Semiparametric Bayesian Estimation and Comparison of Moment Condition Models

Siddhartha Chib* Minchul Shin† Anna Simoni‡

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

This version: February 15, 2016

Abstract

In this paper we consider the problem of inference in statistical models characterized via moment restrictions and develop a semiparametric Bayes procedure for selecting valid and relevant moments. We cast the moment estimation problem in the Exponentially Tilted Empirical Likelihood (ETEL) framework introduced by Schennach [2007]. Because the ETEL has a well-defined probabilistic interpretation and plays the role of a likelihood, a fully Bayesian framework can be developed. We show how the moment selection problem can be tackled on the basis of marginal likelihoods. These are computed exactly (up to simulation error) by Chib [1995]'s method. We show that our proposed marginal likelihood based moment selection procedure is consistent in the sense that it discards misspecified as well as irrelevant moment restrictions. As a byproduct, we prove that a posterior distribution obtained from the ETEL satisfies the Bernstein - von Mises theorem in misspecified moment models. The finite sample properties of our procedure are illustrated in simulation exercises in the settings of linear instrumental regression and quantile instrumental regression. Finally, we apply our method to the habit persistence asset pricing model.

*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, 15, Boulevard Gabriel Péri, 92240 Malakoff, France, e-mail: simoni.anna@gmail.com

1 Introduction

Statistical and econometric models characterized via moment restrictions of the type $\mathbf{E}[g(X,\theta)] = 0$ where $g(X,\theta)$ is a known vector-valued function of a random vector X and an unknown parameter vector θ are very often encountered in many applications. These models are semiparametric in the sense that no restriction is imposed on the data generating process (DGP) except for this set of moment restrictions and one is interested in making inference about the finite dimensional parameter θ.

Models of this type, which do not involve the direct specification of a likelihood function, have, for obvious reasons, posed a long-standing challenge for Bayesian inference. In this paper we cast the moment estimation problem in the Exponentially Tilted Empirical Likelihood (ETEL) framework introduced by Schennach [2007] and develop a Bayes procedure for moment selection. The idea is to use the exponential tilted (ET) implied probabilities to construct a likelihood function. This likelihood function has been shown to have a well-defined probabilistic interpretation by Schennach [2005]. In particular, Schennach [2005] shows that the ETEL naturally arises as the limit of a nonparametric Bayes procedure that places a certain prior on the set of distributions as this prior becomes more diffuse. Given that the ETEL plays the role of a likelihood, a fully Bayesian procedure can be developed. In particular, marginal likelihood (ML) and Bayes factors can be computed.

One of the advantages of moment restriction models is that they avoid any parametric assumption on the data generating process and hence reduce the risk of misspecification and of inconsistent estimation, compared to parametric models. However, if no value of θ satisfies the moment restrictions simultaneously in the population then the model is misspecified. If this is the case then inference based on the given set of moment restrictions leads to an estimator that, if it is convergent, converges towards a value different from the true one. For this reason it is important to have a selection procedure that discards the misspecified and irrelevant moment restrictions.

This paper proposes a moment selection procedure based on marginal likelihoods. To the best of our knowledge, this is the first time that ML is applied to the framework of moment condition models in conjunction with the ETEL. Our procedure selects the set of moment restrictions that maximizes the ML defined through the ETEL and the prior distribution on the finite dimensional parameter. Each set of moment restrictions defines a particular model and the different sets of restrictions define different models. These models can be compared based on their corresponding ML. In order to make the models comparable we transform the moment functions $g(X,\theta)$ so that all the transformed moments are included in all the

models but the transformation depends on the model. This linear transformation simply consists in adding an extra parameter different from zero to the components of the vector $g(X, \theta)$ that correspond to the restrictions not included in a specific model.

We compute the marginal likelihood based on the method proposed by Chib [1995] and extended by Chib and Jeliazkov [2001]. This method makes computation of the ML extremely simple and is a key feature of our procedure. The main advantage of Chib [1995]'s method is that it avoids numerical integration of the ETEL function with respect to the prior distribution which may be very challenging. We develop asymptotic theory to show that our ML-based selection procedure is consistent in the sense that it discards misspecified as well as irrelevant moment restrictions. In other words, the model that maximizes the ML is the one that contains the maximum number of valid moment restrictions that are not irrelevant. Irrelevant moment restrictions are moment restrictions that do not contain any additional information and leave the ML unchanged. The consistency result is based on: *(i)* the asymptotic behavior of the ETEL function for both correctly and misspecified models and *(ii)* the validity of the Bernstein-von Mises (BvM) theorem for both correctly and misspecified models. The BvM theorem for a posterior distribution obtained from the ETEL has been discussed in Schennach [2005] for correctly specified models whereas this result for misspecified model was unknown in the literature. Therefore, our paper contributes also to the literature on semiparametric Bayesian statistics by proving that a posterior distribution obtained from the ETEL satisfies the BvM theorem in misspecified models. We stress that results available in the literature, like Kleijn and van der Vaart [2012], cannot be directly applied to our framework because the ETEL likelihood contains random quantities.

The finite sample properties of our selection procedure are analyzed through a series of simulation exercises. In particular, we consider the linear and quantile regression problem with an endogenous regressor and multiple instruments. We assume that we are sure that one of the instruments is valid and relevant while we are not sure whether the rest of instruments are good instruments or not. We first transform validity and relevance conditions for each instrument into the unconditional moment restrictions. Then, we impose additional parameter restrictions on moment restrictions related to questionable instruments. Using simulated data, we confirm the theoretical results provided in this paper. More specifically, we show that estimated marginal likelihood can differentiate valid/invalid instruments as well as relevant/irrelevant instruments with reasonable amount of observations. Finally, we illustrate our ML-based selection procedure by applying it to the habit persistence asset pricing model proposed by Campbell and Cochrane [1999].

Previous literature on Bayesian inference in moment condition models can be divided in two broad classes: quasi-Bayesian methods and pure Bayesian methods. The procedures in the first class are constructed by using a likelihood that is not the true one but is obtained by exponentiating the quadratic form associated with the empirical counterpart of the moment restrictions. Even if this is not the true likelihood, it has been shown that the corresponding DGP is the closest to the true one in terms of Kullback-Leibler divergence, among all the DGP satisfying the moment restrictions, see Csiszar [1975]. This class includes the following contributions: Kwan [1999], Kim [2002], Chernozhukov and Hong [2003], Liao and Jiang [2011], Gallant [2015] and Gallant et al. [2015].

