorem 10) (adapted to account for the augmented model).

Assumption 7. *(a) There exists a function $M(\cdot)$ such that $\mathbf{E}^P[M(X)] < \infty$ and $\|g^A(x,\psi)\| \leq M(x)$ for all $\psi \in \Psi$; (b) $\lambda_\circ(\psi) \in int(\Lambda(\psi))$ where $\Lambda(\psi)$ is a compact set and $\lambda_\circ$ is as defined in (2.16); (c) it holds $\mathbf{E}^P\left[\sup_{\psi\in\Psi,\lambda\in\Lambda(\psi)} e^{\{\lambda' g^A(X,\psi)\}}\right] < \infty$.*

Assumption 8. *Let $\psi_\circ$ be as defined in (2.16). (a) The pseudo-true value $\psi_\circ \in int(\Psi)$ is the unique maximizer of*

$$\lambda_\circ(\psi)'\mathbf{E}^P[g^A(X,\psi)] - \log \mathbf{E}^P[\exp\{\lambda_\circ(\psi)' g^A(X,\psi)\}],$$

where Ψ is compact; (b) $S_{jl}(x_i,\psi) := \partial^2 g^A(x_i,\psi)/\partial\psi_j\partial\psi_l$ is continuous in ψ for $\psi \in \mathcal{U}_\circ$, where $\mathcal{U}_\circ$ denotes a ball centred at $\psi_\circ$ with radius $n^{-1/2}$; (c) there exists $b(x_i)$ satisfying $\mathbf{E}^P\left[\sup_{\psi\in\mathcal{U}_\circ}\sup_{\lambda\in\Lambda(\psi)}\exp\{\kappa_1\lambda' g^A(X,\psi)\}b(X)^{\kappa_2}\right] < \infty$ for $\kappa_1 = 0,1,2$ and $\kappa_2 = 0,1,2,3,4$ such that $\|g^A(x_i,\psi)\| < b(x_i)$, $\|\partial g^A(x_i,\psi)/\partial\psi'\|_F \leq b(x_i)$ and $\|S_{jl}(x_i,\psi)\| \leq b(x_i)$ for $j,l = 1,\ldots,p$ for any $x_i \in (\mathcal{X},\mathfrak{B}_\mathcal{X})$ and for all $\psi \in \mathcal{U}_\circ$.

C Proofs for Sections 2.2 and 2.3

In this appendix we prove Theorems 2.1 and 2.2 and Lemma 2.1. It is useful to introduce some notation that will be used hereafter. The estimator $\widehat{\psi} := (\widehat{\theta},\widehat{v})$ denotes Schennach (2007)'ETEL estimator of ψ:

$$\widehat{\psi} := \arg\max_{\psi\in\Psi} \frac{1}{n}\sum_{i=1}^n \left[\widehat{\lambda}(\psi)' g^A(x_i,\psi) - \log\frac{1}{n}\sum_{j=1}^n \exp\{\widehat{\lambda}(\psi)' g^A(x_j,\psi)\}\right] \tag{C.1}$$

where $\widehat{\lambda}(\psi) := \arg\min_{\lambda\in\mathbb{R}^d} \frac{1}{n}\sum_{i=1}^n \left[\exp\{\lambda' g^A(x_i,\psi)\}\right]$. The log-likelihood ratio is:

$$l_{n,\psi}(x) - l_{n,\psi_\circ}(x) \;=\; \log\frac{e^{\widehat{\lambda}(\psi)' g^A(x,\psi)}}{\frac{1}{n}\sum_{j=1}^n\left[e^{\widehat{\lambda}(\psi)' g^A(x_j,\psi)}\right]} - \log\frac{e^{\widehat{\lambda}(\psi_\circ)' g^A(x,\psi_\circ)}}{\frac{1}{n}\sum_{j=1}^n\left[e^{\widehat{\lambda}(\psi_\circ)' g^A(x_j,\psi_\circ)}\right]}. \tag{C.2}$$

C.1 Proof of Theorem 2.1

Denote by $h := \sqrt{n}(\psi - \psi_*)$ the local parameter and $V_{\psi_*} := \Gamma'\Delta^{-1}\Gamma$. We denote by π^h and $\pi^h(\cdot|x_{1:n})$ the prior and posterior distribution, respectively, of the local parameter h. Therefore, $\pi^h(h) = n^{-d_\psi/2}\pi(\psi_* + h/\sqrt{n})$, where $d_\psi := (p + d_v)$, and

$$\pi^h(h|x_{1:n}) = \frac{\pi(\psi_* + h/\sqrt{n}) \exp\{\log \frac{p(x_{1:n}|\psi_* + h/\sqrt{n})}{p(x_{1:n}|\psi_*)}\}}{\int \pi(\psi_* + \tilde{h}/\sqrt{n}) \exp\{\log \frac{p(x_{1:n}|\psi_* + \tilde{h}/\sqrt{n})}{p(x_{1:n}|\psi_*)}\} d\tilde{h}}$$
$$=: C_n^{-1}\pi(\psi_* + h/\sqrt{n}) \exp\left\{\log \frac{p(x_{1:n}|\psi_* + h/\sqrt{n})}{p(x_{1:n}|\psi_*)}\right\}$$

and we need to show (2.15) which is equivalent to

$$\int \left| C_n^{-1}\pi(\psi_* + h/\sqrt{n}) \exp\left\{\log \frac{p(x_{1:n}|\psi_* + h/\sqrt{n})}{p(x_{1:n}|\psi_*)}\right\} - (2\pi)^{-d_\psi/2}|V_{\psi_*}|^{1/2} e^{\{-h'V_{\psi_*}h/2\}} \right| dh \xrightarrow{p} 0. \tag{C.3}$$

Remark that

$$\int \left| C_n^{-1}\pi(\psi_* + h/\sqrt{n}) \exp\left\{\log \frac{p(x_{1:n}|\psi_* + h/\sqrt{n})}{p(x_{1:n}|\psi_*)}\right\} - (2\pi)^{-d_\psi/2}|V_{\psi_*}|^{1/2} e^{\{-h'V_{\psi_*}h/2\}} \right| dh$$
$$\leq C_n^{-1} \int \left| \pi(\psi_* + h/\sqrt{n}) \exp\left\{\log \frac{p(x_{1:n}|\psi_* + h/\sqrt{n})}{p(x_{1:n}|\psi_*)}\right\} - \pi(\psi_*) \exp\{-h'V_{\psi_*}h/2\} \right| dh$$
$$+ C_n^{-1} \int \left| \pi(\psi_*) \exp\{-h'V_{\psi_*}h/2\} - C_n(2\pi)^{-d_\psi/2}|V_{\psi_*}|^{1/2} \exp\{-h'V_{\psi_*}h/2\} \right| dh$$
$$=: \mathcal{I}_1 + \mathcal{I}_2. \tag{C.4}$$

Term $\mathcal{I}_1 \xrightarrow{p} 0$ by Lemma C.1 below. Because Lemma C.1 implies that $C_n \xrightarrow{p} \pi(\psi_*)(2\pi)^{d_\psi/2}|V_{\psi_*}|^{-1/2}$, then term $\mathcal{I}_2 \xrightarrow{p} 0$ and this concludes the proof of the theorem.

□

Lemma C.1. *Under Assumptions 1, 2, 5, 6 and (2.14),*

$$\int \left| \pi(\psi_* + h/\sqrt{n}) \exp\left\{\log \frac{p(x_{1:n}|\psi_* + h/\sqrt{n})}{p(x_{1:n}|\psi_*)}\right\} - \pi(\psi_*) \exp\{-h'V_{\psi_*}h/2\} \right| dh \xrightarrow{p} 0 \tag{C.5}$$

Proof. Given any $\delta, c > 0$ we break the domain of integration into three regions: (I) $A_1 := \{h; \|h\| < c \log \sqrt{n}\}$; (II) $A_2 := \{h; c \log \sqrt{n} < \|h\| < \delta\sqrt{n}\}$; (III) $A_3 := \{h; \|h\| > \delta\sqrt{n}\}$. We begin with A_3:

$$\int_{A_3} \left| \pi(\psi_* + h/\sqrt{n}) \exp\{\log \frac{p(x_{1:n}|\psi_* + h/\sqrt{n})}{p(x_{1:n}|\psi_*)}\} - \pi(\psi_*) \exp\{-h' V_{\psi_*} h/2\} \right| dh$$
$$\leq \int_{A_3} \pi(\psi_* + h/\sqrt{n}) e^{\{\sum_{i=1}^n (l_{n,\psi_* + h/\sqrt{n}}(x_i) - l_{n,\psi_*}(x_i))\}} dh + \int_{A_3} \pi(\psi_*) e^{\{-h' V_{\psi_*} h/2\}} dh.$$

The first integral goes to zero by (2.14). The second integral goes to zero by the properties of the tails of a normal distribution. Let us consider A_1. By (C.9) and (C.7) in Lemma C.2 we have, for a generic constant C:

$$\int_{A_1} \left| \pi(\psi_* + h/\sqrt{n}) \exp\left\{\log \frac{p(x_{1:n}|\psi_* + h/\sqrt{n})}{p(x_{1:n}|\psi_*)}\right\} - \pi(\psi_*) \exp\{-h' V_{\psi_*} h/2\} \right| dh$$
$$\leq e^{Cn^{-1/2} \log \sqrt{n}} \int_{A_1} \pi(\psi_* + h/\sqrt{n}) \left| e^{-h' V_{\psi_*} h/2 + Cn^{-1/2}\|h\|^2} - e^{-h' V_{\psi_*} h/2} \right| dh + o_p(\log \sqrt{n}/\sqrt{n})$$
$$+ \int_{A_1} \left| \pi(\psi_* + h/\sqrt{n}) - \pi(\psi_*) \right| e^{-h' V_{\psi_*} h/2} dh.$$

Because π is continuous at ψ_* by Assumption 2, the second integral goes to zero. Because $\left| e^{Cn^{-1/2}\|h\|^2} - 1 \right| \leq e^{Cn^{-1/2}\|h\|^2} \left| Cn^{-1/2}\|h\|^2 \right|$ for a generic constant C, the first integral is

$$\leq \sup_{h \in A_1} \pi(\psi_* + h/\sqrt{n}) \int_{A_1} e^{-h' V_{\psi_*} h/2} \left| e^{Cn^{-1/2}\|h\|^2} - 1 \right|$$
$$\leq \sup_{h \in A_1} \pi(\psi_* + h/\sqrt{n}) \sup_{h \in A_1} e^{Cn^{-1/2}\|h\|^2} \left| Cn^{-1/2}\|h\|^2 \right| \int_{A_1} e^{-h' V_{\psi_*} h/2} dh = o_p(1).$$

Next, consider the last region of integration and use (C.8) and (C.9):

$$\int_{A_2} \left| \pi(\psi_* + h/\sqrt{n}) \exp\left\{\log \frac{p(x_{1:n}|\psi_* + h/\sqrt{n})}{p(x_{1:n}|\psi_*)}\right\} - \pi(\psi_*) \exp\{-h' V_{\psi_*} h/2\} \right| dh$$

$$\leq e^{Cn^{-1/2}\sqrt{n}} \int_{A_2} \pi(\psi_* + h/\sqrt{n}) e^{-h'V_{\psi_*}h/2 + h'\left(\frac{1}{n}\sum_{i=1}^n \ddot{l}_{n,\psi_t}(x_i) - V_{\psi_*}\right)h/2} dh + \int_{A_2} \pi(\psi_*) e^{-h'V_{\psi_*}h/2} dh. \tag{C.6}$$

The second integral can be upper bounded as (for a generic constant $C > 0$):

$$\int_{A_2} \pi(\psi_*) e^{-h'V_{\psi_*}h/2} dh \leq 2\pi(\psi_*) e^{-c\log(\sqrt{n})\rho_{\min}(V_{\psi_*})/2} (\delta\sqrt{n} - c\log\sqrt{n}) \leq C\pi(\psi_*) \frac{\sqrt{n}}{n^{c\rho_{\min}(V_{\psi_*})/4}}$$

so that by choosing $c > 2\rho_{\min}$, the integral goes to zero because, under Assumptions 5 *(e)* and 6 *(c)*, $\rho_{\min}(V_{\psi_*})$ is strictly positive, where $\rho_{\min}(V_{\psi_*})$ denotes the minimum eigenvalue of the matrix V_{ψ_*}. To control the first integral in (C.6), there exists a N such that for all $n \geq N$: $P\left(-h'V_{\psi_*}h/2 + h'\left(\frac{1}{n}\sum_{i=1}^n \ddot{l}_{n,\psi_t}(x_i) - V_{\psi_*}\right)h/2 < -h'V_{\psi_*}h/4 \text{ for all } h \in A_2\right) > 1 - \epsilon$. Therefore, with probability larger than $1 - \epsilon$,

$$\int_{A_2} \pi(\psi_* + \tilde{h}/\sqrt{n}) e^{-h'V_{\psi_*}h/2 + h'\left(\frac{1}{n}\sum_{i=1}^n \ddot{l}_{n,\psi_t}(x_i) - V_{\psi_*}\right)h/2} dh \leq \sup_{h \in A_2} \pi(\psi_* + h/\sqrt{n}) \int_{A_2} e^{-h'V_{\psi_*}h/4} dh$$

which converges to zero as $n \to \infty$. Finally, by putting these three results together we show (C.5).

