Received: 1 November 2019 | Accepted: 26 September 2021

DOI: 10.1111/rssb.12484

ORIGINAL ARTICLE

ROYAL STATISTICAL SOCIETY Series B Statistical Methodology

Bayesian estimation and comparison of conditional moment models

Siddhartha Chib[1] | **Minchul Shin**[2] | **Anna Simoni**[3]

[1]Olin Business School, Washington University in St. Louis, St. Louis, Missouri, USA

[2]Federal Reserve Bank of Philadelphia, Philadelphia, Pennsylvania, USA

[3]CREST, CNRS, École Polytechnique, ENSAE, Palaiseau, France

Correspondence
Siddhartha Chib, Olin Business School, Washington University in St. Louis, Campus Box 1133, 1 Brookings Drive, St. Louis, Missouri 63130, USA.
Email: chib@wustl.edu

Abstract

We consider the Bayesian analysis of models in which the unknown distribution of the outcomes is specified up to a set of conditional moment restrictions. The non-parametric exponentially tilted empirical likelihood function is constructed to satisfy a sequence of unconditional moments based on an increasing (in sample size) vector of approximating functions (such as tensor splines based on the splines of each conditioning variable). For any given sample size, results are robust to the number of expanded moments. We derive Bernstein–von Mises theorems for the behaviour of the posterior distribution under both correct and incorrect specification of the conditional moments, subject to growth rate conditions (slower under misspecification) on the number of approximating functions. A large-sample theory for comparing different conditional moment models is also developed. The central result is that the marginal likelihood criterion selects the model that is less misspecified. We also introduce sparsity-based model search for high-dimensional conditioning variables, and provide efficient Markov chain Monte Carlo computations for high-dimensional parameters. Along with clarifying examples, the framework is illustrated with real data applications to risk-factor determination in finance, and causal inference under conditional ignorability.

KEYWORDS

Bayesian inference, Bernstein–von Mises theorem, conditional moment restrictions, exponentially tilted empirical likelihood, marginal likelihood, misspecification, posterior consistency

1 | INTRODUCTION

We tackle the problem of prior–posterior inference when the only available information about the unknown parameter $\theta \in \Theta \subset \mathbb{R}^p$ is supplied by a set of *conditional moment* (CM) restrictions

$$\mathbf{E}^P[\rho(X, \theta)|Z] = 0, \tag{1}$$

where $\rho(X, \theta)$ is a d-vector of known functions of a $\mathbb{R}^{d_x}$-valued random vector X and the unknown θ, and P is the unknown conditional distribution of X given a $\mathbb{R}^{d_z}$-valued random vector Z. Such models are important because many standard models in statistics can be recast in terms of CM restrictions. These models also arise naturally in causal inference, missing data problems and in models derived from theory in economics and finance. Because the CM conditions constrain the set of possible distributions P, we say that the model is correctly specified if the true data generating process P_* is in the set of distributions constrained to satisfy these moment conditions for some $\theta \in \Theta$, while the model is misspecified if P_* is not in the set of implied distributions for any $\theta \in \Theta$.

A different starting point is when one is given the *unconditional* moments, say $\mathbf{E}^P[g(X, \theta)] = 0$. Prior–posterior analysis can then be based on the empirical likelihood, for example, Lazar (2003) and many others, or the exponentially tilted empirical likelihood (ETEL), as in Schennach (2005) and Chib et al. (2018). Developing a Bayesian framework for CM models is important. While it is true that the conditional moments imply that $\rho(X, \theta)$ is uncorrelated with Z, that is, $\mathbf{E}^P[\rho(X, \theta) \otimes Z] = 0$, where $\otimes$ is the Kronecker product operator, the conditional moments assert even more, that $\rho(X, \theta)$ is uncorrelated with any measurable, bounded function of Z. Thus, there is an efficiency loss if this information is ignored.