On the contrary, pure Bayesian methods for moment condition models use a proper likelihood by putting a nonparametric prior on it and includes: Florens and Rolin [1994], Chamberlain and Imbens [2003], Lazar [2003], Schennach [2005], Ragusa [2007], Kitamura and Otsu [2011], Shin [2014], Florens and Simoni [2015], Bornn et al. [2015]. The procedure used in this paper is fully Bayesian and enters in this class.

Our paper also relates to the literature on moment selection: Andrews [1999] and Andrews and Lu [2001] use a criterion based on the J statistic penalized by the number of moment conditions; Hong et al. [2003] construct a moment selection criterion based on a penalized Generalized Empirical Likelihood (GEL) function; Hong and Preston [2012] construct GEL Bayes factors where the penalization for the number of moment restriction arises from the integrated prior on the tilting parameter; Inoue and Shintani [2015] show the consistency of the model selection procedure based on the marginal likelihood obtained from Laplace-type estimators of Chernozhukov and Hong [2003]. Other important contributions on moment selection includes: Donald and Newey [2001], Inoue [2006], Hall et al. [2007], Liao [2013], Cheng and Liao [2015], and DiTraglia [2015].

2 The setting

Let X be an $\mathbb{R}^{d_x}$-valued random vector with distribution P and $x_1, \ldots, x_n$ be an *i.i.d.* sample of X. We denote by $x^{(n)} = (x_1, \ldots, x_n)$ the vector of observations. Let $g(X, \theta)$ be a vector of known functions with values in $\mathbb{R}^d$, namely $g : \mathbb{R}^{d_x} \times \Theta \mapsto \mathbb{R}^d$ where $\Theta \subset \mathbb{R}^p$. The parameter of interest $\theta \in \Theta$ is characterized by the set of moment restrictions

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

where $\mathbf{E}^P$ denotes the expectation taken with respect to P. We assume that $d \geq p$ and denote by $\pi(\theta)$ the prior distribution on θ. The Lebesgue-density of P is unknown and in a Bayesian framework it is natural to specify a nonparametric prior on it. We consider the nonparametric prior introduced by Schennach [2005]. For convenience we summarize Schennach's prior here.

2.1 The nonparametric Schennach [2005]'s prior

The nonparametric prior on the likelihood function is specified via a vector of parameters $\xi_N = (\xi_{N,1}, \ldots, \xi_{N,N})$ that defines the support and contains all the information about the likelihood provided by the moment restrictions. For a given N and a given vector of parameters $\xi_N = (\xi_{N,1}, \ldots, \xi_{N,N})$, the Lebesgue-density of P, denoted by $p(x|\xi_N)$, is taken to be a mixture of N uniform densities, each supported on a d_x-dimensional hypercube of side $2\rho_N$ centred on $\xi_{N,j}$, $j = 1, \ldots, N$, where $\rho_N \to 0$ as $N \to \infty$:

$$p(x^{(n)}|\xi_N) = \prod_{i=1}^{n} \left[\frac{1}{N} \rho_N^{-d_x} \sum_{j=1}^{N} 1\left\{ \|x_i - \xi_{N,j}\|_\infty \leq \rho_N \right\} \right]. \tag{2.2}$$

In the limit as $N \to \infty$, any distribution can be represented by this mixture. Next, the moment restrictions (2.1) are imposed via the conditional prior of ξ_N, given θ. This prior is constructed in two steps. In a first step, ξ_N is drawn from a prior $\pi(\xi_N)$ with discrete support $(\mathbb{X}_N)^N$ that ignores the moment restrictions and that is described in Schennach [2005, Appendix 1]. In the second step, ξ_N is retained only if the corresponding $p(x|\xi_N)$ satisfies the moment restrictions within a tolerance of ε, for a given value of θ. For this, let us introduce a function $G(\xi_N, \theta, \varepsilon)$:

$$\begin{aligned} G(\xi_N, \theta, \varepsilon) &= 1\left\{ \left\| \int p(x|\xi_N) g(x,\theta) dx \right\|_\infty \leq \varepsilon \right\} \\ &= 1\left\{ \left\| \frac{1}{N} \rho_N^{-d_x} \sum_{j=1}^{N} \int_{\|x - \xi_{N,j}\|_\infty \leq \rho_N} g(x,\theta) dx \right\|_\infty \leq \varepsilon \right\} \end{aligned}$$

where $1\{A\}$ is 1 if A holds true and 0 otherwise, the outer norm $\|\cdot\|_\infty$ denotes the maximum over the moment functions. By summing up these two steps, the prior of ξ_N conditional on the moment conditions and on θ is:

$$\pi(\xi_N|\theta, \varepsilon) = \frac{G(\xi_N, \theta, \varepsilon)\pi(\xi_N)}{\sum_{\xi_{N,1}} \cdots \sum_{\xi_{N,N}} G(\xi_N, \theta, \varepsilon)\pi(\xi_N)} \tag{2.3}$$

where the N sums in the last display are taken over the support $(\mathbb{X}_N)^N$.

2.2 The posterior distribution $p(\theta|X)$

By using the nonparametric prior (2.3) and the Lebesgue-density of P (2.2), the posterior $p(\theta|X)$ is defined as the limit:

$$\begin{aligned}\pi(\theta|x^{(n)}) &\propto \lim_{\varepsilon\to 0}\lim_{N\to\infty}\int\prod_{i=1}^{n} p(x_i|\xi_N)\pi(\xi_N|\theta,\varepsilon)d\xi_N\ \pi(\theta)\\ &\propto \lim_{\varepsilon\to 0}\lim_{N\to\infty}\frac{\sum_{\xi_{N,1}}\cdots\sum_{\xi_{N,N}}\prod_{i=1}^{n} p(x_i|\xi_N)G(\xi_N,\theta,\varepsilon)\pi(\xi_N)}{\sum_{\xi_{N,1}}\cdots\sum_{\xi_{N,N}} G(\xi_N,\theta,\varepsilon)\pi(\xi_N)}\ \pi(\theta). \qquad (2.4)\end{aligned}$$