□

Lemma C.2. *Let Assumptions 1, 2, 5 and 6 hold and denote* $h := \sqrt{n}(\psi - \psi_*)$ *and* $V_{\psi_*} := \Gamma'\Delta^{-1}\Gamma$. *Then,*

$$\log \frac{p(x_{1:n}|\psi_* + h/\sqrt{n})}{p(x_{1:n}|\psi_*)} = -\frac{1}{2}h'V_{\psi_*}h + O_p((\|h\| + \|h\|^2)n^{-1/2}). \tag{C.7}$$

Proof. Denote $d_\psi := (p + d_v)$, $\tau(\widehat{\lambda}, \psi, x) := e^{\widehat{\lambda}(\psi)' g^A(x,\psi)}$ and $\tau_n(\widehat{\lambda}, \psi) := \frac{1}{n}\sum_{i=1}^n \tau(\widehat{\lambda}, \psi, x_i)$. Moreover, let $G^A(x, \psi_*) := \partial g^A(x, \psi)/\partial\psi'\big|_{\psi=\psi_*}$ be a matrix of dimension $d \times d_\psi$. A first order Taylor expansion of $h \mapsto \log p(x_{1:n}|\psi_* + h/\sqrt{n})$ around $h = 0$, with Lagrange remainder, gives:

$$\log \frac{p(x_{1:n}|\psi_* + h/\sqrt{n})}{p(x_{1:n}|\psi_*)} = \frac{1}{\sqrt{n}}\sum_{i=1}^n \dot{l}_{n,\psi_*}(x_i)'h + \frac{1}{2n}\sum_{i=1}^n h'\ddot{l}_{n,\psi_t}(x_i)h \tag{C.8}$$

where $\dot{l}_{n,\psi_*}(x) := \partial l_{n,\psi}(x)/\partial\psi\big|_{\psi=\psi_*}$, $\ddot{l}_{n,\psi_t}(x) := \partial^2 l_{n,\psi}(x)/(\partial\psi\partial\psi')\big|_{\psi=\psi_t}$ and $\psi_t := \psi_* +$

$th/\sqrt{n}$, $t \in [0,1]^{d_\psi}$. Simple computations give:

$$\begin{aligned}\frac{1}{\sqrt{n}}\sum_{i=1}^{n} \dot{l}_{n,\psi_*}(x_i)'h &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(1-\frac{\tau(\widehat{\lambda},\psi_*,x_i)}{\tau_n(\widehat{\lambda},\psi_*)}\right)\left(\widehat{\lambda}(\psi_*)'G^A(x_i,\psi_*)+g^A(x_i,\psi_*)'\frac{d\widehat{\lambda}(\psi_*)}{d\psi'}\right)h \\ &= O_p(n^{-1/2}\|h\|) \end{aligned} \tag{C.9}$$

since under Assumption 5, $\widehat{\lambda}(\psi_*) = O_p(n^{-1/2})$ by Newey and Smith (2004, Theorem 3.1) and $\left|1-\frac{\tau(\widehat{\lambda},\psi,x_i)}{\tau_n(\widehat{\lambda},\psi)}\right| = O_p(n^{-1/2})$ by continuity of $\psi \mapsto \widehat{\lambda}(\psi)$ (due to the Birge's maximum theorem and strict convexity of $\lambda \mapsto \frac{1}{n}\sum_{i=1}\exp\{\lambda' g^A(x_i,\psi)\}$). Denote $\mathcal{A}_1(\widehat{\lambda},\psi,x_i) := \left(\widehat{\lambda}(\psi)'G^A(x_i,\psi)+g^A(x_i,\psi)'\frac{d\widehat{\lambda}(\psi)}{d\psi'}\right)$. Then,

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n} h'\ddot{l}_{n,\psi_t}(x_i)h = \frac{h'}{n}\sum_{i=1}^{n}\left(1-\frac{\tau(\widehat{\lambda},\psi_t,x_i)}{\tau_n(\widehat{\lambda},\psi_t)}\right)\Big(\sum_{j=1}^{d}\widehat{\lambda}_j(\psi_t)\frac{\partial^2 g_j^A(x_i,\psi_t)}{\partial\psi\partial\psi'} \\ + \frac{d\widehat{\lambda}(\psi_t)'}{d\psi}G^A(x_i,\psi_t)+G^A(x_i,\psi_t)'\frac{d\widehat{\lambda}(\psi_t)}{d\psi'}+\sum_{j=1}^{d} g_j^A(x_i,\psi_t)\frac{d^2\widehat{\lambda}_j(\psi_t)}{d\psi d\psi'}\Big)h \\ -h'\frac{1}{n}\sum_{i=1}^{n}\left(\frac{\tau(\widehat{\lambda},\psi_t,x_i)}{\tau_n(\widehat{\lambda},\psi_t)}\mathcal{A}_1(\widehat{\lambda},\psi_t,x_i)'-\frac{\tau(\widehat{\lambda},\psi_t,x_i)}{\tau_n^2(\widehat{\lambda},\psi_t)}\frac{1}{n}\sum_{j=1}^{n}\tau(\widehat{\lambda},\psi_t,x_j)\mathcal{A}_1(\widehat{\lambda},\psi_t,x_j)'\right)\mathcal{A}_1(\widehat{\lambda},\psi_t,x_i)h \\ = -h'\frac{1}{n}\sum_{i=1}^{n}\frac{\tau(\widehat{\lambda},\psi_t,x_i)}{\tau_n(\widehat{\lambda},\psi_t)}\frac{d\widehat{\lambda}(\psi_t)'}{d\psi}g^A(x_i,\psi_t)g^A(x_i,\psi_t)'\frac{d\widehat{\lambda}(\psi_t)}{d\psi'}h+O_p(\|h\|^2 n^{-1/2}) \\ = -h'\Gamma'\Delta^{-1}\Gamma h+O_p(\|h\|^2 n^{-1/2}) \end{aligned} \tag{C.10}$$

because: *(i)* under Assumption 5, $\widehat{\lambda}(\psi_*) = O_p(n^{-1/2})$ by Newey and Smith (2004, Theorem 3.1) so that $\sup_{t\in[0,1]^{d_\psi}}\left|1-\frac{\tau(\widehat{\lambda},\psi_t,x_i)}{\tau_n(\widehat{\lambda},\psi_t)}\right| = O_p(n^{-1/2})$ by continuity of $\psi \mapsto \widehat{\lambda}(\psi)$ (due to the Birge's maximum theorem and strict convexity of $\lambda \mapsto \frac{1}{n}\sum_{i=1}\exp\{\lambda' g^A(x_i,\psi)\}$); *(ii)* $\frac{1}{n}\sum_{j=1}^{n}\tau(\widehat{\lambda},\psi_t,x_j)g^A(x_j,\psi_t) = O_p(n^{-1/2})$ by the results in *(i)* and Newey and Smith (2004, Lemma A.3); *(iii)* $\frac{d\widehat{\lambda}(\psi_t)}{d\psi'} = -\Delta^{-1}\Gamma + O_p(n^{-1/2})$ (by Assumptions 5 *(b)* - *(d)* and 6 *(b)* - *(c)*). By replacing (C.9) and (C.10) in (C.8) we get the result of the lemma.

□

C.2 Proof of Theorem 2.2.

The main steps of this proof proceed as in the proof of Van der Vaart (1998, Theorem 10.1) and Kleijn and van der Vaart (2012, Theorem 2.1) while the proofs of the technical theorems and lemmas that we use all along this proof are new. Let us consider a reparametrization of

the model centred around the pseudo-true value $\psi_\circ$ and define a local parameter $h = \sqrt{n}(\psi - \psi_\circ)$. Denote by π^h and $\pi^h(\cdot|x_{1:n})$ the prior and posterior distribution of h, respectively. Denote by Φ_n the normal distribution $\mathcal{N}_{\Delta_{n,\psi_\circ}, V_{\psi_\circ}^{-1}}$ and by ϕ_n its Lebesgue density. For a compact subset $K \subset \mathbb{R}^p$ such that $\pi^h(h \in K|x_{1:n}) > 0$ define, for any Borel set $B \subseteq \Psi$,

$$\pi_K^h(B|x_{1:n}) := \frac{\pi^h(K \cap B|x_{1:n})}{\pi^h(K|x_{1:n})}$$

and let Φ_n^K be the Φ_n distribution conditional on K. The proof consists of two steps. In the first step we show that the Total Variation (TV) norm of $\pi_K^h(\cdot|x_{1:n}) - \Phi_n^K$ converges to zero in probability. In the second step we use this result to show that the TV norm of $\pi^h(\cdot|x_{1:n}) - \Phi_n$ converges to zero in probability.

Let Assumption 8 *(a)* hold. For every open neighborhood $\mathcal{U} \subset \Psi$ of $\psi_\circ$ and a compact subset $K \subset \mathbb{R}^p$, there exists an N such that for every $n \geq N$:

$$\psi_\circ + K\frac{1}{\sqrt{n}} \subset \mathcal{U}. \tag{C.11}$$

Define the function $f_n : K \times K \to \mathbb{R}$ as, $\forall k_1, k_2 \in K$:

$$f_n(k_1, k_2) := \left(1 - \frac{\phi_n(k_2)s_n(k_1)\pi^h(k_1)}{\phi_n(k_1)s_n(k_2)\pi^h(k_2)}\right)_+$$

where $(a)_+ = \max(a, 0)$, here π^h denotes the Lebesgue density of the prior π^h for h and $s_n(h) := p(x_{1:n}|\psi_\circ + h/\sqrt{n})/p(x_{1:n}|\psi_\circ)$. The function f_n is well defined for n sufficiently large because of (C.11) and Assumption 8 *(a)*. Remark that by (C.11) and since the prior for ψ puts enough mass on $\mathcal{U}$, then π^h puts enough mass on K and as $n \to \infty$: $\pi^h(k_1)/\pi^h(k_2) \to 1$. Because of this and by the stochastic LAN expansion (C.16) in Theorem C.1 below:

$$\begin{aligned}\log\frac{\phi_n(k_2)s_n(k_1)\pi^h(k_1)}{\phi_n(k_1)s_n(k_2)\pi^h(k_2)} &= -\frac{1}{2}(k_2 - \Delta_{n,\psi^\circ})'V_{\psi^\circ}(k_2 - \Delta_{n,\psi^\circ}) + \frac{1}{2}(k_1 - \Delta_{n,\psi^\circ})'V_{\psi_\circ}(k_1 - \Delta_{n,\psi_\circ}) \\ &\quad + k_1'V_{\psi_\circ}\Delta_{n,\psi_\circ} - \frac{1}{2}k_1'V_{\psi_\circ}k_1 - k_2'V_{\psi_\circ}\Delta_{n,\psi_\circ} + \frac{1}{2}k_2'V_{\psi_\circ}k_2 + o_p(1) = o_p(1). \end{aligned} \tag{C.12}$$

Since, for every n, f_n is continuous in (k_1, k_2) and $K \times K$ is compact, then

$$\sup_{k_1,k_2 \in K} f_n(k_1, k_2) \xrightarrow{p} 0, \qquad \text{as } n \to \infty. \tag{C.13}$$

Suppose that the subset K contains a neighborhood of 0 (which guarantees that $\Phi_n(K) > 0$ and then that Φ_n^K is well defined) and let $\Xi_n := \{\pi^h(K|x_{1:n}) > 0\}$. Moreover, for a given $\eta > 0$ define the event $\Omega_n := \left\{\sup_{k_1,k_2\in K} f_n(k_1,k_2) \leq \eta\right\}$. The TV distance $\|\cdot\|_{TV}$ between two probability measures P and Q, with Lebesgue densities p and q respectively, can be expressed as: $\|P - Q\|_{TV} = 2\int(1 - p/q)_+ dQ$. Therefore, by the Jensen inequality and convexity of the functions $(\cdot)_+$,

$$\begin{aligned}\frac{1}{2}\mathbf{E}^P\|\Phi_n^K - \pi_K^h(\cdot|x_{1:n})\|_{TV}1_{\Omega_n\cap\Xi_n} &= \mathbf{E}^P\int_K\left(1 - \frac{d\Phi_n^K(k_2)}{d\pi_K^h(k_2|x_{1:n})}\right)_+ d\pi_K^h(k_2|x_{1:n})1_{\Omega_n\cap\Xi_n}\\ &\leq \mathbf{E}^P\int_K\int_K f_n(k_1,k_2)d\Phi_n^K(k_1)d\pi_K^h(k_2|x_{1:n})1_{\Omega_n\cap\Xi_n}\\ &\leq \mathbf{E}^P\sup_{k_1,k_2\in K} f_n(k_1,k_2)1_{\Omega_n\cap\Xi_n} \quad \text{(C.14)}\end{aligned}$$

that converges to zero by (C.13). By (C.14), it follows that (by remembering that $\|\cdot\|_{TV}$ is upper bounded by 2)

$$\mathbf{E}^P\|\pi_K^h(\cdot|x_{1:n}) - \Phi_n^K\|_{TV}1_{\Xi_n} \leq \mathbf{E}^P\|\pi_K^h(\cdot|x_{1:n}) - \Phi_n^K\|_{TV}1_{\Omega_n\cap\Xi_n} + 2P(\Omega_n^c \cap \Xi_n), \quad \text{(C.15)}$$

where the second term is $o(1)$ by (C.13). In the second step of the proof let K_n be a sequence of closed balls in the parameter space of h centred at 0 with radii $M_n \to \infty$ and redefine Ξ_n accordingly. For each $n \geq 1$, (C.15) holds for these balls. Moreover, by (C.18) in Theorem C.2 below: $P(\Xi_n) \to 1$. Therefore, by the triangular inequality, the TV distance is upper bounded by

$$\begin{aligned}\mathbf{E}^P\|\pi^h(\cdot|x_{1:n}) - \Phi_n\|_{TV} &\leq \mathbf{E}^P\|\pi^h(\cdot|x_{1:n}) - \Phi_n\|_{TV}1_{\Xi_n} + \mathbf{E}^P\|\pi^h(\cdot|x_{1:n}) - \Phi_n\|_{TV}1_{\Xi_n^c}\\ &\leq \mathbf{E}^P\|\pi^h(\cdot|x_{1:n}) - \pi_{K_n}^h(\cdot|x_{1:n})\|_{TV} + \mathbf{E}^P\|\pi_{K_n}^h(\cdot|x_{1:n}) - \Phi_n^{K_n}\|_{TV}1_{\Xi_n}\\ &\quad + \mathbf{E}^P\|\Phi_n^{K_n} - \Phi_n\|_{TV} + 2P(\Xi_n^c)\\ &\leq 2\mathbf{E}^P(\pi_{K_n^c}^h(\cdot|x_{1:n})) + \mathbf{E}^P\|\pi_{K_n}^h(\cdot|x_{1:n}) - \Phi_n^{K_n}\|_{TV}1_{\Xi_n} + o(1) \xrightarrow{p} 0\end{aligned}$$

since $\mathbf{E}^P(\pi^h(K_n^c|x_{1:n})) = o(1)$ by (C.18) and $\mathbf{E}^P\|\pi_{K_n}^h(\cdot|x_{1:n}) - \Phi_n^{K_n}\|_{TV}1_{\Xi_n} = o_P(1)$ by (C.15) and (C.14), and where in the third line we have used the fact that: $\mathbf{E}^P\|\pi^h(\cdot|x_{1:n}) - \pi_{K_n}^h(\cdot|x_{1:n})\|_{TV} = 2\mathbf{E}^P(\pi_{K_n^c}^h(\cdot|x_{1:n}))$ and $\|\Phi_n^{K_n} - \Phi_n\|_{TV} = \|\Phi_n^{K_n^c}\|_{TV} = o_p(1)$ by Kleijn and van der Vaart (2012, Lemma 5.2) since Δ_{n,ψ_0} is uniformly tight.

□

The next theorem establishes that the misspecified model satisfies a stochastic Local Asymptotic Normality (LAN) expansion around the pseudo-true value $\psi_\circ$.

Theorem C.1 (Stochastic LAN)**.** *Assume that the matrix $V_{\psi_\circ}$ is nonsingular and that Assumptions 5 (a)-(d), 6 (b), 3, 7, and 8 hold. Then for every compact set $K \subset \mathbb{R}^p$,*

$$\sup_{h \in K} \left| \log \frac{p(x_{1:n}|\psi_\circ + h/\sqrt{n})}{p(x_{1:n}|\psi_\circ)} - h' V_{\psi_\circ} \Delta_{n,\psi_\circ} + \frac{1}{2} h' V_{\psi_\circ} h \right| \xrightarrow{p} 0 \tag{C.16}$$

where $\psi_\circ$ is as defined in (2.16), $V_{\psi_\circ} := -\mathbf{E}^P[\ddot{\mathfrak{L}}_{n,\psi_\circ}]$ and $\Delta_{n,\psi_\circ} := \frac{1}{\sqrt{n}} \sum_{i=1}^n V_{\psi_\circ}^{-1} \dot{\mathfrak{L}}_{n,\psi_\circ}(x_i)$ is bounded in probability.