We approach this problem by first constructing K unconditional moments

$$\mathbf{E}^P[\rho(X, \theta) \otimes q^K(Z)] = 0 \tag{2}$$

based on an increasing (in sample size) vector of approximating functions, $q^K(Z) := (q_1^K(Z), \ldots, q_K^K(Z))'$, obtained, for instance, from splines of each variable in Z (Donald et al., 2003). Efficiency loss is avoided as the number of moments increases with sample size. Next, for each sample size and for each θ, the non-parametric ETEL function is constructed to satisfy these unconditional moments. Unlike the empirical likelihood, the ETEL function has a fully Bayesian interpretation. It is the likelihood that emerges from integrating out P with respect to a non-parametric prior that satisfies the CMs. The posterior of interest is then this non-parametric likelihood multiplied by a prior distribution of the parameters. Due to the fact that the non-parametric likelihood is limited to a set $H_{n,K}$ of θ values for which the empirical counterpart of the moment conditions (2) is equal to 0, the posterior (equivalently, the prior) is truncated to the set $H_{n,K}$.

We study the prior–posterior mapping on many fronts, taking up the question of misspecified models, model comparisons and computations, combining careful theoretical work with the needs of applications. The posterior distribution is shown to satisfy Bernstein–von Mises (BvM) theorems in both the correct and misspecified cases. In the former case the growth of K (for approximating functions given by splines) is at most $n^{1/6}$, where n is the sample size. The asymptotic posterior variance is then equal to the semiparametric efficiency bound derived in Chamberlain (1987). In the latter case, in parallel with Kleijn and van der Vaart (2012), the

posterior distribution of the centred and scaled parameter $\sqrt{n}(\theta - \theta_\circ)$, where $\theta_\circ$ is the pseudo-true value, converges to a normal distribution with variance that now is different from the variance of the frequentist estimator. Interestingly, this convergence holds only if K increases more slowly than in the correctly specified case. This can be interpreted as limiting the number of implied unconditional moments to limit the magnification of the misspecification.

We informally use these rate conditions from the theoretical analysis to guide the range of choice of K for any given n. Due to the fact that for a fixed n the volume (prior probability content) of the region of truncation $H_{n,K}$ decreases with K (a result of more restrictions), values of K beyond the range recommended by the theory amplify the Bayesian bias, and, hence, should be avoided. Large values of K can also produce rank deficiency of the approximating functions basis matrix and, in the event of a misspecified model, increase misspecification. Around the values of K we recommend, the posterior distribution is generally robust to K, and little fine-tuning is necessary.

Finite sample summaries of the posterior distribution are obtained by Markov chain Monte Carlo (MCMC) methods. Since the posterior is underpinned by a non-parametric likelihood, and the effective prior is truncated, efficient sampling is not automatic. However, after extensive study, we have produced a near-black-box MCMC approach (available as a R-package) that is based on the tailored Metropolis–Hastings (M-H) algorithm of Chib and Greenberg (1995) and its randomized version in Chib and Ramamurthy (2010).

The entire paper is interspersed with examples of pedagogical importance and practical relevance. Real data applications to risk-factor determination in finance, and causal inference under conditional ignorability, are included.

It is worth noting that previous Bayesian work on conditional moments, for example, Liao and Jiang (2011), Florens and Simoni (2012, 2016), Kato (2013), Chen et al. (2018) and Liao and Simoni (2019), has little overlap with the discussion here. A major difference is that none of these papers adopt the fully Bayesian ETEL framework. Another is that these papers examine a different class of CM models. Finally, none of these papers takes up the question of model comparisons. Nonetheless, these papers and the current work, taken together, represent an important broadening of the Bayesian enterprise to new classes of models.

The rest of the paper is organized as follows. Section 2 has the sketch of the conditional moment setting. Section 3 discusses the prior–posterior analysis and the large-sample properties of the posterior distribution. Section 4 is concerned with the problem of comparing CM models via marginal likelihoods. In Section 5 two extensions are considered and Section 6 has real data applications to finance and causal inference. Section 7 concludes. Proofs are in the online supplementary appendix.