Under the assumptions that: (i) $g(x,\theta)$ is continuous in x for every $\theta\in\Theta$, and (ii) the interior of the convex hull of $\bigcup_{i=1}^{n} g(x_i,\theta)$ contains the origin, Schennach [2005, Theorem 1] shows that the posterior distribution of θ given in (2.4) writes

$$\pi(\theta|x^{(n)}) \propto \pi(\theta)\prod_{i=1}^{n}\frac{e^{\widehat{\lambda}(\theta)'g(x_i,\theta)}}{\sum_{j=1}^{n} e^{\widehat{\lambda}(\theta)'g(x_j,\theta)}} \qquad (2.5)$$

where $\widehat{\lambda}(\theta) = \arg\min_{\lambda\in\mathbb{R}^d}\frac{1}{n}\sum_{i=1}^{n}\exp\left(\lambda' g(x_i,\theta)\right)$ is the estimated tilting parameter. The weights $w_i^*(\theta) := \frac{e^{\hat{\lambda}(\theta)'g(x_i,\theta)}}{\sum_{j=1}^{n} e^{\hat{\lambda}(\theta)'g(x_j,\theta)}}$, $i=1,\ldots,n$, are the maximizers of the entropy:

$$\begin{aligned}&\max_{w_1,\ldots,w_n}\sum_{i=1}^{n}\left[-w_i\log(nw_i)\right]\\ &\text{subject to}\quad \sum_{i=1}^{n} w_i = 1 \quad\text{and}\quad \sum_{i=1}^{n} w_i g(x_i,\theta) = 0.\end{aligned}$$

Because the likelihood function can be obtained as the limit of a nonparametric Bayes procedure that places a particular prior on the set of distributions, this posterior is fully Bayesian.

3 Marginal Likelihood and Consistency Results

3.1 Moment Selection

The set of moment restrictions (2.1) might contain restrictions that are missspecified or that are irrelevant.

Definition 3.1 (Misspecified model)**.** *We say that the 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 $\theta \in \Theta$, that is, $P \notin \mathcal{P}$ where $\mathcal{P} = \bigcup_{\theta\in\Theta} \mathcal{P}_\theta$ and $\mathcal{P}_\theta = \{Q \in \mathbb{M};\ \mathbf{E}^Q[g(X,\theta)] = 0\}$ with $\mathbb{M}$ the set of all probability measures on $\mathbb{R}^{d_x}$.*

Including a misspecified moment restrictions causes inconsistency of the corresponding estimator, while including irrelevant moment restrictions increases the burden of the estimation procedure without any gain. It is therefore important to have a procedure that is able to select the correctly specified and relevant moment restrictions. In the spirit of Bayesian model selection, we propose in this paper to use the Marginal Likelihood (ML) obtained by using the ETEL to select the moment restrictions and discard the misspecified and irrelevant ones.

Let $c \in \mathbb{R}^d$ be a d-vector of zeros and ones that selects the moment conditions: if the jth element of c is a one, then the jth moment condition is included; if the jth element of c is a zeros, then it is not included. For a given c, $|c|$ denotes the total number of moments selected by c. We assume that the model is identified: $p \leq |c| \leq d$. We denote by $g_c(x,\theta)$ the subvector of $g(x,\theta)$ made of $|c|$ moment conditions selected by c, and denote by λ_c the $|c|$-dimensional vector of tilting parameters corresponding to the $|c|$ moment conditions selected by c. Every vector c characterizes a model and we denote by $m(x^{(n)};c)$ the corresponding ML:

$$m(x^{(n)};c) = \int p(x^{(n)}|\theta;c)\pi(\theta)d\theta \tag{3.1}$$

where

$$p(x^{(n)}|\theta;c) = \prod_{i=1^n} w_i^*(\theta) = \prod_{i=1}^{n} \frac{e^{\hat{\lambda}_c(\theta)' g_c(x_i,\theta)}}{\sum_{j=1}^{n} e^{\hat{\lambda}_c(\theta)' g_c(x_j,\theta)}}. \tag{3.2}$$

As shown by Chib [1995], the log-ML can be written as follows: $\forall x^{(n)} = (x_1,\ldots,x_n)$ and $\forall\theta \in \Theta$,

$$\log m(x^{(n)};c) = \log \pi(\theta) + \log p(x^{(n)}|\theta;c) - \log \pi(\theta|x^{(n)};c). \tag{3.3}$$

Remark that the left hand side of (3.3) does not depend on θ therefore, the right hand side of (3.3) can be evaluated at any value of θ without changing the value of the log-ML. The ML-based selection procedure selects the set of moments that maximizes the ML for a given

observed sample $x^{(n)}$:

$$\hat{c} = \arg\max_{c \in \mathbb{R}^d} \log m(x^{(n)}; c).$$

Each set of moment restrictions defines a model. The ML-based criterion retains the model with maximum value of ML among all the models. However, in order to make comparable the models based on different set of moment restrictions, we need to make a linear transformation of the moment functions by adding extra parameters. In the following, we refer to this transformation as *augmentation* because it is based on the addition of extra parameters. We now explain how this augmentation idea works.

Denote by M_d the model that considers the d moment restrictions $\mathbf{E}^P[g(X,\theta)] = 0$ to estimate θ and by λ_d the corresponding tilting parameter. For every $p \leq |c| \leq d$ and every $\ell = 1, \ldots, d!$, let $p^\ell_{|c|}$ be the d-vector of zeros and ones that has $|c|$ ones and that corresponds to the ℓ-th permutation of the set that contains $|c|$ ones and $(d-|c|)$ zeros. Let $p^\ell_{|c|}(j)$ denote the j-th component of the vector $p^\ell_{|c|}$. For each $p^\ell_{|c|}$, the augmentation idea consists in the following steps.

1. Augment the parameter vector to $(\theta', v')'$ where $v \in \mathbb{R}^{d-|c|}$ is a nuisance parameter;

2. for every $j \in \{j = 1, \ldots, d;\, p^\ell_{|c|}(j) = 0\}$, transform the moment function $g_j(x,\theta)$ in $g_j(x,\theta) - \breve{v}_j$, for some parameter $\breve{v}_j$ to be defined below;

3. estimate the parameter vector $(\theta', v')'$, as well as the corresponding tilting parameter, by using the $|c|$ moment restrictions $\mathbf{E}^P[g_j(X,\theta)] = 0$ with $j \in \{j = 1, \ldots, d;\, p^\ell_{|c|}(j) = 1\}$ and the $d-|c|$ transformed moment restrictions $\mathbf{E}^P[g_j(X,\theta)] - v_j = 0$, for $j \in \{j = 1, \ldots, d;\, p^\ell_{|c|}(j) = 0\}$.