Proof. See Appendix E.

□

The next theorem establishes that the posterior of ψ concentrates and puts all its mass on $\Psi_n := \{\|\psi - \psi_\circ\| \leq M_n/\sqrt{n}\}$ as $n \to \infty$.

Theorem C.2 (Posterior Consistency)**.** *Assume that the stochastic LAN expansion (C.16) holds for $\psi_\circ$ defined in (2.16). Moreover, let Assumptions 2 (a), 3 and 4 hold and assume that there exists a constant $C > 0$ such that for any sequence $M_n \to \infty$,*

$$P\left(\sup_{\psi \in \Psi_n^c} \frac{1}{n} \sum_{i=1}^n (l_{n,\psi}(x_i) - l_{n,\psi_\circ}(x_i)) \leq -\frac{C M_n^2}{n} \right) \to 1 \tag{C.17}$$

as $n \to \infty$ where $\Psi_n := \{\|\psi - \psi_\circ\| \leq M_n/\sqrt{n}\}$. Then,

$$\pi\left(\sqrt{n}\|\psi - \psi_\circ\| > M_n \middle| x_{1:n} \right) \xrightarrow{p} 0 \tag{C.18}$$

for any $M_n \to \infty$, as $n \to \infty$.

Proof. See Appendix E.

□

C.3 Proof of Lemma 2.1.

By Theorem 10 of Schennach (2007), which is valid under Assumptions 5 *(a)-(c)*, 3, 7 *(c)*, *(e)* and 8: $\sqrt{n}(\widehat{\psi} - \psi_\circ) = O_p(1)$. Denote $\widehat{h} := \sqrt{n}(\widehat{\psi} - \psi_\circ)$ and $\widetilde{h} := \Delta_{n,\psi_\circ}$. Because of

(C.16), we have:

$$\sum_{i=1}^{n}\left(l_{n,\psi_\circ+\widehat{h}/\sqrt{n}}-l_{n,\psi_\circ}\right)(x_i) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\widehat{h}'\dot{\mathfrak{L}}_{n,\psi_\circ}(x_i)-\frac{1}{2}\widehat{h}'V_{\psi_\circ}\widehat{h}+o_p(1) \quad (C.19)$$

$$\sum_{i=1}^{n}\left(l_{n,\psi_\circ+\widetilde{h}/\sqrt{n}}-l_{n,\psi_\circ}\right)(x_i) = \frac{1}{2\sqrt{n}}\sum_{i=1}^{n}\widetilde{h}'\dot{\mathfrak{L}}_{n,\psi_\circ}(x_i)+o_p(1). \quad (C.20)$$

By definition of $\widehat{\psi}$ as the maximizer of $\sum_{i=1}^{n} l_{n,\psi}(x_i)$, the left hand side of (C.19) is not smaller than the left hand side of (C.20). It follows that the same relation holds for the right hand sides of (C.19) and (C.20), and by taking their difference we obtain:

$$-\frac{1}{2}\left(\widehat{h}-\frac{1}{\sqrt{n}}\sum_{i=1}^{n}V_{\psi_\circ}^{-1}\dot{\mathfrak{L}}_{n,\psi_\circ}(x_i)\right)'V_{\psi_\circ}\left(\widehat{h}-\frac{1}{\sqrt{n}}\sum_{i=1}^{n}V_{\psi_\circ}^{-1}\dot{\mathfrak{L}}_{n,\psi_\circ}(x_i)\right)+o_p(1)\geq 0. \quad (C.21)$$

Because $-V_{\psi_\circ}$ is negative definite then

$$-\frac{1}{2}\left(\widehat{h}-\frac{1}{\sqrt{n}}\sum_{i=1}^{n}V_{\psi_\circ}^{-1}\dot{\mathfrak{L}}_{n,\psi_\circ}(x_i)\right)'V_{\psi_\circ}\left(\widehat{h}-\frac{1}{\sqrt{n}}\sum_{i=1}^{n}V_{\psi_\circ}^{-1}\dot{\mathfrak{L}}_{n,\psi_\circ}(x_i)\right)\leq 0.$$

This and (C.21) imply that $\left\|V_{\psi_\circ}^{-1/2}\left(\widehat{h}-\frac{1}{\sqrt{n}}\sum_{i=1}^{n}V_{\psi_\circ}^{-1}\dot{\mathfrak{L}}_{n,\psi_\circ}(x_i)\right)\right\|\xrightarrow{p}0$ which in turn implies that

$$\left\|\left(\widehat{h}-\frac{1}{\sqrt{n}}\sum_{i=1}^{n}V_{\psi_\circ}^{-1}\dot{\mathfrak{L}}_{n,\psi_\circ}(x_i)\right)\right\|\xrightarrow{p}0$$

which establishes the result of the lemma.

□

D Proofs for Section 3.3

In this appendix we prove Theorems 3.1 and 3.2 and Corollary 3.1. The proofs of Theorems 3.1 and 3.2 have already been stated in the Appendix of the paper but for easiness of reading we give them also here. For the same reason we remind the notation already introduced in the Appendix of the paper. Recall the notation $\psi^\ell=(\theta^\ell,v^\ell)$ and the estimator $\widehat{\psi}^\ell:=(\widehat{\theta}^\ell,\widehat{v}^\ell)$ denotes Schennach (2007)'ETEL estimator of ψ^ℓ in model M_ℓ:

$$\widehat{\psi}^\ell:=\arg\max_{\psi^\ell\in\Psi^\ell}\frac{1}{n}\sum_{i=1}^{n}\left[\widehat{\lambda}(\psi^\ell)'g^A(x_i,\psi^\ell)-\log\frac{1}{n}\sum_{j=1}^{n}\exp\{\widehat{\lambda}(\psi^\ell)'g^A(x_j,\psi^\ell)\}\right] \quad (D.1)$$

where $\widehat{\lambda}(\psi^\ell) = \arg\min_{\lambda\in\mathbb{R}^d} \frac{1}{n}\sum_{i=1}^n \big[\exp\{\lambda' g^A(x_i,\psi^\ell)\}\big]$. Denote $\widehat{g}^A(\psi^\ell) := \frac{1}{n}\sum_{i=1}^n g^A(x_i,\psi^\ell)$, $\widehat{g}^A_\ell := \widehat{g}^A(\psi^\ell)$,

$$\widehat{L}(\psi^\ell) := \exp\{\widehat{\lambda}(\psi^\ell)'\widehat{g}^A(\psi^\ell)\}\left[\frac{1}{n}\sum_{i=1}^n \exp\{\widehat{\lambda}(\psi^\ell)' g^A(x_i,\psi^\ell)\}\right]^{-1}$$

and $L(\psi^\ell) = \exp\{\lambda_\circ(\psi^\ell)'\mathbf{E}^P[g^A(x,\psi^\ell)]\}\left(\mathbf{E}^P\left[\exp\{\lambda_\circ(\psi^\ell)' g^A(x,\psi^\ell)\}\right]\right)^{-1}$. Moreover, we use the notation $\Sigma_\ell = \left(\Gamma_\ell'\Delta_\ell^{-1}\Gamma_\ell\right)^{-1}$ where $\Gamma_\ell := \mathbf{E}^P\left[\frac{\partial}{\partial\psi^{\ell\prime}} g^A(X,\psi^\ell_*)\right]$ and $\Delta_\ell := \mathbf{E}^P[g^A(X,\psi^\ell_*)g^A(X,\psi^\ell_*)']$. In the proofs, we omit measurability issues which can be dealt with in the usual manner by replacing probabilities with outer probabilities.

D.1 Proof of Theorem 3.1

By (3.5) and Lemmas D.1 and D.2 below we obtain

$$\begin{aligned} P\left(\max_{\ell\neq j}\log m(x_{1:n};M_\ell) < \log m(x_{1:n};M_j)\right) &= P\Big(\max_{\ell\neq j}\Big[-\frac{n}{2}\widehat{g}_\ell^{A\prime}\Delta^{-1}\widehat{g}_\ell^A + \log\pi(\widehat{\psi}^\ell|M_\ell) \\ &\quad - \frac{(p_\ell+d_{v_\ell})}{2}(\log n - \log(2\pi)) + \frac{1}{2}\log|\Sigma_\ell|\Big] + \frac{n}{2}\widehat{g}_j^{A\prime}\Delta^{-1}\widehat{g}_j^A + o_p(1) \\ &\quad < \log\pi(\widehat{\psi}^j|M_j) - \frac{(p_j+d_{v_j})}{2}(\log n - \log(2\pi)) + \frac{1}{2}\log|\Sigma_j|\Big). \quad \text{(D.2)} \end{aligned}$$

Remark that $n\widehat{g}_j^{A\prime}\Delta^{-1}\widehat{g}_j^A \xrightarrow{d} \chi^2_{d-(p_j+d_{v_j})}$, $\forall j$, so that $n\widehat{g}_j^{A\prime}\Delta^{-1}\widehat{g}_j^A = O_p(1)$. Suppose first that $(p_\ell + d_{v_\ell} > p_j + d_{v_j})$, $\forall \ell \neq j$. Since $-n\widehat{g}_\ell^{A\prime}\Delta^{-1}\widehat{g}_\ell^A < 0$ for every ℓ, we lower bound (D.2) as

$$\begin{aligned} P\left(\max_{\ell\neq j}\log m(x_{1:n};M_\ell) < \log m(x_{1:n};M_j)\right) &\geq P\Big(\frac{n}{2}\widehat{g}_j^{A\prime}\Delta^{-1}\widehat{g}_j^A + o_p(1) \\ &\quad < \log n\Big[\frac{\min_{\ell\neq j}(p_\ell+d_{v_\ell}) - p_j - d_{v_j}}{2} - \frac{\min_{\ell\neq j}(p_\ell+d_{v_\ell}) - p_j - d_{v_j}}{2\log n}\log(2\pi) \\ &\quad - \frac{\log[\max_{\ell\neq j}\pi(\widehat{\psi}^\ell|M_\ell)/\pi(\widehat{\psi}^j|M_j)]}{\log n} - \frac{1}{2\log n}\left(\max_{\ell\neq j}\log|\Sigma_\ell| - \log|\Sigma_j|\right)\Big]\Big) \\ &= P\Big(\underbrace{\frac{n}{2}\widehat{g}_j^{A\prime}\Delta^{-1}\widehat{g}_j^A + o_p(1)}_{=:\mathcal{I}_n} < \underbrace{\log n\Big[\frac{\min_{\ell\neq j}(p_\ell+d_{v_\ell}) - p_j - d_{v_j}}{2} + \mathcal{O}_p((\log n)^{-1})\Big]}_{=:\mathcal{II}_n}\Big). \quad \text{(D.3)} \end{aligned}$$

Because $\mathcal{I}_n = \mathcal{O}_p(1)$ (and is asymptotically positive) and $\mathcal{II}_n$ is strictly positive as $n\to\infty$ (since $(p_\ell + d_{v_\ell}) > (p_j + d_{v_j})$, $\forall \ell \neq j$) and converges to $+\infty$, then the probability converges to 1. This proves one direction of the statement.

To prove the second direction of the statement, suppose that $\lim_{n\to\infty} P(\max_{\ell\neq j} \log m(x_{1:n}; M_\ell) < \log m(x_{1:n}; M_j)) = 1$ and consider the following upper bound (which follows from (D.2) and the fact that $n\widehat{g}_j^{A'}\Delta^{-1}\widehat{g}_j^{A} > 0$, $\forall n$):

$$
\begin{aligned}
&P\left(\max_{\ell\neq j} \log m(x_{1:n}; M_\ell) < \log m(x_{1:n}; M_j)\right) \\
&\qquad\qquad \leq P\left(\log m(x_{1:n}; M_\ell) < \log m(x_{1:n}; M_j)\right), \qquad \forall \ell \neq j \\
&\leq P\Big(-\frac{n}{2}\widehat{g}_\ell^{A'}\Delta^{-1}\widehat{g}_\ell^{A} + o_p(1) + \log n\Big[\frac{(p_j + d_{v_j}) - (p_\ell + d_{v_\ell})}{2} + \mathcal{O}_p\left(\frac{1}{\log n}\right)\Big] < 0\Big), \quad \forall \ell \neq j.
\end{aligned}
\tag{D.4}
$$

Because the probability in the first line of (D.4) converges to 1 as $n \to \infty$ then, necessarily, the probability in the last line of (D.4) converges to 1 which is possible only if $(p_j + d_{v_j}) < (p_\ell + d_{v_\ell})$ because $\log n \left[\frac{(p_j+d_{v_j})-(p_\ell+d_{v_\ell})}{2}\right]$ is the dominating term since $-\frac{n}{2}\widehat{g}_\ell^{A'}\Delta^{-1}\widehat{g}_\ell^{A} < 0$ and it remains bounded as $n \to \infty$. Since the first inequality in (D.4) holds $\forall \ell \neq j$ then convergence to 1 of the probability in the last line of (D.4) is possible only if $(p_j + d_{v_j}) < (p_\ell + d_{v_\ell})$, $\forall \ell \neq j$.

□

D.2 Proof of Theorem 3.2

We can write $\log p(x_{1:n}|\psi^\ell; M_\ell) = -n \log n + n \log \widehat{L}(\psi^\ell)$. Then, we have:

$$
\begin{aligned}
P\left(\log m(x_{1:n}; M_j) > \max_{\ell\neq j} \log m(x_{1:n}; M_\ell)\right) &= P\Big(n \log \widehat{L}(\psi_\circ^j) + \log \pi(\psi_\circ^j|M_j) - \log \pi(\psi_\circ^j|x_{1:n}, M_j) \\
&\qquad > \max_{\ell\neq j}[n \log \widehat{L}(\psi_\circ^\ell) + \log \pi(\psi_\circ^\ell|M_\ell) - \log \pi(\psi_\circ^\ell|x_{1:n}, M_\ell)]\Big) \\
= P\Big(n \log L(\psi_\circ^j) + n \log \frac{\widehat{L}(\psi_\circ^j)}{L(\psi_\circ^j)} + \mathcal{B}_j &> \max_{\ell\neq j}\left[n \log L(\psi_\circ^\ell) + \mathcal{B}_\ell + n \log \frac{\widehat{L}(\psi_\circ^\ell)}{L(\psi_\circ^\ell)}\right]\Big)
\end{aligned}
\tag{D.5}
$$

where $\forall \ell$, $\mathcal{B}_\ell := \log \pi(\psi_\circ^\ell|M_\ell) - \log \pi(\psi_\circ^\ell|x_{1:n}, M_\ell)$ and $\mathcal{B}_\ell = O_p(1)$ under the assumptions of Theorem 2.2. By definition of $dQ^*(\psi)$ in Section 2.3 we have that: $\log L(\psi_\circ^\ell) = \mathbf{E}^P[\log dQ^*(\psi_\circ^\ell)/dP] = -\mathbf{E}^P[\log dP/dQ^*(\psi_\circ^\ell)] = -K(P||Q^*(\psi_\circ^\ell))$. Remark that $\mathbf{E}^P[\log(dP/dQ^*(\psi_\circ^2))] > \mathbf{E}^P[\log(dP/dQ^*(\psi_\circ^1))]$ means that the KL divergence between P and $Q^*(\psi_\circ^\ell)$, is smaller for model M_1 than for model M_2, where $Q^*(\psi_\circ^\ell)$ minimizes the KL divergence between $Q \in \mathcal{P}_{\psi_\circ^\ell}$ and P for $\ell \in \{1, 2\}$ (notice the inversion of the two probabilities).