2 | SETTING AND MOTIVATION

Let $X := (X_1', X_z')'$ be an $\mathbb{R}^{d_x}$-valued random vector and $Z := (Z_1', X_z')'$ be an $\mathbb{R}^{d_z}$-valued random vector. The vectors Z and X have elements in common if the dimension of the subvector X_z is non-zero. Moreover, we denote $W := (X', Z_1')' \in \mathbb{R}^{d_w}$ and its (unknown) joint distribution by P. By abuse of notation, let P also denote the associated conditional distribution. Suppose that we are given a random sample $W_{1:n} = (W_1, \ \dots \ , W_n)$ of W. Hereafter, $\mathbf{E}^P[\cdot]$ is the expectation with respect to P and $\mathbf{E}^P[\cdot|\cdot]$ is the conditional expectation with respect to the conditional distribution associated with P.

The parameter of interest is $\theta \in \Theta \subset \mathbb{R}^p$, which is related to the conditional distribution P through the conditional moment restrictions

$$\mathbf{E}^P[\rho(X, \theta)|Z] = 0, \tag{3}$$

where $\rho(X, \theta)$ is a d-vector of known functions. Many interesting and important models in statistics fall into this framework.

Example 1 (Linear model with heteroscedasticity of unknown form). Suppose that

$$\mathbf{E}^P[(Y - \theta_0 - \theta_1 X)|Z] = 0, \tag{4}$$

where $\rho(X, \theta) = (Y - \theta_0 - \theta_1 X)$, $Z = (1, X)$ and $d = 1$. This CM model is consistent with the data generating process (DGP) $Y = \theta_0 + \theta_1 X + \varepsilon$, where $\varepsilon = h(X)U$, and (X, U) (independent) follow some unknown distribution P, with $E(U) = 0$, and the heteroscedasticity function $h(X)$ is unknown. The restrictions

$$\mathbf{E}^P[(Y - \theta_0 - \theta_1 X)\,|Z] = 0 \quad \text{and} \quad \mathbf{E}^P[(Y - \theta_0 - \theta_1 X)^3|Z] = 0, \tag{5}$$

where now $\rho(X, \theta)$ is a (2×1) vector of functions, additionally impose that ε is conditionally symmetric.

Note that in the foregoing example, the two unconditional moment conditions

$$\mathbf{E}^P[(Y - \theta_0 - \theta_1 X) \otimes (1, X)'] = 0, \tag{6}$$

which assert that: (i) ε has mean zero and (ii) ε is uncorrelated with X, are weaker but, if the CM model is correct, less informative about θ.

3 | PRIOR–POSTERIOR ANALYSIS

3.1 | Expanded moment conditions

The starting point, as in the frequentist approaches of Donald and Newey (2001), Ai and Chen (2003) and Carrasco and Florens (2000), is a transformation of the CM restrictions into unconditional moment restrictions. Following Donald et al. (2003), let $q^K(Z) := (q_1^K(Z), \ \dots \ , q_K^K(Z))'$, $K > 0$, denote a K-vector of real-valued functions of Z, for instance, splines. Suppose that these functions satisfy the following condition for the distribution P.

Assumption 3.1 For all K, $\mathbf{E}^P[q^K(Z)'q^K(Z)]$ is finite, and for any function $a(z) : \mathbb{R}^{d_z} \to \mathbb{R}$ with $\mathbf{E}^P[a(Z)^2] < \infty$ there are $K \times 1$ vectors γ_K such that as $K \to \infty$,

$$\mathbf{E}^P[(a(Z) - q^K(Z)'\gamma_K)^2] \to 0.$$

Now, let θ_* be the value of θ that satisfies (3) for the true P. If $\mathbf{E}^P[\rho(X, \theta_*)'\rho(X, \theta)] < \infty$, then Donald et al. (2003 Lemma 2.1) established that: (1) if Equation (3) is satisfied with $\theta = \theta_*$, then $\mathbf{E}^P[\rho(X, \theta_*) \otimes q^K(Z)] = 0$ for all K; (2) if Equation (3) is not satisfied by $\theta = \theta_*$, then $\mathbf{E}^P[\rho(X, \theta_*) \otimes q^K(Z)] \neq 0$, for all large enough K.