For every $p \leq |c| \leq d$ and every $\ell = 1, \ldots, d!$, let $\breve{v} = (\breve{v}_1, \ldots \breve{v}_d)' \in \mathbb{R}^d$, v be the $(d-|c|)$-vector of nonzero components of $diag(\iota - p^\ell_{|c|})\breve{v}$ where ι is a d-vector of ones[1]. Moreover, let

$$g_{p^\ell_{|c|}}(x,\theta,v) = g(x,\theta) - diag(\iota - p^\ell_{|c|})\breve{v}$$

be the d-vector of augmented moment functions. Denote by $M_{p^\ell_{|c|}}$ the model that uses the d augmented moments restrictions $\mathbf{E}^P[g_{p^\ell_{|c|}}(X,\theta,v)] = 0$ to estimate $(\theta', v')'$ and by $\lambda_{|c|,\ell}$ the corresponding tilting parameter. For illustration, suppose that for a given c and ℓ the ℓ-th permutation is

$$p^\ell_{|c|} = (\underbrace{1, \ldots, 1}_{|c| \text{ elements}}, \underbrace{0, \ldots, 0}_{d-|c| \text{ elements}})$$

[1]For simplicity, we do not explicit the dependence of v on ℓ and $|c|$.

then,

$$g_{p^{\ell}_{|c|}}(x,\theta,v) = \left(g_1(x,\theta),\ldots,g_{|c|}(x,\theta), g_{|c|+1}(x,\theta)-v_1,\ldots,g_d(x,\theta)-v_{d-|c|}\right)'. \tag{3.4}$$

Remark that $\widehat{\lambda}_{|c|,\ell}(\theta,v) = \arg\max_{\lambda\in\mathbb{R}^d}\left\{-\frac{1}{n}\sum_{i=1}^n \exp\{\lambda' g(x_i,\theta) - \lambda' diag(\iota - p^{\ell}_{|c|})\breve{v}\}\right\}$.

The prior for the parameter $(\theta', v')'$ can be specified as an independent prior: $\pi(\theta,v) = \pi(\theta)\times\pi(v)$. Then, the log-ML for the augmented model $M_{p^{\ell}_{|c|}}$ writes:

$$\log m(x^{(n)}; M_{p^{\ell}_{|c|}}) = \log\pi(\theta) + \log\pi(v) + \log p\left(x^{(n)}|\theta,v; M_{p^{\ell}_{|c|}}\right) - \log\pi(\theta,v|x^{(n)}; M_{p^{\ell}_{|c|}}). \tag{3.5}$$

To sum up, for practical implementation of our ML-based selection procedure, one has first to perform this simple augmentation of the moment functions and then to use the log-ML (3.5) as the selection criterion .

3.2 Consistency

Let θ_* be the true value of the parameter of interest θ. In order to prove consistency of the ML-based selection procedure we define $\Delta := \mathbf{E}^P[g(X,\theta_*)g(X,\theta_*)']$, $\Gamma := \mathbf{E}^P\left[\frac{\partial}{\partial\theta}g(X,\theta_*)\right]$ and make the following assumptions.

Assumption 1. *The permutation $p^{\ell}_{|c|}$ is such that $(\theta_*, v_*) \in \Theta\times\mathbb{R}^{d-|c|}$ is the unique solution to $\mathbf{E}^P[g_{p^{\ell}_{|c|}}(X,\theta,v)] = 0$;*

Assumption 2. *(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) Θ is compact; (c) $g(x,\theta)$ is continuous at each $\theta\in\Theta$ with probability one; (d) $\mathbf{E}^P[\sup_{\theta\in\Theta}\|g(X,\theta)\|^{\alpha}] < \infty$ for some $\alpha > 2$; (e) Δ is nonsingular.*

Assumption 3. *(a) $\theta_*\in int(\Theta)$; (b) $g(x,\theta)$ is continuously differentiable in a neighborhood $\mathcal{U}$ of θ_* and $\mathbf{E}^P[\sup_{\theta\in\mathcal{N}}\|\partial g(x_i,\theta)/\partial\theta\|] < \infty$; (c) $rank(\Gamma) = p$.*

We assume that the prior distribution π on Θ satisfies the following assumption.

Assumption 4. *(a) π is a continuous probability measure that admits a density with respect to the Lebesgue measure; (b) π is positive on a neighborhood of θ_*.*

The first theorem states that, if the moment restrictions are all valid, then the ML selects the model with the maximum number of moment restrictions. We use the notation $P^n = \bigotimes_{i=1}^n P$ for the product measure.

Theorem 3.1. *Let Assumptions 2 – 4 hold and consider two different models* $M_{p^{\ell}_{|c_1|}}$ *and* $M_{p^{\ell}_{|c_2|}}$ *where* $p^{\ell}_{|c_1|}$ *and* $p^{\ell}_{|c_2|}$ *satisfy Assumption 1. Then, if* $|c_1| < |c_2|$*:*

$$\lim_{n\to\infty} P^n\left(\log m(x^{(n)}; M_{p^{\ell}_{|c_1|}}) < \log m(x^{(n)}; M_{p^{\ell}_{|c_2|}})\right) = 1.$$

Next, suppose that among the d moment restrictions there are some restrictions that are invalid and or irrelevant. Then, some of the models that we consider is misspecified. In order to analyze the behavior of the ML in this case, we have to analyze the behavior of the likelihood and the posterior under misspecification. Let $\widetilde{\lambda}_{|c|,\ell}(\theta, v)$ be the solution of $\mathbf{E}^P[\exp\{\lambda' g_{p^{\ell}_{|c|}}(X,\theta,v)\} g_{p^{\ell}_{|c|}}(X,\theta,v)] = 0$ which is unique by the strict convexity of $\mathbf{E}^P[\exp\{\lambda' g_{p^{\ell}_{|c|}}(X,\theta,v)\}]$ in λ. We refer to $\widetilde{\lambda}_{|c|,\ell}(\theta, v)$ as the pseudo-true value of the tilting parameter. The pseudo-true value of (θ, v) is:

$$(\widetilde{\theta}_{|c|,\ell}, \widetilde{v}) = \arg\max_{(\theta,v)\in\Theta\times\mathcal{V}} \mathbf{E}^P\left[\exp\{\widetilde{\lambda}'_{|c|,\ell} g_{p^{\ell}_{|c|}}(X,\theta,v)\}\right].$$

Assumption 5. *For a fixed* $\theta \in \Theta$*, there exists* $Q \in \mathcal{P}_\theta$ *such that* Q *is mutually absolutely continuous with respect to* P*.*

Assumption 5 is important in a misspecified model since, if the model is misspecified, the true data generating process P does not belong to $\mathcal{P}$. This assumption guarantees that the dual representation for the Kullback-Leibler information criterion (KLIC) minimization holds and that the pseudo-true values $(\widetilde{\theta}_{|c|,\ell}, \widetilde{v})$ and $\widetilde{\lambda}_{|c|,\ell}(\theta, v)$ exist, see Sueishi [2013]. Moreover, this assumption is used to establish the Bernstein-von Mises theorem under misspecification, see Kleijn and van der Vaart [2012].