First, suppose that $\min_{\ell\neq j}\mathbf{E}^P\left[\log\left(dP/dQ^*(\psi_\circ^\ell)\right)\right] > \mathbf{E}^P\left[\log\left(dP/dQ^*(\psi_\circ^j)\right)\right]$. By (D.5):

$$P\left(\log m(x_{1:n};M_j) > \max_{\ell\neq j}\log m(x_{1:n};M_\ell)\right) \geq$$
$$P\left(\log\frac{\widehat{L}(\psi_\circ^j)}{L(\psi_\circ^j)} - \max_{\ell\neq j}\log\frac{\widehat{L}(\psi_\circ^\ell)}{L(\psi_\circ^\ell)} + \frac{1}{n}(\mathcal{B}_j - \max_{\ell\neq j}\mathcal{B}_\ell) > \underbrace{\max_{\ell\neq j}\log L(\psi_\circ^\ell) - \log L(\psi_\circ^j)}_{=:\mathcal{I}_n}\right). \quad \text{(D.6)}$$

This probability converges to 1 because $\mathcal{I}_n = K(P||Q^*(\psi_\circ^j)) - \min_{\ell\neq j}K(P||Q^*(\psi_\circ^\ell)) < 0$ by assumption, and $\left[\log\widehat{L}(\psi^\ell) - \log L(\psi^\ell)\right] \xrightarrow{p} 0$, for every $\psi^\ell \in \Psi^\ell$ and every $\ell \in \{1,2\}$ by Lemma D.3 below.

To prove the second direction of the statement, suppose that $\lim_{n\to\infty}P(\log m(x_{1:n};M_j) > \max_{\ell\neq j}\log m(x_{1:n};M_\ell)) = 1$. By (D.5) it holds, $\forall\ell\neq j$

$$P\left(\log m(x_{1:n};M_j) > \max_{\ell\neq j}\log m(x_{1:n};M_\ell)\right) \leq$$
$$P\Big(\log\frac{\widehat{L}(\psi_\circ^j)}{L(\psi_\circ^j)} - \log\frac{\widehat{L}(\psi_\circ^\ell)}{L(\psi_\circ^\ell)} + \frac{1}{n}(\mathcal{B}_j - \mathcal{B}_\ell) > \log\frac{L(\psi_\circ^\ell)}{L(\psi_\circ^j)}\Big). \quad \text{(D.7)}$$

Convergence to 1 of the left hand side implies convergence to 1 of the right hand side which is possible only if $\log L(\psi_\circ^\ell) - \log L(\psi_\circ^j) < 0$. Since this is true for every model ℓ, then this implies that $K(P||Q^*(\psi_\circ^j)) < \min_{\ell\neq j}K(P||Q^*(\psi_\circ^\ell))$ which concludes the proof.

□

D.3 Proof of Corollary 3.1

We can write $\log p(x_{1:n}|\psi^\ell;M_\ell) = -n\log n + n\log\widehat{L}(\psi^\ell)$. Moreover, denote by $S_m := \{j;\ M_j \text{ does not satisfy Assumption 1}\}$ the set of indices of the models that are misspecified and by S_m^c its complement in $\{1,2,\ldots,J\}$.

First, suppose that $\lim_{n\to\infty}P\left(\log m(x_{1:n};M_1) > \max_{j\neq 1}\log m(x_{1:n};M_j)\right) = 1$. Then, because $\max_{j\neq 1}\log m(x_{1:n};M_j) \geq \max_{j\neq 1;j\in S_m^c}\log m(x_{1:n};M_j)$,

$$P\left(\log m(x_{1:n};M_1) > \max_{j\neq 1}\log m(x_{1:n};M_j)\right)$$
$$\leq P\left(\log m(x_{1:n};M_1) > \max_{j\neq 1;\,j\in S_m^c}\log m(x_{1:n};M_j)\right) \quad \text{(D.8)}$$

which implies that the probability on the right hand side converges to 1 as $n \to \infty$. Then by Theorem 3.1, we necessarily have $(p_1 + d_{v_1}) < (p_j + d_{v_j})$, $\forall j \neq 1$, $j \in S_m^c$.

Next, suppose that $(p_1 + d_{v_1}) < (p_j + d_{v_j})$, $\forall j \neq 1$. Define the event

$$\mathcal{A} := \left\{ \max_{j \neq 1;\, j \in S_m^c} \log m(x_{1:n}; M_j) > \max_{j \neq 1; j \in S_m} \log m(x_{1:n}; M_j) \right\}.$$

Because all the models M_j with $j \in S_m^c$ have $K(P||Q^*(\psi^j)) = 0$ (because they are correctly specified) then $\lim_{n\to\infty} P(\mathcal{A}) = 1$ by Theorem 3.2. By the Law of Total Probability we can write

$$\begin{aligned} P\left(\log m(x_{1:n}; M_1) > \max_{j\neq 1} \log m(x_{1:n}; M_j)\right) &= P\Big(\log m(x_{1:n}; M_1) > \\ \max_{j\neq 1} \log m(x_{1:n}; M_j) \,\Big|\, \mathcal{A}\Big) P(\mathcal{A}) &+ P\left(\log m(x_{1:n}; M_1) > \max_{j\neq 1} \log m(x_{1:n}; M_j) \,\middle|\, \mathcal{A}^c\right) P(\mathcal{A}^c) \\ &\geq P\left(\log m(x_{1:n}; M_1) > \max_{j\neq 1} \log m(x_{1:n}; M_j) \,\middle|\, \mathcal{A}\right) P(\mathcal{A}) \\ &= P\left(\log m(x_{1:n}; M_1) > \max_{j\neq 1;\, j\in S_m^c} \log m(x_{1:n}; M_j)\right) P(\mathcal{A}) \quad \text{(D.9)} \end{aligned}$$

which converges to 1 by Theorem 3.1.

□

D.4 Technical Lemmas

Lemma D.1. *Let Assumptions 1, 5 and 6 hold for ψ^ℓ. Then,*

$$\begin{aligned} \log p(x_{1:n}|\widehat{\psi}^\ell; M_\ell) &= -n \log n - \frac{n}{2} \widehat{g}_\ell^{A\prime} \Delta_\ell^{-1} \widehat{g}_\ell^{A} + o_p(1) \\ &= -n \log n - \frac{\chi^2_{d-(p_\ell + d_{v_\ell})}}{2} + o_p(1) \quad \text{(D.10)} \end{aligned}$$

where $\chi^2_{d-(p_\ell+d_{v_\ell})}$ denotes a chi square distribution with $(d - (p_\ell + d_{v_\ell}))$ degrees of freedom.

Proof. See Appendix F.

□

Lemma D.2. *Let Assumptions 1, 2, 5, 6 and (2.15) hold for ψ^ℓ. Then,*

$$-\log \pi(\widehat{\psi}^\ell | x_{1:n}; M_\ell) = -\frac{(p_\ell + d_{v_\ell})}{2}[\log n - \log(2\pi)] + \frac{1}{2} \log |\Sigma_\ell| + o_p(1).$$

Proof. See Appendix F.

□

Lemma D.3. *Let M_ℓ be a misspecified model (that is, a model that does not satisfy Assumption 1) and let $g^A(x,\psi^\ell)$ and ψ^ℓ be the corresponding moment functions and parameters. Then, under Assumptions 5 (a)-(d), 3 and 7,*

$$\sup_{\psi^\ell\in\Psi^\ell}\left|\log\frac{\exp\{\widehat{\lambda}(\psi^\ell)'\widehat{g}^A(\psi^\ell)\}}{\frac{1}{n}\sum_{i=1}^n\exp\{\widehat{\lambda}(\psi^\ell)'g^A(x_i,\psi^\ell)\}}-\log\frac{\exp\{\lambda_\circ(\psi^\ell)'\mathbf{E}^P[g^A(x,\psi^\ell)]\}}{\mathbf{E}^P\left[\exp\{\lambda_\circ(\psi^\ell)'g^A(x,\psi^\ell)\}\right]}\right|\xrightarrow{p}0.$$

Proof. See Appendix F.

□

E Proof of Theorems C.1 and C.2

E.1 Proof of Theorem C.1

For a vector z and a scalar $\delta>0$ we denote by $B(z,\delta)$ the closed ball centred on z with radius δ. In this proof we use standard notation in empirical process theory: $\mathbb{P}_n:=\frac{1}{n}\sum_{i=1}^n\delta_{x_i}$ where δ_x is the Dirac measure at x, and $\mathbb{G}_nf:=\sqrt{n}(\mathbb{P}_nf-\mathbf{E}^Pf)$ for every function f. Moreover, we use Van der Vaart (2002, Theorem 6.16) which we report here for convenience (in a slightly modified version):

Theorem E.1 (Theorem 6.16 in Van der Vaart (2002))**.** *Let $\mathcal{F}_n$ be classes of measurable functions (that may change with n) such that $P(\widehat{f}_n\in\mathcal{F}_n)\to1$ and such that: (i) the bracketing integral $J_{[]}(\delta_n,\mathcal{F}_n,L_2(P))\to0$, for every $\delta_n\downarrow0$, and (ii) its envelope functions satisfy the Lindeberg condition. If $\mathbf{E}^P[(\widehat{f}_n-f_0)^2]\to0$ in probability for some $f_0\in L_2(P)$ then $\mathbb{G}_n(\widehat{f}_n-f_0)\xrightarrow{p}0$.*

The maps $x\mapsto l_{n,\psi_\circ}(x)$ and $x\mapsto l_{n,\psi_\circ+h/\sqrt{n}}(x)$ are random functions, that is, measurable functions that, for a fixed x, are functions of the observations $x_1,\dots,x_n$. By writing $\mathbb{P}_n[l_{n,\psi_\circ}]$ and $\mathbf{E}^P[l_{n,\psi_\circ}]$ we mean the (empirical and true) expectations of the function $x\mapsto l_{n,\psi_\circ}(x)$ with $(x_1,\dots,x_n)$ kept fixed (and similarly for $l_{n,\psi_\circ+h/\sqrt{n}}(x)$). Denote by $\dot{l}_{n,\psi_\circ}(x)$ and $\ddot{l}_{n,\psi_\circ}(x)$ the first and second order derivatives of the function $\psi\mapsto l_{n,\psi}(x)$ evaluated at $\psi_\circ$ (where we leave implicit the argument x).

A second order Taylor expansion of $l_{n,\psi_\circ+h/\sqrt{n}}$ around $h=0$, for a fixed x, gives

$$l_{n,\psi_\circ+h/\sqrt{n}} = l_{n,\psi_\circ} + \frac{h'}{\sqrt{n}}\dot{l}_{n,\psi_\circ} + \frac{1}{2n}h'\ddot{l}_{n,\psi_\circ}h + Rem. \tag{E.1}$$

By continuity of the map $\psi \mapsto l_{n,\psi}$ (which is valid under Assumption 5 *(c)* and by the Birge's maximum theorem and strict convexity of $\mathbb{P}_n \exp\{\lambda' g^A(x,\psi)\}$), the reminder term Rem is of order $o_p(\|h\|^2/n)$ since $\ddot{l}_{n,\psi_\circ} = \ddot{\mathfrak{L}}_{n,\psi_\circ} + o_p(1)$ and $\ddot{\mathfrak{L}}_{n,\psi_\circ} = O_p(1)$ under Assumptions 7 and 8 (see Schennach (2007, proof of Theorem 10)).

We consider the empirical process

$$\begin{aligned}&\mathbb{G}_n\left(\sqrt{n}(l_{n,\psi_\circ+h/\sqrt{n}} - l_{n,\psi_\circ}) - h'\dot{l}_{n,\psi_\circ}\right)\\ &\qquad := n\left(\mathbb{P}_n\left(l_{n,\psi_\circ+h/\sqrt{n}} - l_{n,\psi_\circ}\right) - \mathbf{E}^P\left(l_{n,\psi_\circ+h/\sqrt{n}} - l_{n,\psi_\circ}\right)\right) - h'\mathbb{G}_n\dot{l}_{n,\psi_\circ}\end{aligned} \tag{E.2}$$

where, according to the definition of random functions given just above:

$$\begin{aligned}\mathbb{P}_n[l_{n,\psi_\circ}] &= \frac{1}{n}\sum_{i=1}^n \widehat{\lambda}(\psi_\circ)' g^A(x_i,\psi_\circ) - \log \mathbb{P}_n\left[e^{\widehat{\lambda}(\psi_\circ)' g^A(x_j,\psi_\circ)}\right],\\ \text{and}\quad \mathbf{E}^P[l_{n,\psi_\circ}] &= \mathbf{E}^P\left[\widehat{\lambda}(\psi_\circ)' g^A(X,\psi_\circ)\right] - \log \mathbf{E}_n\left[e^{\widehat{\lambda}(\psi_\circ)' g^A(x_j,\psi_\circ)}\right]\end{aligned}$$

(and similarly for the other functions). The Markov's inequality and (E.1) imply that

$$P\left(\left|\mathbb{G}_n\left(\sqrt{n}(l_{n,\psi_\circ+h/\sqrt{n}} - l_{n,\psi_\circ}) - h'\dot{l}_{n,\psi_\circ}\right)\right| > \epsilon\right) \le \frac{1}{\epsilon\sqrt{n}}\mathbf{E}^P\left|\mathbb{G}_n\left(\frac{1}{2}h'\ddot{\mathfrak{L}}_{n,\psi_\circ}h\right) + o_p(\|h\|^2)\right|$$

that converges to zero since $\ddot{\mathfrak{L}}_{n,\psi_\circ} = O_p(1)$ under Assumptions 7 and 8. This shows that the sequence $\mathbb{G}_n\left(\sqrt{n}(l_{n,\psi_\circ+h/\sqrt{n}} - l_{n,\psi_\circ}) - h'\dot{l}_{n,\psi_\circ}\right)$ (seen as a stochastic process indexed by h) converges in probability and then (marginally) in distribution to zero. Next, we have to make this result uniform in h, that is, we have to show that the sequence of processes $\mathbb{G}_n\left(\sqrt{n}(l_{n,\psi_\circ+h/\sqrt{n}} - l_{n,\psi_\circ}) - h'\dot{l}_{n,\psi_\circ}\right)$ converges weakly in the space $l^\infty(h; h\in K)$ for a compact set $K \subset \mathbb{R}^p$. To show this we intend to apply Theorem E.1 given above. The proof consists of three steps where each step verifies the assumptions of Theorem E.1.