Henceforth, we let $g(W, \theta) := \rho(X, \theta) \otimes q^K(Z)$ denote the *expanded functions* and refer to

$$\mathbf{E}^P[g(W, \theta)] = 0, \tag{7}$$

as the *expanded moments*. Under the stated assumptions, the expanded moments are equivalent to the CM restrictions (3), as $K \to \infty$.

In our numerical examples, we construct $q^K(Z)$ using the natural cubic spline basis of Chib and Greenberg (2010), with K fixed at a given value, as in sieve estimation. If Z consists of more than one element, say (Z_1, Z_2, Z_3) where Z_1 and Z_2 are continuous variables and Z_3 is binary, then the basis matrix B is constructed as follows. Let $\boldsymbol{z}_j$ denote the $n \times 1$ sample data on Z_j $(j \leq 3)$. Let $\boldsymbol{Z} = (\boldsymbol{z}_1, \boldsymbol{z}_2, \boldsymbol{z}_1 \odot \boldsymbol{z}_2, \boldsymbol{z}_1 \odot \boldsymbol{z}_3, \boldsymbol{z}_2 \odot \boldsymbol{z}_3)$ denote the $n \times 5$ matrix of the continuous data and interactions of the continuous data and the binary data. Now suppose $(\tau_{j1}, \ldots, \tau_{jK})$, for $j = 1, \ldots, 5$ are K knots based on each column of $\boldsymbol{Z}$ and let B_j denote the corresponding $n \times K$ matrix of cubic spline basis functions. Then, B is given by

$$B = \left[B_1 \vdots B_2^* \vdots B_3^* \vdots B_4^* \vdots B_5^* \vdots \boldsymbol{Z}_3\right],$$

where B_j^* $(j = 2, 3, 4, 5)$ is the $n \times (K - 1)$ matrix in which each column of B_j is subtracted from its first and then the first column is dropped, see Chib and Greenberg (2010). Thus, the dimension of this B matrix is $n \times K^*$, where $K^* = (5K - 4 + 1)$. If K^* is large, in relation to n, data-compression methods can be employed. Specifically, let R denote the $K^* \times K^*$ orthogonal matrix of eigenvectors from the singular value decomposition of B, and let e denote the corresponding $K^* \times 1$ vector of eigenvalues. Then, after employing the rotation BR, the columns of BR corresponding to small values of e are dropped, and the resulting column-reduced BR matrix is taken as the basis matrix. We refer to this as the *rotated column reduced* basis matrix. To define the expanded functions, let $\rho_l(\boldsymbol{X}, \theta)$ $(l \leq d)$ denote a $n \times 1$ vector of the lth element of $\rho(X, \theta)$ evaluated at the sample data matrix $\boldsymbol{X}$. Then, the expanded functions for the sample observations are obtained by multiplying $\rho_l(\boldsymbol{X}, \theta)$ by the matrix B (or by the rotated column reduced B) and concatenating. We use versions of this approach in our examples.

3.2 | Posterior distribution

We base the prior–posterior mapping, for each sample size and θ, on the non-parametric ETEL function. The ETEL has a fully Bayesian interpretation (Schennach, 2005) as an integrated likelihood, integrated over the prior on the data distribution P that satisfies the given moments. Other such priors exist, for example, Kitamura and Otsu (2011), Shin (2014) and Florens and Simoni (2021), that lead to different integrated likelihoods.