Assumption 6. *(a) for every* $0 < |c| < d$*,* $(\theta, v) \in \Theta \times \mathcal{V} \subset \mathbb{R}^p \times \mathbb{R}^{d-|c|}$ *where* Θ *and* $\mathcal{V}$ *are compact; (b)* $g_{p^{\ell}_{|c|}}(x_i, \theta, v)$ *is continuous in* (θ, v) *at each* $(\theta, v) \in \Theta \times \mathcal{V}$ *with probability 1; (c) there exists a function* $M(\cdot)$ *such that* $\mathbf{E}^P[M(X)] < \infty$ *and* $\|g_{p^{l}_{|c|}}(x, \theta, v)\| \leq M(x)$*; (d)* $\widetilde{\lambda}_{|c|,\ell}(\theta, v) \in int(\Lambda(\theta, v))$ *where* $\Lambda(\theta, v)$ *is a compact set; (e)* $\mathbf{E}^P\left[\sup_{(\theta,v)\in\Theta\times\mathcal{V},\lambda\in\Lambda(\theta,v)} e^{\{\lambda' g_{p^{\ell}_{|c|}}(X,\theta,v)\}}\right] < \infty$*.*

The following theorem establishes that the ML selection criterion does not select models that contain misspecified moment restrictions with probability approaching one.

Theorem 3.2. *Let Assumptions 2 - 6 hold and consider two different models* $M_{p^{\ell}_{|c_1|}}$ *and* $M_{p^{\ell}_{|c_2|}}$ *where* $M_{p^{\ell}_{|c_1|}}$ *does not use misspecified moments while* $M_{p^{\ell}_{|c_2|}}$ *does, that is,* $p^{\ell}_{|c_1|}$ *satisfies*

Assumption 1 whereas $p^{\ell}_{|c_2|}$ does not. Moreover, assume that Assumption 4 (b) holds with θ_* replaced with the pseudo-true value $\widetilde{\theta}_{|c_2|,\ell}$. Then,

$$\lim_{n\to\infty} P^n\left(\log m(x^{(n)}; M_{p^{\ell}_{|c_1|}}) > \log m(x^{(n)}; M_{p^{\ell}_{|c_2|}})\right) = 1.$$

4 Monte Carlo Experiments

In this section, we explore the finite sample properties of the moment selection procedure proposed in the previous section. In the first two subsections, we simulate data from a linear instrumental regression model with one endogenous regressor and multiple instruments. In the third subsection, we consider the quantile regression with one endogenous regressor and multiple instruments. IN both cases, we assume that assume that the first instrument is valid and relevant while there is uncertainty about validity of the remaining instruments.

4.1 Linear IV regression with normally distributed errors

Data generating process. We generate data from the following data generating process,

$$\begin{aligned} y_t &= \beta x_t + e_{1,t} \\ x_t &= c_x + \delta_1 z_{1,t} + \delta_2 z_{2,t} + e_{2,t} \\ z_{1,t} &= c_{z1} + e_{3,t} \\ z_{2,t} &= c_{z2} + e_{4,t}, \end{aligned} \tag{4.1}$$

where $e_{1,t}, e_{2,t}, e_{3,t}$, and $e_{4,t}$ are normally distributed (standard normal) with $corr(e_{1,t}, e_{2,t}) = 0.6$ so that x_t is endogenous. We set $c_{z1} = c_{z2} = 0.5$, and $c_x = 0$. For the rest of this section, we assume that z_1 is a valid and relevant instrument. This implies that $\delta_1 = 1$ and $cov(e_{1,t}, e_{3,t}) = 0$. On the other hand we are unsure whether z_2 is a good instrument and, therefore, we treat δ_2 and $cov(e_{1,t}, z_{2,t}) = v_2$ as potentially non-zero unknown parameters.

We can also translate this model in terms of unconditional moment restrictions:

$$\begin{aligned} \mathbf{E}[(y_t - x_t\beta)z_{1,t}] &= 0 \\ \mathbf{E}[(y_t - x_t\beta)z_{2,t}] &= v_2 \\ \mathbf{E}[(x_t - \delta_1 z_{1,t} - \delta_2 z_{2,t})z_{1,t}] &= 0 \\ \mathbf{E}[(x_t - \delta_1 z_{1,t} - \delta_2 z_{2,t})z_{2,t}] &= 0. \end{aligned} \tag{4.2}$$

Table 1: Model configuration

	Name	Parameter configuration	# of parameters	# of moments
M1	$z1$-model	$\delta_2 = 0, v_2 = 0$	3	5
M2	$z1z2$-model	$\delta_2 \neq 0, v_2 = 0$	4	5
M3	$z1z2v2$-model	$\delta_2 \neq 0, v_2 \neq 0$	5	5
M4	$z1v2$-model	$\delta_2 = 0, v_2 \neq 0$	4	5

Under this setting, we have four possible model specifications. The first model assumes that the second instrument ($z_{2,t}$) is valid ($v_2 = 0$) but irrelevant ($\delta_2 = 0$). In the second model, we assume that $z_{2,t}$ is valid ($v_2 = 0$) and relevant ($\delta_2 \neq 0$). The third model assumes that $z_{2,t}$ is invalid ($v_2 \neq 0$) but relevant ($\delta_2 \neq 0$). Lastly, the fourth model assumes that $z_{2,t}$ is invalid ($v_2 \neq 0$) and irrelevant ($\delta_2 = 0$). These specifications are summarized in the Table 1.

The prior distributions for finite dimensional parameters, $(\beta, \delta_1, \delta_2, v_2)$, are relatively weak. Each element in this vector is set to be an independent normal distribution with mean zero and variance 10.

In what follows, we generate 200 observations from each of four models and compute the marginal likelihood based on four model specifications. This leads to 16 combinations. All results are summarized in Table 2. In this table, the first column presents the marginal likelihood estimated based on the method proposed by Chib and Jeliazkov [2001]. For all four cases, estimated marginal likelihood picks up the true data generating process. In the next column, we present maximum and minimum value of marginal likelihood estimates based on multiple independent Metropolis-Hastings chains (see Chib and Greenberg [1995]). For now, we compute marginal likelihood using $30,000$ draws from the posterior distribution (after discarding 10,000 draws). Then, we repeat this computation for three to seven times depending on the model specification[2]. This min/max range roughly gauges the accuracy of the marginal likelihood estimates. The last two columns present the posterior mean and 90% credible sets for the structural parameter (β).