In the first step, we verify *(i)* in Theorem E.1 and we define a suitable class of functions that changes with the sample size n. Denote $\tau_n(\widehat{\lambda},\psi) := \mathbb{P}_n\left[e^{\widehat{\lambda}(\psi)' g^A(x,\psi)}\right]$, $\tau(\widehat{\lambda},\psi_\circ) := \mathbf{E}^P\left[e^{\widehat{\lambda}(\psi_\circ)' g^A(X,\psi_\circ)}\right]$ and consider the class of functions

$$\mathcal{F}_{1/\sqrt{n}} := \Big\{ \lambda_1' g^A(x,\psi) - \lambda_2' g^A(x,\psi_\circ) - \log(\tau_1/\tau_2);\ \tau_1 \in B\left(\tau(\lambda_1,\psi), \frac{1}{\sqrt{n}}\right);\ \lambda_1 \in B(\lambda_\circ(\psi), 1/\sqrt{n});$$
$$\tau_2 \in B\left(\tau(\lambda_2,\psi_\circ), \frac{1}{\sqrt{n}}\right);\ \lambda_2 \in B(\lambda_\circ(\psi_\circ), 1/\sqrt{n});\ \|\psi - \psi_\circ\| \leq \frac{1}{\sqrt{n}} \Big\}. \quad \text{(E.3)}$$

By Lemma E.1, the bracketing integral $J_{[]}(\delta_n, \sqrt{n}\mathcal{F}_{1/\sqrt{n}}, L_2(P))$ of the class $\sqrt{n}\mathcal{F}_{1/\sqrt{n}}$ converges to zero as $\delta_n \to 0$. This satisfies assumption (i) of Theorem E.1.

The second step of the proof consists in finding an envelope function for the class $\sqrt{n}\mathcal{F}_{1/\sqrt{n}}$ and in showing that it satisfies the Lindeberg condition (this is assumption *(ii)* in Theorem E.1). Lemma E.2 shows that $F_{2,n}(x)$ is an envelope function for the class $\mathcal{F}_{1/\sqrt{n}}$, where

$$F_{2,n}(x) := \frac{1}{\sqrt{n}}\left(C_{1,n} b(x) + C_{2,n} \|g^A(x,\psi_\circ)\| + \frac{1}{C_{4,n}} C_{3,n} \right)$$

where $b(x)$ is the function defined in Assumption 8 *(c)* and $C_{i,n} > 0$, $i = 1,\ldots,4$ are sequences of positive and bounded constants that depend on $\psi_\circ$ and $\lambda_\circ(\psi_\circ)$. It follows that $\sqrt{n}F_{2,n}(x)$ is an envelope function for the class $\sqrt{n}\mathcal{F}_{1/\sqrt{n}}$. The function $\sqrt{n}F_{2,n}(x)$ satisfies the Lindeberg condition if:

$$\begin{aligned} n\mathbf{E}^P[F_{2,n}(X)^2] &< \infty, \\ n\mathbf{E}^P\left[F_{2,n}(X)^2 1\{\sqrt{n}F_{2,n}(X) > \varepsilon\sqrt{n}\}\right] &\to 0, \qquad \text{for every } \varepsilon > 0. \end{aligned}$$

Under Assumption 8 *(c)*, $n\mathbf{E}^P[F_{2,n}(X)^2] < \infty$ holds true. The second Lindeberg condition is easily satisfied since $n\mathbf{E}^P\left[F_{2,n}(X)^2 1\{\sqrt{n}F_{2,n}(X) > \varepsilon\sqrt{n}\}\right] \leq n\sqrt{\mathbf{E}^P\left[F_{2,n}(X)^4\right]}\sqrt{P\left(F_{2,n} > \varepsilon\right)}$ which converges to zero for every $\varepsilon > 0$ because $P\left(F_{2,n} > \varepsilon\right) \to 0$ as $n \to \infty$ and, under Assumption 8 *(c)*, $\mathbf{E}^P\left[F_{2,n}^4(X)\right] = O(1)$.

Finally, we verify the last requirement of Theorem E.1. Remark that under Assumption 8 *(c)*

$$\mathbf{E}^P[h'\dot{l}_{n,\psi_\circ}]^2 = \mathbf{E}^P[h'\dot{\mathfrak{L}}_{n,\psi_\circ}\dot{\mathfrak{L}}'_{n,\psi_\circ}h] + o(\|h\|) < \infty \quad \text{(E.4)}$$

because $\mathbf{E}^P tr(\dot{\mathfrak{L}}_{n,\psi_\circ}\dot{\mathfrak{L}}'_{n,\psi_\circ})$ can be shown to be bounded under Assumption 8 *(c)* by following the last part of the proof of Schennach (2007, Theorem 10). Moreover, by a first order Taylor expansion of $l_{n,\psi_\circ+h/\sqrt{n}}$ around h, by continuity of the map $\psi \mapsto l_{n,\psi}$, Assumption 6 *(b)* and (E.4), we have: $\mathbf{E}^P\left[\sqrt{n}\left(l_{n,\psi_\circ+h/\sqrt{n}} - l_{n,\psi_\circ}\right) - h'\dot{l}_{n,\psi_\circ}\right]^2 = o(1)$. Therefore, by Theorem E.1 we conclude that

$$\mathbb{G}_n\left(\sqrt{n}(l_{n,\psi_\circ+h/\sqrt{n}} - l_{n,\psi_\circ}) - h'\dot{l}_{n,\psi_\circ}\right) \xrightarrow{p} 0$$

uniformly in h over a bounded set. Hence, by rewriting this as in (E.2) we see that

$$\sum_{i=1}^{n}(l_{n,\psi_\circ+h/\sqrt{n}}-l_{n,\psi_\circ})(x_i)-\mathbb{G}_n h'\dot{l}_{n,\psi_\circ}-n\mathbf{E}^P(l_{n,\psi_\circ+h/\sqrt{n}}-l_{n,\psi_\circ})=o_p(1). \tag{E.5}$$

By using (E.1) we obtain:

$$\begin{aligned}-\mathbb{G}_n h'\dot{l}_{n,\psi_\circ}-n\mathbf{E}^P(l_{n,\psi_\circ+h/\sqrt{n}}-l_{n,\psi_\circ}) &= -\sqrt{n}\mathbb{P}_n h'\dot{l}_{n,\psi_\circ}+\sqrt{n}\mathbf{E}^P h'\dot{l}_{n,\psi_\circ}-n\mathbf{E}^P(l_{n,\psi_\circ+h/\sqrt{n}}-l_{n,\psi_\circ})\\ &= -\sqrt{n}\mathbb{P}_n h'\dot{l}_{n,\psi_\circ}+\sqrt{n}\mathbf{E}^P[h'\dot{l}_{n,\psi_\circ}]-\sqrt{n}\mathbf{E}^P[h'\dot{l}_{n,\psi_\circ}]-\frac{1}{2}h'\mathbf{E}^P[\ddot{l}_{n,\psi_\circ}]h+o_p(1)\\ &= -\sqrt{n}\mathbb{P}_n h'\dot{l}_{n,\psi_\circ}-\frac{1}{2}h'\mathbf{E}^P[\ddot{l}_{n,\psi_\circ}]h+o_p(1) \quad \text{(E.6)}\end{aligned}$$

and by replacing this in (E.5) we get:

$$\sum_{i=1}^{n}(l_{n,\psi_\circ+h/\sqrt{n}}-l_{n,\psi_\circ})(x_i)-\frac{1}{\sqrt{n}}\sum_{i=1}^{n}h'\dot{l}_{n,\psi_\circ}(x_i)-\frac{1}{2}h'\mathbf{E}^P[\ddot{l}_{n,\psi_\circ}]h=o_p(1). \tag{E.7}$$

Because the $o_p(1)$ is uniform in h, this establishes (C.16) with $V_{\psi_\circ}=-\mathbf{E}^P[\ddot{\mathfrak{L}}_{n,\psi_\circ}]$ and $\Delta_{n,\psi_\circ}=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}V_{\psi_\circ}^{-1}\dot{\mathfrak{L}}_{n,\psi_\circ}(x_i)$ if $V_{\psi_\circ}$ is nonsingular since $\ddot{l}_{n,\psi_\circ}=\ddot{\mathfrak{L}}_{n,\psi_\circ}+o_p(1)$ and $\dot{l}_{n,\psi_\circ}=\dot{\mathfrak{L}}_{n,\psi_\circ}+o_p(1)$.

□

Lemma E.1. *Denote* $\tau_n(\widehat{\lambda},\psi)=\mathbb{P}_n\left[e^{\widehat{\lambda}(\psi)'g^A(x,\psi)}\right]$, $\tau(\widehat{\lambda},\psi)=\mathbf{E}^P\left[e^{\widehat{\lambda}(\psi)'g^A(X,\psi)}\right]$ *and consider the class of functions* $\mathcal{F}_{1/\sqrt{n}}$ *defined in* (E.3) *where* $\lambda_\circ$ *and* $\psi_\circ$ *are as defined in* (2.16)*. Under Assumptions 5 (a)-(d), 3, 7 (a), (c), and 8, the bracketing integral* $J_{[]}(\delta_n,\sqrt{n}\mathcal{F}_{1/\sqrt{n}},L_2(P))$ *of the class* $\sqrt{n}\mathcal{F}_{1/\sqrt{n}}$ *converges to zero as* $\delta_n\to 0$*:*

$$J_{[]}(\delta_n,\sqrt{n}\mathcal{F}_{1/\sqrt{n}},L_2(P))=\int_0^{\delta_n}\sqrt{\log N_{[]}(\varepsilon\|F\|_{P,2},\sqrt{n}\mathcal{F}_{1/\sqrt{n}},L_2(P))}d\varepsilon\to 0 \tag{E.8}$$

as $\delta_n\to 0$.

Proof. The class $\mathcal{F}_{1/\sqrt{n}}$ is indexed by a compact subset $B^\otimes:=B\left(\tau(\lambda_1,\psi),\frac{1}{\sqrt{n}}\right)\times B(\lambda_\circ(\psi),\frac{1}{\sqrt{n}})\times B\left(\tau(\lambda_2,\psi_\circ),\frac{1}{\sqrt{n}}\right)\times B(\lambda_\circ(\psi_\circ),\frac{1}{\sqrt{n}})\times B(\psi_\circ,\frac{1}{\sqrt{n}})\subset\mathbb{R}^{p+d_v+2d+2}$. Theorem 10 in Schennach (2007), which is valid under Assumptions 5 *(a)-(c)*, 7 *(b)-(c)* and 8 *(a)-(c)*, shows that $\tau_n(\widehat{\lambda},\psi)\xrightarrow{p}\tau(\widehat{\lambda},\psi)$ and $\widehat{\lambda}(\psi)\xrightarrow{p}\lambda_\circ(\psi)$ at the rate $1/\sqrt{n}$, hence we can write $\tau_n(\psi)\in B(\tau(\widehat{\lambda}(\psi),\psi),1/\sqrt{n})$ and $\widehat{\lambda}(\psi)\in B(\lambda_\circ(\psi),1/\sqrt{n})$ with probability approaching 1. There-

fore, $P\left(l_{n,\psi_\circ+h/\sqrt{n}} - l_{n,\psi_\circ} \in \mathcal{F}_{1/\sqrt{n}}\right) \to 1$. For every $f_a, f_b \in \mathcal{F}_{1/\sqrt{n}}$:

$$
\begin{aligned}
&|f_a(x) - f_b(x)| = \\
&\left|\lambda'_{1,a} g^A(x,\psi_a) - \lambda'_{2,a} g^A(x,\psi_\circ) - \log(\tau_{1,a}/\tau_{2,a}) - \lambda'_{1,b} g^A(x,\psi_b) + \lambda'_{2,b} g^A(x,\psi_\circ) + \log(\tau_{1,b}/\tau_{2,b})\right| \\
&\quad \leq \|\lambda_{1,a}\| \, \|g^A(x,\psi_a) - g^A(x,\psi_b)\| + \|\lambda_{1,a} - \lambda_{1,b}\| \, \|g^A(x,\psi_b)\| + \|\lambda_{2,a} - \lambda_{2,b}\| \, \|g^A(x,\psi_\circ)\| \\
&\qquad\qquad - |\log \tau_{1,a} - \log \tau_{1,b}| + |\log \tau_{2,a} - \log \tau_{2,b}|.
\end{aligned}
$$

The following results hold by compactness of $B^\otimes$ and continuity of $\psi \mapsto g^A(x,\psi)$ (under Assumption 5 *(c)*): *(i)* $\|\lambda_{1,a}\| \leq C$ for a generic constant $C > 0$ since $|\lambda_\circ(\psi)| < \infty$; *(ii)* $\|g^A(x,\psi_a) - g^A(x,\psi_b)\| \leq \|\partial g^A(x,\overline{\psi})/\partial\psi\| \, \|\psi_a - \psi_b\|$ for some $\overline{\psi}$ on the line joining ψ_a and ψ_b by the Mean Value theorem; *(iii)* $\|\lambda_{1,a} - \lambda_{1,b}\| \leq 2/\sqrt{n}$ because $\lambda_{1,a}, \lambda_{1,b} \in B(\lambda_\circ(\psi), 1/\sqrt{n})$; *(iv)* $|\log \tau_{1,a} - \log \tau_{1,b}| \leq |\tau_{1,a} - \tau_{1,b}|/\overline{\tau}_1$ for some $\overline{\tau}_1 > 0$ between $\tau_{1,a}$ and $\tau_{1,b}$ by the Mean Value Theorem (and similarly for $|\log \tau_{2,a} - \log \tau_{2,b}|$). By using all these results:

$$
\begin{aligned}
|f_a(x) - f_b(x)| &\leq \left[C\|\partial g^A(x,\overline{\psi})/\partial\psi\| + 2\left(\|g^A(x,\psi_b)\| + \|g^A(x,\psi_\circ)\| + \overline{\tau}_1^{-1} + \overline{\tau}_2^{-1}\right)\right] \frac{1}{\sqrt{n}} \\
&\leq \left[(C + 2/\sqrt{n})b(x) + 2\left(\|g^A(x,\psi_\circ)\| + \overline{\tau}_1^{-1} + \overline{\tau}_2^{-1}\right)\right] \frac{1}{\sqrt{n}} =: F(x)\frac{1}{\sqrt{n}}
\end{aligned}
$$

where the second inequality follows by the Mean Value Theorem applied to $\|g^A(x,\psi_\circ)\|$ and Assumption 8 *(c)* that implies that $\|\partial g^A(x,\psi)/\partial\psi\| \leq b(x)$ for every $\psi \in B(\psi_\circ, 1/\sqrt{n})$. Remark that under Assumptions 5 *(d)* and 8 *(c)*:

$$
\mathbf{E}^P[F(x)^2] \leq 2(C + 2/\sqrt{n})^2 \mathbf{E}^P[b(X)^2] + 16\mathbf{E}\|g^A(x,\psi_\circ)\|^2 + 16(\overline{\tau}_1^{-1} + \overline{\tau}_2^{-1}) < \infty.
$$