The ETEL function takes the form

$$p(W_{1:n}|\theta, K) = \prod_{i=1}^{n} \hat{p}_i(\theta), \tag{8}$$

where $\{\hat{p}_i(\theta), i = 1, \ldots, n\}$ are the probabilities that minimize the Kullback–Leibler 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 satisfy the given unconditional moment conditions (7).

Specifically, $\{\hat{p}_i(\theta), i = 1, \ldots, n\}$ are the solution of the following problem:

$$\max_{p_1, \cdots, p_n} \sum_{i=1}^{n} [-p_i \log(np_i)] \quad \text{subject to: } \sum_{i=1}^{n} p_i = 1, \quad \sum_{i=1}^{n} p_i g(w_i, \theta) = 0, \quad p_i \geq 0 \tag{9}$$

see Schennach (2005) for a proof). In practice, the solution of this problem emerges from the dual (saddlepoint) representation (see e.g. Csiszar, 1984) as

$$\hat{p}_i(\theta) := \frac{e^{\hat{\lambda}(\theta)' g(w_i, \theta)}}{\sum_{j=1}^{n} e^{\hat{\lambda}(\theta)' g(w_j, \theta)}}, \quad i = 1, \ldots, n, \tag{10}$$

where $\hat{\lambda}(\theta) = \arg \min_{\lambda \in \mathbb{R}^{dK}} \frac{1}{n} \sum_{i=1}^{n} e^{\lambda' g(w_i, \theta)}$ is the estimated tilting parameter.

Let $Co(\theta) := \left\{ \sum_{i=1}^{n} p_i g(w_i, \theta), p_i \geq 0, \sum_{i=1}^{n} p_i = 1 \right\}$ be the convex hull of $\{g(w_i, \theta)\}_{i=1}^{n}$ for a given θ and $\overline{Co(\theta)}$ denote its interior. Let $H_{n,K} := \{\theta \in \Theta; 0 \in \overline{Co(\theta)}\}$ denote the set of θ values for which the empirical moment conditions hold. Then, the posterior distribution is the truncated distribution given by

$$\pi(\theta | w_{1:n}, K) \propto \pi(\theta) \prod_{i=1}^{n} \frac{e^{\hat{\lambda}(\theta)' g(w_i, \theta)}}{\sum_{j=1}^{n} e^{\hat{\lambda}(\theta)' g(w_j, \theta)}} 1\{\theta \in H_{n,K}\}, \tag{11}$$

where $1\{\cdot\}$ is the indicator function.

Combining the indicator function with the prior, we see that the (effective) prior is truncated to $\theta \in H_{n,K}$. This fact can be used to argue that, for fixed n, it is not desirable to have a large K. This is because as K increases for a given n, the support of the prior shrinks (equivalently, the prior probability content of the region of truncation decreases), due to the fact that more restrictions are imposed. We refer to this prior probability content by the shorthand, volume. Reduction in the volume tends to increase the Bayesian bias and reduce the posterior spread, without any change in the data, with deleterious impact on the posterior. In practice, we use the rule $2n^{1/6}$ to fix K. Larger values than this can, of course, be tried, but one should make sure that the volume of $H_{n,K}$ does not become much smaller than one. Around the values of K we recommend, the posterior distribution is generally robust to K, and little fine-tuning is necessary.

Example 1 (continued). To illustrate the role of K in the prior–posterior analysis, and its impact on the volume (prior probability content) of $H_{n,K}$, we create a set of simulated data $\{y_i, x_i\}_{i=1}^{n}$, $n = 250$, with covariates $X \sim \mathcal{U}(-1, 2.5)$, intercept $\theta_0 = 1$, slope $\theta_1 = 1$, and ε_i is distributed according to $\varepsilon_i \sim \mathcal{SN}(m(x_i), h(x_i), s(x_i))$, where $\mathcal{SN}(m, h, s)$ is the skew normal distribution with location, scale and shape parameters given by (m, h, s), each depending on x_i. When s is zero, ε_i is normal with mean m and standard deviation h. We set $m(x_i) = -h(x_i)\sqrt{2/\pi} s(x_i)/(\sqrt{1 + s(x_i)^2})$, so that $\mathbf{E}^P[\varepsilon | X] = 0$.