4.2 Linear IV regression with non-Gaussian errors

One of the advantages of the Bayesian ETEL estimation procedure is that one does not need to make a strong parameteric assumption about the underlying distribution. To illustrate this point, we generate e_t from the log-normal distribution. More specifically, we simulate the vector of innovations, $e_t = [e_{1,t}, e_{2,t}, e_{3,t}, e_{4,t}]$ from the following log-normal

[2]We plan to increase the number of repetitions.

Table 2: IV Regression with Normally distributed errors

	$log(p(Y\|M_i))$	(min, max)	$mean(\beta)$	90% Credible Set
		(a) M1 is DGP		
M1	**-1069.74**	(-1069.75, -1069.72)	0.924	(0.808, 1.037)
M2	-1071.84	(-1071.87, -1071.80)	0.919	(0.806, 1.035)
M3	-1075	(-1075.04, -1074.98)	0.915	(0.802, 1.035)
M4	-1071.67	(-1071.72, -1071.63)	0.912	(0.797, 1.028)
		(b) M2 is DGP		
M1	-1314.77	(-1314.78, -1314.75)	1.094	(1.032, 1.159)
M2	**-1073.04**	(-1073.09, -1073.00)	0.991	(0.929, 1.058)
M3	-1075.16	(-1075.17, -1075.13)	0.922	(0.823, 1.028)
M4	-1220.43	(-1220.47, -1220.38)	0.951	(0.878, 1.019)
		(c) M3 is DGP		
M1	-1218.7	(-1218.71, -1218.69)	1.046	(0.993, 1.098)
M2	-1116.44	(-1116.46, -1116.39)	1.096	(1.046, 1.146)
M3	**-1075.03**	(-1075.04, -1075.03)	0.911	(0.802, 1.026)
M4	-1209.8	(-1209.81, -1209.79)	1.018	(0.974, 1.063)
		(d) M4 is DGP		
M1	-1298.63	(-1298.65, -1298.62)	0.85	(0.768, 0.937)
M2	-1104.2	(-1104.24, -1104.16)	0.635	(0.479, 0.802)
M3	-1075.06	(-1075.09, -1075.04)	0.916	(0.802, 1.035)
M4	**-1072.06**	(-1072.09, -1072.02)	0.916	(0.798, 1.033)

distribution

$$v_t = [v_{1,t}, v_{2,t}, v_{3,t}, v_{4,t}]' \sim \log \mathcal{N}\left(0,\ c \exp\left(s\Sigma\right)\right) \tag{4.3}$$

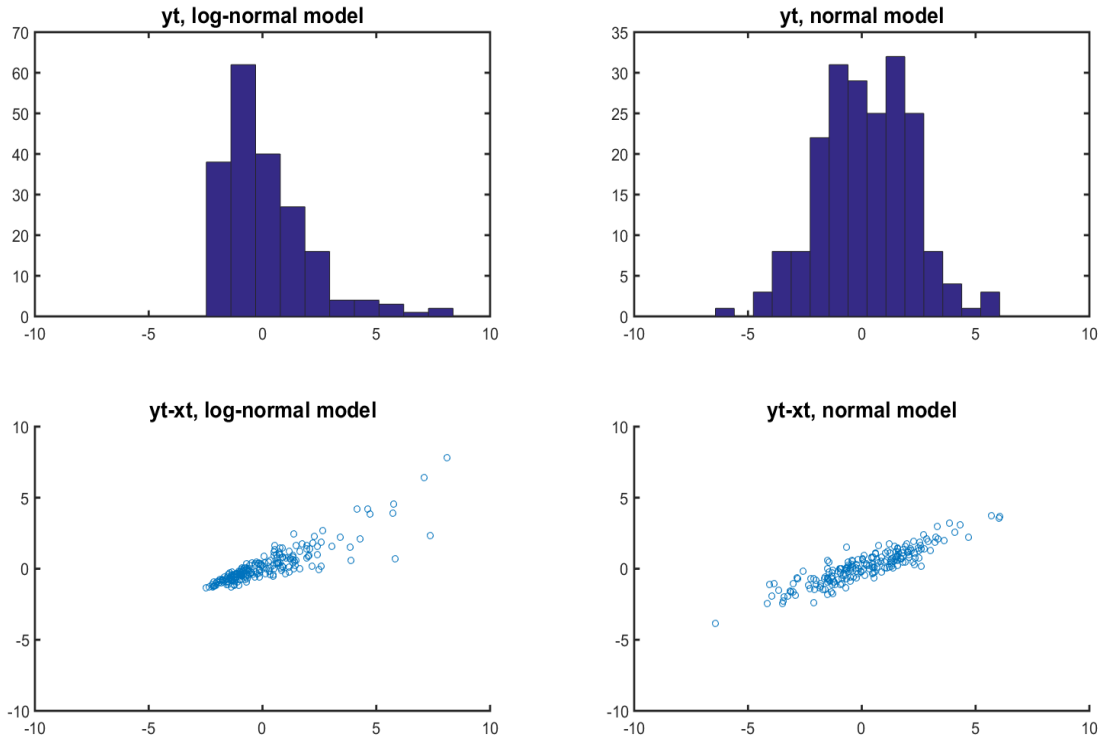
Then e_t is defined as

$$e_t = v_t - \mathbf{E}[v_t] \tag{4.4}$$

so that $\mathbf{E}[e_t] = 0$. The matrix Σ is the same covariance matrix that we used in the previous experiment. The parameter s is set to be 0.6 and c is chosen so that the interquartile range of $[y_t, x_t, z_{1,t}, z_{2,t}]$ is similar to the one based on the normal distribution. This formulation is very close to the one employed in Conley et al. [2008]. With this specification of the DGP, the instruments, y_t, and x_t are skewed and clearly deviate from the normal distribution (see Figure 1).

We repeat the model selection exercise with 16 different configurations as we did for the

Figure 1: Data generated from two different models (log-normal vs. normal model, one realization for each)



normal distribution case of Section 4.1. The results are summarized in Table 2. As in the normal distribution case, marginal likelihood correctly picks the right model.