Therefore, by example 19.7 in Van der Vaart (1998), there exists a constant K independent of ε and n such that the bracketing numbers of the class of functions $\mathcal{F}_{1/\sqrt{n}}$ satisfy

$$
N_{[]}\left(\varepsilon \frac{1}{\sqrt{n}} \|F\|_{P,2}, \mathcal{F}_{1/\sqrt{n}}, L_2(P)\right) \leq K \left(\frac{diam\widetilde{B}}{\frac{\varepsilon}{\sqrt{n}}}\right)^{p+d_v+2d+2}, \qquad 0 < \varepsilon < \frac{1}{\sqrt{n}}
$$

where $L_2(P)$ denotes the L_2 space of square integrable functions with respect to P and $\|\cdot\|_{P,2}$ denotes the norm in this space. Remark that $diam\widetilde{B} = 2/\sqrt{n}$ so that $\left(\frac{diam\widetilde{B}}{\frac{\varepsilon}{\sqrt{n}}}\right)^{p+d_v+2d+2} = (2/\varepsilon)^{p+d_v+2d+2}$. Then, the bracketing numbers of the class of functions $\sqrt{n}\mathcal{F}_{1/\sqrt{n}}$ satisfy

$$
N_{[]}(\varepsilon\|F\|_{P,2}, \sqrt{n}\mathcal{F}_{1/\sqrt{n}}, L_2(P)) \leq K \, (2/\varepsilon)^{p+d_v+2d+2}, \qquad 0 < \varepsilon < \frac{1}{\sqrt{n}}. \tag{E.9}
$$

Let us compute the bracketing integral of the class $\sqrt{n}\mathcal{F}_{1/\sqrt{n}}$:

$$\begin{aligned} J_{[]}(\delta_n, \sqrt{n}\mathcal{F}_{1/\sqrt{n}}, L_2(P)) &= \int_0^{\delta_n} \sqrt{\log N_{[]}(\varepsilon \|F\|_{P,2}, \sqrt{n}\mathcal{F}_{1/\sqrt{n}}, L_2(P))} d\varepsilon \\ &\leq \int_0^{\delta_n} \sqrt{\log K + \log (2/\varepsilon)^{p+d_v+2d+2}} d\varepsilon \to 0 \quad \text{as } \delta_n \to 0 \end{aligned}$$

where the last inequality follows from (E.9) and proves the lemma.

□

Lemma E.2. *Denote* $\tau_n(\widehat{\lambda}, \psi) = \mathbb{P}_n \left[e^{\widehat{\lambda}(\psi)' g^A(x,\psi)} \right]$, $\tau(\widehat{\lambda}, \psi_\circ) = \mathbf{E}^P \left[e^{\widehat{\lambda}(\psi_\circ)' g^A(X,\psi_\circ)} \right]$ *and consider the class of functions* $\mathcal{F}_{1/\sqrt{n}}$ *defined in* (E.3) *where* $\lambda_\circ$ *and* $\psi_\circ$ *are as defined in* (2.16). *Under Assumptions 5 (a)-(c), 3, 7, and 8, the function*

$$F_{2,n}(x) := \frac{1}{\sqrt{n}} \left(C_{1,n} b(x) + C_{2,n} \|g^A(x, \psi_\circ)\| + \frac{1}{C_{4,n}} C_{3,n} \right).$$

is an envelope function for the class $\mathcal{F}_{1/\sqrt{n}}$, *where* $b(x)$ *is the function defined in Assumption 8 (c) and* $C_{i,n} > 0$, $i = 1, \ldots, 4$ *are sequences of positive and bounded constants that depend on* $\psi_\circ$ *and* $\lambda_\circ(\psi_\circ)$.

Proof. First, remark that every $f \in \mathcal{F}_{1/\sqrt{n}}$ satisfies

$$\begin{aligned} |f(x)| \leq & \sup_{\|\psi - \psi_\circ\| \leq n^{-1/2}} \sup_{\lambda \in B(\lambda_\circ(\psi), 1/\sqrt{n})} \|\lambda\| \|g^A(x,\psi) - g^A(x,\psi_\circ)\| \\ & + \sup_{\|\psi - \psi_\circ\| \leq n^{-1/2}} \sup_{\lambda_1 \in B(\lambda_\circ(\psi), 1/\sqrt{n})} \sup_{\lambda_2 \in B(\lambda_\circ(\psi_\circ), 1/\sqrt{n})} \|\lambda_1 - \lambda_2\| \|g^A(x,\psi_\circ)\| + \frac{|\tau_1 - \tau_2|}{\tau_2} \end{aligned} \quad \text{(E.10)}$$

for $\tau_1 \in B(\tau(\lambda_1, \psi), 1/\sqrt{n})$, $\tau_2 \in B(\tau(\lambda_2, \psi_\circ), 1/\sqrt{n})$, $\lambda_1 \in B(\lambda_\circ(\psi), 1/\sqrt{n})$ and $\lambda_2 \in B(\lambda_\circ(\psi_\circ), 1/\sqrt{n})$ since $\log \tau_2/\tau_1 \leq (\tau_2 - \tau_1)/\tau_1$. Next, we bound each of these terms separately.

Let $\sup_{\|\psi - \psi_\circ\| \leq n^{-1/2}} \sup_{\lambda \in B(\lambda_\circ(\psi), 1/\sqrt{n})} \|\lambda\| =: C_{1,n} < \infty$. Remark that by the implicit function theorem for vector valued functions applied to the first order condition for $\lambda_\circ$, the function $\psi \mapsto \lambda_\circ(\psi)$ is continuously differentiable in a neighborhood of $\psi_\circ$. Therefore, for every $\lambda_1 \in B(\lambda_\circ(\psi), 1/\sqrt{n})$, $\lambda_2 \in B(\lambda_\circ(\psi_\circ), 1/\sqrt{n})$ with $\psi \in B(\psi_\circ, 1/\sqrt{n})$, by the triangular inequality and the continuity of $\lambda_\circ$, there exists a N such that $\forall n \geq N$

$$\|\lambda_1 - \lambda_2\| \leq \|\lambda_1 - \lambda_\circ(\psi)\| + \|\lambda_\circ(\psi) - \lambda_\circ(\psi_\circ)\| + \|\lambda_2 - \lambda_\circ(\psi_\circ)\|$$

$$
\begin{aligned}
&\leq \frac{2}{\sqrt{n}} + \left\| \frac{\partial \lambda_\circ(\overline{\psi})}{\partial \psi'} \right\| \|\psi - \psi_\circ\| \\
&\leq \frac{1}{\sqrt{n}} \left(2 + \sup_{\psi \in B(\psi_\circ, \frac{1}{\sqrt{n}})} \left\| \frac{\partial \lambda_\circ(\psi)}{\partial \psi'} \right\| \right) =: \frac{1}{\sqrt{n}} C_{2,n}
\end{aligned}
$$

where the second inequality follows from the Mean Value theorem with $\overline{\psi}$ being on the line joining ψ and $\psi_\circ$. Remark that $C_{2,n}$ is a sequence of positive and bounded constants which depends on $\psi_\circ$. By similar arguments we have that for every $\tau_1 \in B\left(\tau(\lambda_1, \psi), \frac{1}{\sqrt{n}}\right)$ and $\tau_2 \in B\left(\tau(\lambda_2, \psi_\circ), \frac{1}{\sqrt{n}}\right)$ with $\lambda_1 \in B(\lambda_\circ(\psi), 1/\sqrt{n})$, $\lambda_2 \in B(\lambda_\circ(\psi_\circ), 1/\sqrt{n})$ and $\psi \in B(\psi_\circ, 1/\sqrt{n})$:

$$
\begin{aligned}
|\tau_1 - \tau_2| &\leq |\tau_1 - \tau(\lambda_1, \psi)| + |\tau(\lambda_1, \psi) - \tau(\lambda_2, \psi_\circ)| + |\tau_2 - \tau(\lambda_2, \psi_\circ)| \\
&\leq \frac{2}{\sqrt{n}} + \mathbf{E}^P \left[e^{\lambda_1' g^A(X,\psi)} \|\lambda_1\| \left\| \frac{\partial g^A(X, \tilde{\psi})}{\partial \psi} \right\| \right] \|\psi - \psi_\circ\| \\
&\quad + \mathbf{E}^P [e^{\tilde{\lambda}' g^A(X,\psi_\circ)} \|g^A(X, \psi_\circ)\|] \|\lambda_1 - \lambda_2\| \\
&\leq \frac{2}{\sqrt{n}} + \frac{C_{1,n}}{\sqrt{n}} \mathbf{E}^P \left[\sup_{\|\psi - \psi_\circ\| \leq 1/\sqrt{n}} \sup_{\lambda_1 \in B(\lambda_\circ(\psi), 1/\sqrt{n})} e^{\lambda_1' g^A(X,\psi)} b(X) \right] \\
&\quad + \mathbf{E}^P \left[\sup_{t \in (0,1)} \sup_{\psi \in B(\psi_\circ, 1/\sqrt{n})} \sup_{\lambda_1 \in B(\lambda_\circ(\psi), 1/\sqrt{n})} \sup_{\lambda_2 \in B(\lambda_\circ(\psi_\circ), 1/\sqrt{n})} e^{\tilde{\lambda}' g^A(X,\psi_\circ)} b(X) \right] \frac{C_{2,n}}{\sqrt{n}} \\
&=: \frac{1}{\sqrt{n}} C_{3,n}
\end{aligned}
$$

where $\tilde{\psi}$ is between ψ and $\psi_\circ$, $\tilde{\lambda} = t\lambda_1 + (1-t)\lambda_2$, $t \in (0,1)$ and $C_{3,n}$ is a sequence of positive and bounded constants by Assumption 8 *(c)* which depends on $\psi_\circ$ and $\lambda_\circ$. Therefore, by this result and since $\log \tau_2/\tau_1 \leq (\tau_2 - \tau_1)/\tau_1$, τ_1 is uniformly bounded away from zero over a compact set: $|\log \tau_2/\tau_1| \leq \frac{1}{C_{4,n}\sqrt{n}} C_{3,n}$ for some strictly positive constant $0 < C_{4,n} < \infty$ that lower bounds τ_2 uniformly. Therefore, by replacing everything in (E.10) we get, $\forall f \in \mathcal{F}_{1/\sqrt{n}}$:

$$
\begin{aligned}
|f(x)| &\leq C_{1,n} \left\| \partial g^A(x, \tilde{\psi}) / \partial \psi \right\| \|\psi - \psi_\circ\| + \frac{1}{\sqrt{n}} C_{2,n} \|g^A(x, \psi_\circ)\| + \frac{C_{3,n}}{C_{4,n}\sqrt{n}} \\
&\leq \frac{1}{\sqrt{n}} \left(C_{1,n} b(x) + C_{2,n} \|g^A(x, \psi_\circ)\| + \frac{1}{C_{4,n}} C_{3,n} \right) =: F_{2,n}(x)
\end{aligned}
$$

where in the last inequality we have used Assumption 8 *(c)* that holds for every ψ in $B(\psi_\circ, 1/\sqrt{n})$. Therefore, $F_{2,n}(x)$ is an envelope function for the class $\mathcal{F}_{1/\sqrt{n}}$ and this concludes the proof.

□

E.2 Proof of Theorem C.2

Define the events $A_{n,1} := \left\{\sup_{\psi\in\Psi_n^c} \frac{1}{n}\sum_{i=1}^n (l_{n,\psi}(x_i) - l_{n,\psi_\circ}(x_i)) \le -CM_n^2/n\right\}$ and

$$A_{n,2} := \left\{\int_\Psi \frac{p(x_{1:n}|\psi)}{p(x_{1:n}|\psi_\circ)}\pi(\psi)d(\psi) \ge e^{-CM_n^2/2}\right\}.$$

By (C.17), $P(A_{n,1}^c) \to 0$ and by Lemma E.3 below, $P(A_{n,2}^c) \to 0$. Therefore,

$$\begin{aligned}
\mathbf{E}^P\left[\pi\left(\sqrt{n}\|\psi-\psi_*\| > M_n\middle|\, x_{1:n}\right)\right] &\le \mathbf{E}^P\left[\pi\left(\sqrt{n}\|\psi-\psi_*\| > M_n\middle|\, x_{1:n}\right)\middle|\, A_{n,1}\cap A_{n,2}\right] \\
&\quad \times P(A_{n,1}\cap A_{n,2}) + o(1) \\
&= \mathbf{E}^P\left[\left.\frac{\int_{\Psi_n^c}\frac{p(x_{1:n}|\psi)}{p(x_{1:n}|\psi_*)}\pi(\psi)d\psi}{\int_\Psi \frac{p(x_{1:n}|\psi)}{p(x_{1:n}|\psi_*)}\pi(\psi)d\psi}\right| A_{n,1}\cap A_{n,2}\right] P(A_{n,1}\cap A_{n,2}) + o(1) \\
&\le e^{-CM_n^2}\pi(\Psi_n^c)\mathbf{E}^P\left[\left.\left(\int_\Psi \frac{p(x_{1:n}|\psi)}{p(x_{1:n}|\psi_*)}\pi(\psi)d\psi\right)^{-1}\right| A_{n,1}\cap A_{n,2}\right] P(A_{n,1}\cap A_{n,2}) + o(1) \\
&\le e^{-CM_n^2}e^{CM_n^2/2}\pi(\Psi_n^c)P(A_{n,1}\cap A_{n,2}) + o(1) = o(1) \quad \text{(E.11)}
\end{aligned}$$

which proves the result of the theorem.

□

Lemma E.3. *Assume that the stochastic LAN expansion (C.16) holds for $\psi_\circ$ defined in (2.16) and that Assumptions 2 (a), 3, 4 and 8 are satisfied. Then,*

$$P\left(\int_\Psi \frac{p(x_{1:n}|\psi)}{p(x_{1:n}|\psi_\circ)}\pi(\psi)d\psi < a_n\right) \to 0 \qquad \text{(E.12)}$$

for every sequence $a_n \to 0$.

Proof. For a given $M > 0$ define $\mathfrak{C} = \{h \in \mathbb{R}^{d_v+p} : \|h\| \le M\}$. Denote by $h \mapsto Rem(h)$ the remaining term in (C.16) and remark that $\sup_{h\in\mathfrak{C}} Rem(h) \xrightarrow{p} 0$ by (C.16) and compactness of $\mathfrak{C}$. Therefore, for a sequence κ_n that converges to zero slowly enough, the event $B_n := \{\sup_{h\in\mathfrak{C}} Rem(h) \le \kappa_n\}$ has probability $P(B_n) \to 1$. Let $K_n \to \infty$. By considering the local parameter $h = \sqrt{n}(\psi - \psi_\circ)$ and by denoting by π^h both its prior distribution and prior

Lebesgue density (under Assumption 2 *(a)*), we upper bound the probability in (E.12) as follows:

$$P\left(\int_{\Psi}\frac{p(x_{1:n}|\psi)}{p(x_{1:n}|\psi_\circ)}\pi(\psi)d\psi<e^{-K_n^2}\right)\leq P\left(\int_{\mathfrak{C}}\frac{p(x_{1:n}|\psi_\circ+h/\sqrt{n})}{p(x_{1:n}|\psi_\circ)}\pi^h(h)dh<e^{-K_n^2}\right)$$
$$=P\left(\left\{\int_{\mathfrak{C}}e^{\sum_{i=1}^n(l_{n,\psi_\circ+h/\sqrt{n}}-l_{n,\psi_\circ})}\pi^h(h)dh<e^{-K_n^2}\right\}\cap B_n\right)+o_p(1). \quad \text{(E.13)}$$

By replacing the LAN expansion (C.16) and by noting that for n sufficiently large, $\kappa_n\leq\frac{1}{2}K_n^2$ on B_n and $\sup_{h\in\mathfrak{C}}h'V_{\psi_\circ}h\leq\sup_{h\in\mathfrak{C}}\|h\|^2\|V_{\psi_\circ}\|\leq M^2\|V_{\psi_\circ}\|\leq\kappa_n\leq\frac{1}{2}K_n^2$ (since $M^2\|V_{\psi_\circ}\|$ has the same order as $Rem(h)$ and where $\|V_{\psi_\circ}\|$ denotes the operator norm) we obtain:

$$P\left(\int_{\Psi}\frac{p(x_{1:n}|\psi)}{p(x_{1:n}|\psi_\circ)}\pi(\psi)d\psi<e^{-K_n^2}\right)\leq P\left(\int_{\mathfrak{C}}e^{h'V_{\psi_\circ}\Delta_{n,\psi_\circ}}\pi^h(h)dh<e^{-3K_n^2/4}\right)+o_p(1)$$
$$=P\left(\int_{\mathfrak{C}}e^{h'V_{\psi_\circ}\Delta_{n,\psi_\circ}}\pi^h(h|\mathfrak{C})dh<e^{-\log\pi^h(\mathfrak{C})}e^{-3K_n^2/4}\right)+o_p(1)$$
$$\leq P\left(\exp\left\{\int_{\mathfrak{C}}h'V_{\psi_\circ}\Delta_{n,\psi_\circ}\pi^h(h|\mathfrak{C})dh\right\}<e^{K_n^2/8}e^{-3K_n^2/4}\right)+o_p(1)$$
$$\leq P\left(\int_{\mathfrak{C}}h'V_{\psi_\circ}\Delta_{n,\psi_\circ}\pi^h(h|\mathfrak{C})dh<-5K_n^2/8\right)+o_p(1)$$
$$\leq\frac{64}{25K_n^4}E^P\left(\int_{\mathfrak{C}}\left(h'V_{\psi_\circ}\Delta_{n,\psi_\circ}\right)^2\pi^h(h|\mathfrak{C})dh\right)+o_p(1)\to 0 \quad \text{(E.14)}$$

where in the third line we have used that, for n large enough, $-\log\pi^h(\mathfrak{C})\leq K_n^2/8$ and the Jensen's inequality. In the last line we have used the Markov's inequality and then the Jensen's inequality. The result follows by (E.4) and Assumption 4.

□

F Proof of the technical Lemmas for the proof of Theorems 3.1-3.2

For a vector z and a scalar $\delta>0$ we denote by $B(z,\delta)$ the closed ball centred on z with radius δ. When the Mean Value theorem is applied to a vector of functions it must be understood that it is applied componentwise.

F.1 Proof of Lemma D.1

Let us consider the expression for the likelihood given in (2.6)-(2.9) and evaluated at $\widehat{\psi}^{\ell}$:

$$\log p(x_{1:n}|\widehat{\psi}^{\ell}; M_{\ell}) \;=\; -n\log n \;+\; \sum_{i=1}^{n} \widehat{\lambda}(\widehat{\psi}^{\ell})' g^{A}(x_i, \widehat{\psi}^{\ell}) \;-\; n\log \frac{1}{n}\sum_{j=1}^{n} e^{\widehat{\lambda}(\widehat{\psi}^{\ell})' g^{A}(x_j, \widehat{\psi}^{\ell})}. \quad \text{(F.1)}$$

To shorten notation, in the rest of this proof we eliminate the superscripts and subscripts and just write: g, $\widehat{\psi}$ instead of g^{A} and $\widehat{\psi}^{\ell}$.

Let $\tilde{\lambda}$ be on the line joining 0 and $\widehat{\lambda}(\widehat{\psi})$, then a second order Taylor expansion around $\widehat{\lambda} = 0$ gives

$$\frac{1}{n}\sum_{j=1}^{n} e^{\widehat{\lambda}(\widehat{\psi})' g(x_j, \widehat{\psi})} = 1 + \frac{1}{n}\sum_{i=1}^{n} \widehat{\lambda}(\widehat{\psi})' g\left(x_i, \widehat{\psi}\right) + \frac{1}{2}\widehat{\lambda}(\widehat{\psi})' \frac{1}{n}\sum_{j=1}^{n} e^{\tilde{\lambda}' g(x_j, \widehat{\psi})} g\left(x_j, \widehat{\psi}\right) g\left(x_j, \widehat{\psi}\right)' \widehat{\lambda}(\widehat{\psi}). \quad \text{(F.2)}$$

Under Assumption 5 and because $\tilde{\lambda} = \mathcal{O}_p(n^{-1/2}) = o_p(n^{-\zeta})$ for any $\zeta < 1/2$ (since by Newey and Smith (2004, Lemma A.2) $\widehat{\lambda}(\widehat{\psi}) = \mathcal{O}_p(n^{-1/2}) = o_p(n^{-\zeta})$ for any $\zeta < 1/2$ and $\tilde{\lambda}$ is between 0 and $\widehat{\lambda}(\widehat{\psi})$) we can apply Newey and Smith (2004, Lemma A.1) that implies: $\max_{1\leq i\leq n} |\tilde{\lambda}' g(x_i, \widehat{\psi})| \xrightarrow{p} 0$. Therefore, $\max_{1\leq i\leq n} \left| -e^{\tilde{\lambda}' g(x_i, \widehat{\psi})} + 1 \right| \xrightarrow{p} 0$ which in turn implies:

$$\frac{1}{n}\sum_{j=1}^{n} e^{\tilde{\lambda}' g(x_j, \widehat{\psi})} g\left(x_j, \widehat{\psi}\right) g\left(x_j, \widehat{\psi}\right)' \xrightarrow{p} \Delta. \quad \text{(F.3)}$$

By replacing this in (F.2) we obtain:

$$\frac{1}{n}\sum_{j=1}^{n} e^{\widehat{\lambda}(\widehat{\psi})' g(x_j, \widehat{\psi})} = 1 + \frac{1}{n}\sum_{i=1}^{n} \widehat{\lambda}(\widehat{\psi})' g\left(x_i, \widehat{\psi}\right) + \frac{1}{2}\widehat{\lambda}(\widehat{\psi})' \Delta \widehat{\lambda}(\widehat{\psi}) + o_p\left(n^{-1}\right). \quad \text{(F.4)}$$

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 (F.4) in it to obtain:

$$\log\left(\frac{1}{n}\sum_{j=1}^{n} e^{\widehat{\lambda}(\widehat{\psi})' g(x_j, \widehat{\psi})}\right) = \frac{1}{n}\sum_{i=1}^{n} \widehat{\lambda}(\widehat{\psi})' g\left(x_i, \widehat{\psi}\right) + \frac{1}{2}\widehat{\lambda}(\widehat{\psi})' \Delta \widehat{\lambda}(\widehat{\psi}) + o_p\left(n^{-1}\right) + o\left(\left|\frac{1}{n}\sum_{j=1}^{n} e^{\widehat{\lambda}(\widehat{\psi})' g(x_j, \widehat{\psi})} - 1\right|\right). \quad \text{(F.5)}$$

In order to simplify (F.5) further and to find the rate of the last term in the right hand side of (F.5) we approximate $\widehat{g}(\psi) := \frac{1}{n}\sum_{i=1}^{n} g(x_i, \psi)$ as follows:

$$\begin{aligned}
\widehat{g}(\widehat{\psi}) =& \widehat{g}(\psi_*) + \frac{1}{n}\sum_{i=1}^{n} \frac{\partial g(x_i, \psi_*)}{\partial \psi'}(\widehat{\psi} - \psi_*) + o(\|(\widehat{\psi} - \psi_*)\|) \\
=& \widehat{g}(\psi_*) + \Gamma_\ell(\widehat{\psi} - \psi_*) + o_p(n^{-1/2}) && \text{(F.6)} \\
=& \widehat{g}(\psi_*) - \Gamma_\ell H \widehat{g}(\psi_*) + o_p(n^{-1/2}) && \text{(F.7)} \\
=& -\Delta_\ell \widehat{\lambda}(\widehat{\psi}) + o_p(n^{-1/2}) && \text{(F.8)}
\end{aligned}$$

where to get (F.7) we have used the fact that, under Assumptions 1, 5 and 6: $\sqrt{n}\left(\widehat{\psi} - \psi_*\right) = -H\sqrt{n}\widehat{g}(\psi_*) + o_p(1)$ with $H := (\Gamma_\ell'\Delta_\ell^{-1}\Gamma_\ell)^{-1}\Gamma_\ell'\Delta_\ell^{-1}$ (see Schennach (2007, Proof of Theorem 3)) and to get (F.6) we have used the fact that $\|\frac{1}{n}\sum_{i=1}^{n}\frac{\partial g(x_i,\psi_*)}{\partial \psi'} - \Gamma_\ell\| = \mathcal{O}_p(n^{-1/2})$ under Assumption 6 (b) by the Markov's inequality. Finally, (F.8) is obtained by using the fact that $I - \Gamma_\ell H = \Delta_\ell \Phi_\ell$ where $\Phi_\ell := \Delta_\ell^{-1} - \Delta_\ell^{-1}\Gamma_\ell\Sigma_\ell\Gamma_\ell'\Delta_\ell^{-1}$ and $\Sigma_\ell = (\Gamma_\ell'\Delta_\ell^{-1}\Gamma_\ell)^{-1}$, and that, under Assumptions 1, 5 and 6, $\sqrt{n}\widehat{\lambda}(\widehat{\psi}) = -\Phi_\ell\sqrt{n}\widehat{g}(\psi_*) + o_p(1)$ (see Schennach (2007, Proof of Theorem 3)). By substituting this result in (F.4) we obtain:

$$\begin{aligned}
\left|\frac{1}{n}\sum_{j=1}^{n} e^{\widehat{\lambda}(\widehat{\psi})' g(x_j,\widehat{\psi})} - 1\right| &= \left|-\widehat{\lambda}(\widehat{\psi})\Delta_\ell\widehat{\lambda}(\widehat{\psi}) + \frac{1}{2}\widehat{\lambda}(\widehat{\psi})'\Delta_\ell\widehat{\lambda}(\widehat{\psi})\right| + o_p(n^{-1}) \\
&= \left|-\frac{1}{2}\widehat{\lambda}(\widehat{\psi})'\Delta_\ell\widehat{\lambda}(\widehat{\psi})\right| + o_p(n^{-1})
\end{aligned}$$

which is $\mathcal{O}_p(n^{-1})$ since $\left\|\widehat{\lambda}(\widehat{\psi})\right\|^2 = \mathcal{O}_p(n^{-1})$. By replacing this result and (F.8) in (F.5) we obtain:

$$\log\left(\frac{1}{n}\sum_{j=1}^{n} e^{\widehat{\lambda}(\widehat{\psi})' g(x_j,\widehat{\psi})}\right) = \widehat{\lambda}(\widehat{\psi})'\widehat{g}(\widehat{\psi}) + \frac{1}{2}\widehat{g}(\widehat{\psi})'\Delta_\ell^{-1}\widehat{g}(\widehat{\psi}) + o_p\left(n^{-1}\right). \quad \text{(F.9)}$$

Next, we replace (F.9) in (F.1) to get:

$$\begin{aligned}
\log p(x_{1:n}|\widehat{\psi}; M_\ell) = -n\log n + \sum_{i=1}^{n}\widehat{\lambda}(\widehat{\psi})' g(x_i,\widehat{\psi}) - \sum_{i=1}^{n}\widehat{\lambda}(\widehat{\psi})' g\left(x_i,\widehat{\psi}\right) & \\
- n\frac{1}{2}\widehat{g}(\widehat{\psi})'\Delta_\ell^{-1}\widehat{g}(\widehat{\psi}) + o_p(1) & \\
= -n\log n - \frac{\chi^2_{d-(p_\ell+d_{v_\ell})}}{2} + o_p(1) & \text{(F.10)}
\end{aligned}$$

where the last equality follows from standard arguments as in Hansen (1982) that show that $n\widehat{g}(\widehat{\psi})'\Delta_\ell^{-1}\widehat{g}(\widehat{\psi}) \to^d \chi^2_{d-(p_\ell+d_{v_\ell})}$.