4.3 Quantile IV regression

Another interesting application of our framework is quantile regression and quantile IV regression. There are at least two reasons why our framework is attractive for quantile regression models. First, the objective function for the quantile models is not smooth and this makes the numerical optimization problem hard. This is one of the motivations for Chernozhukov and Hong [2003] to consider integration-based estimation methods such as the Laplace type estimation. Lancaster and Jun [2010] also apply the ETEL likelihood to quantile regression models. Second, the ETEL framework provides a formal way to incorporate external information (such as commonality assumption) into the estimation of quantile models. It turns out that such an informative prior plays an important role in the

Table 3: IV Regression with log-Normal errors

	$log(p(Y\|M_i))$	(min, max)	$mean(\beta)$	90% Credible Set
		(a) M1 is DGP		
M1	**-1067.02**	(-1067.04, -1067.01)	0.949	(0.726, 1.172)
M2	-1070.4	(-1070.42, -1070.38)	0.978	(0.760, 1.190)
M3	-1074.23	(-1074.24, -1074.22)	0.979	(0.761, 1.209)
M4	-1071.07	(-1071.11, -1071.03)	0.973	(0.745, 1.190)
		(b) M2 is DGP		
M1	-1141.04	(-1141.05, -1141.02)	1.128	(1.060, 1.197)
M2	**-1071.29**	(-1071.31, -1071.27)	0.994	(0.907, 1.084)
M3	-1074.48	(-1074.50, -1074.46)	0.98	(0.803, 1.154)
M4	-1135.81	(-1135.84, -1135.78)	1.055	(0.984, 1.125)
		(c) M3 is DGP		
M1	-1174.92	(-1174.93, -1174.90)	1.064	(1.012, 1.114)
M2	-1081.02	(-1081.04, -1080.98)	1.221	(1.137, 1.305)
M3	**-1074.3**	(-1074.33, -1074.27)	0.957	(0.787, 1.124)
M4	-1177.32	(-1177.35, -1177.31)	1.051	(1.000, 1.102)
		(d) M4 is DGP		
M1	-1171.03	(-1171.03, -1171.02)	1.715	(1.613, 1.822)
M2	-1086.65	(-1086.67, -1086.61)	0.447	(0.232, 0.674)
M3	-1074.08	(-1074.09, -1074.05)	0.98	(0.762, 1.201)
M4	**-1071.28**	(-1071.30, -1071.25)	0.994	(0.764, 1.230)

context of quantile regressions Yang and He [2012].

In this subsection, we follow Chernozhukov and Hansen [2006] and Lancaster and Jun [2010] and consider the following data generating process,

$$\begin{aligned} y_t &= \beta(U_t)x_t + \alpha(U_t) \\ x_t &= c_{x,t} + \delta_1 z_{1,t} + \delta_2 z_{2,t} + e_{2,t} \\ z_{1,t} &= c_{z1} + e_{3,t} \\ z_{2,t} &= c_{z2} + e_{4,t} \end{aligned} \tag{4.5}$$

where $\alpha(U_t), e_{2,t}, e_{3,t}$, and $e_{4,t}$ are (standard) normally distributed and $corr(\alpha(U_t), e_{2,t}) = 0.6$ so that x_t is endogenous. We set $c_{z1} = c_{z2} = 0.5$, and $c_x = 0$. Moreover, $\beta(s) = 1$ for all $s \in [0, 1]$. Note that under this parameter configuration and distributional assumption,

this data generating process is the same as the linear IV regression example in the previous section. However, it differs in terms of moment conditions we are going to use to estimate the model. For any arbitrary $\tau \in [0,1]$, we can set the moment conditions as follows:

$$\begin{aligned}
\mathbf{E}[(1\{y_t \leq \alpha_\tau + x_t\beta_\tau\} - \tau)] &= 0 \\
\mathbf{E}[(1\{y_t \leq \alpha_\tau + x_t\beta_\tau\} - \tau)z_{1,t}] &= 0 \\
\mathbf{E}[(1\{y_t \leq \alpha_\tau + x_t\beta_\tau\} - \tau)z_{2,t}] &= v_2 \\
\mathbf{E}[(x_t - \delta_1 z_{1,t} - \delta_2 z_{2,t})z_{1,t}] &= 0 \\
\mathbf{E}[(x_t - \delta_1 z_{1,t} - \delta_2 z_{2,t})z_{2,t}] &= 0
\end{aligned} \tag{4.6}$$

The first three moment restrictions are based on so-called structural quantile function. The last two restrictions are the same as in the case of linear IV regression example. Note that the above moment conditions are satisfied for any value of $\tau \in [0,1]$. In the following exercise, we set $\tau = 0.5$.

As in the previous case, we suppose that we are confident that the first instrument, $z_{1,t}$ is valid and relevant. But we are not sure about the second instrument ($z_{2,t}$). Depending on the assumptions about v_2 and δ_2, we have four models. The prior distributions for the finite dimensional parameters $(\alpha_\tau, \beta_\tau, \delta_1, \delta_2, v_2)$ are relatively diffuse. Each element in this vector is set to be an independent normal distribution with mean zero and variance 10.

In what follows, we generate 500 observations from each of four models and compute the marginal likelihood based on four model specifications. This leads to 16 combinations. All results are summarized in Table 4. In this table, the first column presents the marginal likelihood estimated based on the method proposed by Chib and Jeliazkov [2001]. For all four cases, estimated marginal likelihood picks the true data generating process. In the next column, we present maximum and minimum value of marginal likelihood estimates based on multiple independent Metropolis-Hastings chains. For now, we compute marginal likelihood using $70,000$ draws from the posterior distribution. Then, we repeat this computation for three to seven times depending on the model specification[3]. This min/max range roughly gauges the accuracy of the marginal likelihood estimates. The last two columns present the posterior mean and 90% credible sets for the structural parameter β.

[3]We plan to increase the number of repetitions.

Table 4: Quantile IV Regression

	$log(p(Y\|M_i))$	(min, max)	$mean(\alpha)$	90% Credible Set	$mean(\beta)$	90% Credible Set
			(a) M1 is DGP			
M1	**-3121.65**	(-3121.74, -3121.55)	-0.002	(-0.116, 0.112)	1.03	(0.931, 1.115)
M2	-3124.24	(-3124.41, -3124.12)	-0.02	(-0.138, 0.097)	1.034	(0.938, 1.124)
M3	-3129.05	(-3129.39, -3128.75)	-0.022	(-0.141, 0.092)	1.034	(0.937, 1.123)
M4	-3126.64	(-3126.73, -3126.50)	-0.003	(-0.122, 0.112)	1.032	(0.934, 1.122)
			(b) M2 is DGP			
M1	-3428.52	(-3430.19, -3426.67)	-0.856	(-1.163, -0.499)	0.894	(0.790, 1.030)
M2	**-3124.83**	(-3124.96, -3124.72)	-0.007	(-0.148, 0.125)	1.001	(0.934, 1.065)
M3	-3129.06	(-3129.41, -3128.77)	-0.031	(-0.184, 0.110)	1.027	(0.939, 1.111)
M4	-3384.36	(-3384.50, -3384.10)	-0.649	(-0.738, -0.550)	1.196	(1.153, 1.241)
			(c) M3 is DGP			
M1	-3462.97	(-3464.03, -3462.70)	-0.58	(-0.694, -0.466)	1.191	(1.164, 1.222)
M2	-3184.73	(-3185.19, -3183.99)	-0.104	(-0.291, 0.017)	1.161	(1.085, 1.272)
M3	**-3129.28**	(-3129.57, -3129.02)	-0.021	(-0.172, 0.123)	1.018	(0.926, 1.108)
M4	-3467.49	(-3469.02, -3464.33)	-0.464	(-0.582, -0.361)	1.15	(1.088, 1.205)
			(d) M4 is DGP			
M1	-3205.81	(-3205.93, -3205.74)	0.275	(0.258, 0.300)	0.669	(0.647, 0.696)
M2	-3177.46	(-3178.36, -3176.85)	0.221	(0.128, 0.304)	0.645	(0.527, 0.730)
M3	-3129.25	(-3129.48, -3128.92)	-0.021	(-0.139, 0.095)	1.033	(0.937, 1.122)
M4	**-3125.2**	(-3125.27, -3125.14)	-0.014	(-0.135, 0.102)	1.029	(0.932, 1.120)

References

D. W. Andrews and B. Lu. Consistent model and moment selection procedures for GMM estimation with application to dynamic panel data models. *Journal of Econometrics*, 101 (1):123–164, 2001.

D. W. K. Andrews. Consistent moment selection procedures for generalized method of moments estimation. *Econometrica*, 67(3):543–563, 1999.

L. Bornn, N. Shephard, and R. Solgi. Moment conditions and bayesian nonparametrics. Technical report, arXiv:1507.08645, 2015.

J. Y. Campbell and J. H. Cochrane. By force of habit: A consumption based explanation of aggregate stock market behavior. *Journal of Political Economy*, 107(2):205–251, 1999.

G. Chamberlain and G. W. Imbens. Nonparametric applications of bayesian inference. *Journal of Business & Economic Statistics*, 21(1):12–18, 2003.

X. Cheng and Z. Liao. Select the valid and relevant moments: An information-based LASSO for GMM with many moments. *Journal of Econometrics*, 186(2):443–464, 2015.

V. Chernozhukov and C. B. Hansen. Instrumental quantile regression inference for structural and treatment effect models. *Journal of Econometrics*, 132(2):491–525, 2006.

V. Chernozhukov and H. Hong. An MCMC approach to classical estimation. *Journal of Econometrics*, 115(2):293–346, 2003.

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

S. Chib and E. Greenberg. Understanding the metropolis-hastings algorithm. *The American Statistician*, 49(4):327–335, 1995.

S. Chib and I. Jeliazkov. Marginal likelihood from the metropolis hastings output. *Journal of the American Statistical Association*, 96(453):270–281, 2001.

T. G. Conley, C. B. Hansen, R. E. McCulloch, and P. E. Rossi. A semi-parametric bayesian approach to the instrumental variable problem. *Journal of Econometrics*, 144(1):276–305, 2008.

I. Csiszar. i-divergence geometry of probability distributions and minimization problems. *Ann. Probab.*, 3(1):146–158, 1975.

F. J. DiTraglia. Using invalid instruments on purpose: Focused moment selection and avaeraging for gmm. Technical report, University of Pennsylvania, 2015.

S. G. Donald and W. K. Newey. Choosing the number of instruments. *Econometrica*, 69(5): 1161–1191, 2001.

J.-P. Florens and J.-M. Rolin. Bayes, bootstrap, moments. Technical report, Université Catholique de Louvain, 1994.

J.-P. Florens and A. Simoni. Gaussian processes and bayesian moment estimation. Technical report, Crest, 2015.

A. R. Gallant. Reflections on the probability space induced by moment conditions with implications for bayesian inference. *Journal of Financial Econometrics*, 2015.

A. R. Gallant, R. Giacomini, and G. Ragusa. Bayesian estimation of state space models using moment conditions. Technical report, 2015.

A. R. Hall, A. Inoue, K. Jana, and C. Shin. Information in generalized method of moments estimation and entropy-based moment selection. *Journal of Econometrics*, 138(2):488–512, 2007.

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

H. Hong, B. Preston, and M. Shum. Generalized empirical likelihood-based model selection criteria for moment condition models. *Econometric Theory*, pages 923–943, 12 2003.

A. Inoue. A bootstrap approach to moment selection. *Econometrics Journal*, 9(1):48–75, 2006.

A. Inoue and M. Shintani. Quasi-bayesian model selection. Technical report, Vanderbilt University, 2015.

J.-Y. Kim. Limited information likelihood and bayesian analysis. *Journal of Econometrics*, 107(1-2):175–193, 2002.

Y. Kitamura and T. Otsu. Bayesian analysis of moment condition models using nonparametric priors. Technical report, 2011.

B. Kleijn and A. van der Vaart. The bernstein-von-mises theorem under misspecification. *Electron. J. Statist.*, 6:354–381, 2012.

Y. K. Kwan. Asymptotic bayesian analysis based on a limited information estimator. *Journal of Econometrics*, 88(1):99–121, 1999.

T. Lancaster and S. J. Jun. Bayesian quantile regression methods. *Journal of Applied Econometrics*, 25(2):287–307, 2010.

N. A. Lazar. Bayesian empirical likelihood. *Biometrika*, 90(2):319–326, 2003.

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

Z. Liao. Adaptive gmm shrinkage estimation with consistent moment selection. *Econometric Theory*, 29:857–904, 2013. ISSN 1469-4360.

G. Ragusa. Bayesian likelihoods for moment condition models. Working Papers 060714, University of California-Irvine, Department of Economics, 2007.

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

S. M. Schennach. Point estimation with exponentially tilted empirical likelihood. *Ann. Statist.*, 35(2):634–672, 04 2007.

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

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

Y. Yang and X. He. Bayesian empirical likelihood for quantile regression. *The Annals of Statistics*, 40(2):1102–1131, 2012.