□

F.2 Proof of Lemma D.2

Result (2.15) and consistency of the ETEL estimator $\widehat{\psi}^\ell$ (which is guaranteed under Assumptions 1, 5 and 6, see Schennach (2007, Theorem 3)) implies that the posterior $\pi(\psi^\ell|x_{1:n}; M_\ell)$ of ψ^ℓ converges in total variation towards a $\mathcal{N}_{\widehat{\psi}^\ell, n^{-1}\Sigma_\ell}$ distribution, where $\Sigma_\ell = (\Gamma_\ell'\Delta^{-1}\Gamma_\ell)^{-1}$. Hence, the negative logarithm of the posterior density evaluated at $\widehat{\psi}^\ell$ is

$$-\log \pi(\widehat{\psi}^\ell|x_{1:n}; M_\ell) = -\frac{(p_\ell + d_\ell)}{2}[\log n - \log(2\pi)] + \frac{1}{2}\log|\Sigma_\ell| + o_p(1).$$

□

F.3 Proof of Lemma D.3

In the proof we eliminate the superscript ℓ for simplicity. The proof proceeds in two steps. In the first step we show uniform convergence of $\widehat{g}^A(\psi)$ and $\widehat{\lambda}(\psi)$. By the uniform strong Law of Large Numbers, which is valid under Assumptions 5 *(a)-(b)* and 7 *(a)* (see *e.g.* Newey and McFadden (1994, Lemma 2.4)), it holds:

$$\sup_{\psi\in\Psi} \left\|\widehat{g}^A(\psi) - \mathbf{E}^P[g^A(x,\psi)]\right\| \xrightarrow{p} 0. \tag{F.11}$$

Let $\overline{\lambda}(\psi) = \arg\min_{\lambda\in\Lambda(\psi)} \frac{1}{n}\sum_{i=1}^n \exp\{\lambda' g^A(x_i,\psi)\}$. Schennach (2007, page 668) shows that $\sup_{\psi\in\Psi} \|\overline{\lambda}(\psi) - \lambda_\circ(\psi)\| \xrightarrow{p} 0$ (under Assumptions 5 *(a)-(b)* and 7). Moreover, if $\overline{\lambda}(\psi)$ lies in the interior of $\Lambda(\psi)$ then the minimum of $\frac{1}{n}\sum_{i=1}^n \exp\{\lambda' g^A(x_i,\psi)\}$ is unique by strict convexity of the latter function. As $\overline{\lambda}(\psi) \xrightarrow{p} \lambda_\circ(\psi)$ uniformly in $\psi \in \Psi$ and $\lambda_\circ(\psi) \in int(\Lambda(\psi))$ by Assumption 7 (b), it follows that $\widehat{\lambda}(\psi) - \overline{\lambda}(\psi) \xrightarrow{p} 0$ as $n \to \infty$. By continuity of both $\overline{\lambda}(\psi)$ and $\widehat{\lambda}(\psi)$ in (ψ) (due to the Birge's maximum theorem and strict convexity of $\frac{1}{n}\sum_{i=1}^n \exp\{\lambda' g^A(x_i,\psi)\}$ in λ) and compactness of Ψ we conclude that

$$\sup_{\psi\in\Psi} \|\widehat{\lambda}(\psi) - \lambda_\circ(\psi)\| \xrightarrow{p} 0. \tag{F.12}$$

In the second step of the proof, we use the results of the first step to show the result of the lemma. By the triangular inequality and the Cauchy-Schwartz inequality

$$\begin{aligned}
\sup_{\psi\in\Psi}\left|\log\frac{\exp\{\widehat{\lambda}(\psi)'\widehat{g}^A(\psi)\}}{\frac{1}{n}\sum_{i=1}^n\exp\{\widehat{\lambda}(\psi)'g^A(x_i,\psi)\}}-\log\frac{\exp\{\lambda_\circ(\psi)'\mathbf{E}^P[g^A(x,\psi)]\}}{\mathbf{E}^P\left[\exp\{\lambda_\circ(\psi)'g^A(x,\psi)\}\right]}\right| \\
\leq \sup_{\psi\in\Psi}\|\widehat{\lambda}(\psi)-\overline{\lambda}(\psi)\|\sup_{\psi\in\Psi}\|\widehat{g}^A(\psi)\| \\
+\sup_{\psi\in\Psi}\|\overline{\lambda}(\psi)\|\sup_{\psi\in\Psi}\|\widehat{g}^A(\psi)-\mathbf{E}^P[g^A(x,\psi)]\| \\
+\sup_{\psi\in\Psi}\|\overline{\lambda}(\psi)-\lambda_\circ(\psi)\|\sup_{\psi\in\Psi}\|\mathbf{E}^P[g^A(x,\psi)]\| \\
+\sup_{\psi\in\Psi}\left|\log\frac{1}{n}\sum_{i=1}^n\exp\{\widehat{\lambda}(\psi)'g^A(x_i,\psi)\}-\log\mathbf{E}^P\left[\exp\{\lambda_\circ(\psi)'g^A(x,\psi)\}\right]\right| \\
=: \mathcal{A}_1+\mathcal{A}_2+\mathcal{A}_3+\mathcal{A}_4. \quad \text{(F.13)}
\end{aligned}$$

By continuity of $\psi\mapsto\overline{\lambda}(\psi)$ and compactness of Ψ and of $\Lambda(\psi)$ for all $\psi\in\Psi$: $\sup_{\psi\in\Psi}\|\overline{\lambda}(\psi)\|<\infty$. By Assumption 7 (b): $\sup_{\psi\in\Psi}\|\mathbf{E}^P[g^A(x,\psi)]\|\leq\mathbf{E}^P[M(x)]<\infty$ and $\sup_{\psi\in\Psi}\|\widehat{g}^A(\psi)\|<\infty$ (by using (F.11)). Therefore, $\mathcal{A}_i\xrightarrow{p}0$ for $i=1,2,3$ by (F.11) and (F.12).
In order to show the convergence to zero of $\mathcal{A}_4$ remark that because $\log(a)\leq a-1$ for every $a>0$, then,

$$\mathcal{A}_4\leq\sup_{\psi\in\Psi}\frac{\left|\frac{1}{n}\sum_{i=1}^n\exp\{\widehat{\lambda}(\psi)'g^A(x_i,\psi)\}-\mathbf{E}^P\left[\exp\{\lambda_\circ(\psi)'g^A(X,\psi)\}\right]\right|}{\mathbf{E}^P\left[\exp\{\lambda_\circ(\psi)'g^A(X,\psi)\}\right]}$$

which converges to zero by the result in Lemma F.1 below.

□

Lemma F.1. *Let M_ℓ be a misspecified model (that is, a model that does not satisfy Assumption 1) and let $g^A(x,\psi^\ell)$ and ψ^ℓ be the corresponding moment functions and parameters. Then, under Assumptions 5 (a)-(d), 3, and 7:*

$$\sup_{\psi\in\Psi}\left|\frac{1}{n}\sum_{i=1}^n\exp\{\widehat{\lambda}(\psi)'g^A(x_i,\psi)\}-\mathbf{E}^P\left[\exp\{\lambda_\circ(\psi)'g^A(X,\psi)\}\right]\right|\xrightarrow{p}0.$$

Proof. In this proof, let A_n denote the following event: $A_n:=\{\sup_{\psi\in\Psi}\|\widehat{\lambda}(\psi)-\lambda_\circ(\psi)\|\leq\delta_n\}$, for a $\delta_n>0$ converging to zero as $n\to\infty$, and let $B(\lambda_\circ,\delta_n)$ be the closed ball around $\lambda_\circ(\psi)$ with radius δ_n. Then, under Assumption 7 *(b)* there exists an $N>0$ such that $\forall n>N$:

$\widehat{\lambda}(\psi) \in \Lambda(\psi)$ on the event A_n.

Next, we prove the intermediate result

$$\sup_{\psi\in\Psi,\lambda\in\Lambda(\psi)} \left| \frac{1}{n}\sum_{i=1}^{n} e^{\lambda' g^A(x_i,\psi)} - \mathbf{E}^P\left[e^{\lambda' g^A(X,\psi)}\right] \right| \xrightarrow{a.s.} 0. \tag{F.14}$$

Consider the class of functions (on $\mathcal{X}$) $\mathcal{F} := \left\{\exp\{\lambda' g^A(\cdot,\psi)\}; \lambda \in \Lambda(\psi),\ \psi \in \Psi\right\}$. Since: (I) the function $(\lambda,\psi) \mapsto \exp\{\lambda' g^A(x,\psi)\}$ is continuous for P-almost all x (under Assumption 5 *(c)*); (II) Ψ is compact and $\Lambda(\psi)$ is compact for every $\psi \in \Psi$ (by Assumptions 5 *(b)* and 7 *(b)*); (III) the envelope of $\mathcal{F}$, $\sup_{\psi\in\Psi,\lambda\in\Lambda(\psi)} \exp\{\lambda' g^A(x,\psi)\}$ is in $L_1(P)$ (by Assumption 7 *(c)*), then (F.14) holds (see van de Geer (2010, Lemma 3.10, page 38)).

With this result, and the fact that $P(A_n^c) = o(1)$ by (F.12), we are now ready to show the result of the lemma. Let $h(\widehat{\lambda},\psi) := \mathbf{E}_X^P\left[e^{\widehat{\lambda}(\psi)' g^A(X,\psi)}\right]$, where $\mathbf{E}_X^P[\cdot]$ denotes the expectation taken with respect to the distribution of X only (so, we do not integrate out $\widehat{\lambda}$). Moreover, let $\eta > 0$ and denote by B_n the event $B_n := \{\sup_{\psi\in\Psi} \left| h(\widehat{\lambda},\psi) - \mathbf{E}^P\left[e^{\lambda_\circ(\psi)' g^A(X,\psi)}\right]\right| \leq \eta/2\}$. To upper bound $P(B_n^c)$ we use the Markov's inequality and the Mean value theorem applied to the function $h(\cdot,\psi)$ which is defined on $B(\lambda_\circ(\psi),\delta_n)$ on the event A_n:

$$\begin{aligned} P(B_n^c) &= P(B_n^c|A_n)P(A_n) + P(B_n^c|A_n^c)P(A_n^c) \\ &\leq \frac{2}{\eta}\mathbf{E}^P\left(\sup_{\psi\in\Psi}\left|\mathbf{E}_X^P\left[e^{\{\tilde{\lambda}(\psi)' g^A(X,\psi)\}} g^A(X,\psi)'\right](\widehat{\lambda}(\psi) - \lambda_\circ(\psi))\right| A_n\right) P(A_n) + P(A_n^c) \\ &= \frac{2\delta_n}{\eta}(\mathbf{E}^P\left(\left(\sup_{\psi\in\Psi}\left\|\mathbf{E}_X^P\left[e^{\{\tilde{\lambda}(\psi)' g^A(X,\psi)\}} g^A(X,\psi)'\right]\right\|\right)^2 1_{A_n}\right))^{1/2} + o(1) \end{aligned} \tag{F.15}$$

where $\tilde{\lambda}(\psi)$ is on the line joining $\widehat{\lambda}(\psi)$ and $\lambda_\circ(\psi)$ and 1_{A_n} is the indicator function of the event A_n. By applying the Cauchy-Schwartz inequality we get

$$\begin{aligned} &\mathbf{E}^P\left(\sup_{\psi\in\Psi}\left\|\mathbf{E}_X^P\left[e^{\tilde{\lambda}(\psi)' g^A(X,\psi)} g^A(X,\psi)\right]\right\|^2 1_{A_n}\right) \\ &\qquad \leq \mathbf{E}^P\left(\sup_{\psi\in\Psi}\mathbf{E}_X^P\left[e^{2\tilde{\lambda}(\psi)' g^A(X,\psi)}\right] 1_{A_n}\right) \sup_{\psi\in\Psi}\mathbf{E}^P\left\|g^A(X,\psi)\right\|^2. \end{aligned} \tag{F.16}$$

The first term in the product in the right hand side is bounded by uniform convergence of $\tilde{\lambda}(\psi)$ towards $\lambda_\circ(\psi)$ on A_n, uniform continuity of the function $\psi \mapsto e^{\lambda' g^A(\psi)}$ on $B(\lambda_\circ(\psi),2\delta_n)\times\Psi$, Assumption 7 *(c)* and by the Dominated Convergence theorem. The second term in the

product is bounded by Assumption 5 *(d)* because

$$\sup_{\psi\in\Psi}\mathbf{E}_X^P\left\|g^A(X,\psi)\right\|^2 \leq \mathbf{E}^P\left[\sup_{\psi\in\Psi}\left\|g^A(X,\psi)\right\|^2\right] \leq \mathbf{E}^P\left[\sup_{\psi\in\Psi}\left\|g^A(X,\psi)\right\|^\alpha\right], \qquad \forall\alpha>2.$$

Therefore, (F.15) and (F.16) and the fact that $\delta_n \to 0$ show that $P(B_n^c) = o(1)$.

Next,

$$\begin{aligned}
&P\left(\sup_{\psi\in\Psi}\left|\frac{1}{n}\sum_{i=1}^n e^{\widehat{\lambda}(\psi)'g^A(x_i,\psi)} - \mathbf{E}^P\left[e^{\lambda_\circ(\psi)'g^A(X,\psi)}\right]\right| > \eta\right) \leq \\
&\quad P\left(\sup_{\psi\in\Psi}\left|\frac{1}{n}\sum_{i=1}^n e^{\widehat{\lambda}(\psi)'g^A(x_i,\psi)} - h(\widehat{\lambda},\psi)\right| > \eta - \sup_{\psi\in\Psi}\left|h(\widehat{\lambda},\psi) - \mathbf{E}^P\left[e^{\widetilde{\lambda}(\psi)'g^A(X,\psi)}\right]\right|\right) \\
&\quad\leq P\left(\sup_{\psi\in\Psi}\left|\frac{1}{n}\sum_{i=1}^n e^{\widehat{\lambda}(\psi)'g^A(x_i,\psi)} - h(\widehat{\lambda},\psi)\right| > \eta - \frac{\eta}{2}\right)P(B_n) + P(B_n^c) \\
&\leq P\left(\sup_{\psi\in\Psi}\left|\frac{1}{n}\sum_{i=1}^n e^{\widehat{\lambda}(\psi)'g^A(x_i,\psi)} - h(\widehat{\lambda},\psi)\right| > \frac{\eta}{2}\middle| A_n\right)P(A_n)P(B_n) + P(A_n^c)P(B_n) + P(B_n^c) \\
&\quad\leq P\left(\sup_{\psi\in\Psi,\lambda\in\Lambda(\psi)}\left|\frac{1}{n}\sum_{i=1}^n e^{\lambda'g^A(x_i,\psi)} - \mathbf{E}^P\left[e^{\lambda'g^A(X,\psi)}\right]\right| > \frac{\eta}{2}\right) + o(1) \quad \text{(F.17)}
\end{aligned}$$

where to get the last line we have used the fact that the probability is conditional on the event A_n and $P(A_n^c) = o(1)$. Finally, this probability goes to zero by (F.14).

□

References

Chib, S. (1995), 'Marginal Likelihood from the Gibbs Output', *Journal of the American Statistical Association* **90**(432), 1313–1321.

Kleijn, B. and van der Vaart, A. (2012), 'The Bernstein-Von-Mises Theorem Under Misspecification', *Electronic Journal Statistics* **6**, 354–381.

Newey, W. K. and McFadden, D. (1994), Chapter 36: Large sample estimation and hypothesis testing, Vol. 4 of *Handbook of Econometrics*, Elsevier, pp. 2111 – 2245.

Newey, W. K. and Smith, R. J. (2004), 'Higher Order Properties of GMM and Generalized Empirical Likelihood Estimators', *Econometrica* **72**(1), 219–255.

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

van de Geer, S. A. (2010), *Empirical Processes in M-Estimation*, Cambridge Series in Statistical and Probabilistic Mathematics.

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

Van der Vaart, A. W. (2002), Semiparametric Statistics, *in* P. Bernard, ed., 'Lectures on Probability Theory and Statistics - Ecole d'Eté de Probabilités de Saint-Flour XXIX - 1999', Vol. 1781 of *Lecture Notes in Mathematics*, Springer-Verlag Berlin Heidelberg.