Suppose that $h(x) = \sqrt{\exp(1 + 0.7x + 0.2x^2)}$ and $s(x) = 1 + x^2$. Model parameters are estimated solely from the condition $\mathbf{E}^P[\varepsilon | Z] = 0$, $Z = (1, X)$. The prior is the default independent student-t distribution with location 0, dispersion 5 and degrees of freedom 2.5,

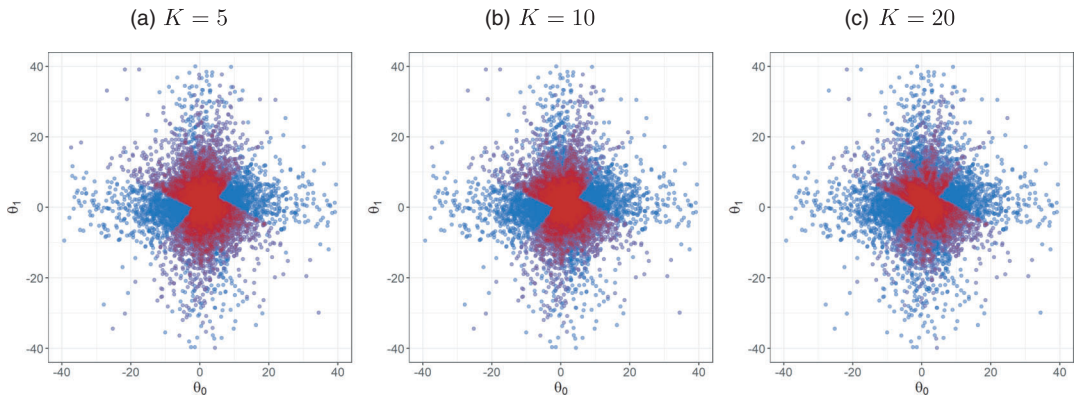


FIGURE 1 Example 1 ($n = 250$): This figure visualizes the volume of the convex hull, $H_{250,K}$, for different values of K. Blue dots are 10,000 independent draws from the default student-t prior. Red dots are the corresponding points in $H_{n,K}$. The volume of $H_{n,K}$ shrinks as K increases (the prior support decreases), just as shown by the volume estimates presented in Table 1. The range in the figures is set to $(-40, 40)$ for visualization clarity

truncated to $H_{n,K}$. The posterior is computed for K given by 2, $2n^{1/6}$, 9 and 20 (the value $2n^{1/6}$ is based on the theory below for splines approximating functions). The results are shown in Table 1. Importantly, when K is close to the value suggested by theory, the posterior distribution is robust to K. However, when $K = 20$, quite different from the recommended value, the Bayesian bias is larger and the posterior standard deviation is smaller, without any change in the data. This is due to the effect of the prior, in the following way. As K increases for a fixed n, the volume of $H_{n,K}$ decreases, equivalently, the support of the prior distribution shrinks, as illustrated in Figure 1. This explains why values of K close to the recommended value are preferred.

As an aside, if the true model was unconditional (i.e. without conditional heteroscedasticity), then there is little loss in using more expanded moments—the extra moments are superfluous and hence do not change the effective support of the prior. In that case, no tangible cost in imposed, apart from the computational burden of carrying those moments along.

3.3 | Asymptotic properties

Consider now the large-sample behaviour of the posterior distribution of θ. We let θ_* and P_*, respectively, denote the true value of θ and of the data distribution P. As notation, when the true distribution P_* is involved, expectations $\mathbf{E}^P[\cdot]$ (resp. $\mathbf{E}^P[\cdot|\cdot]$) are taken with respect to P_* (resp. the conditional distribution associated with P_*). In addition, we denote $\ell_{n,\theta}(W_i) := \log \hat{p}_i(\theta) = \log \frac{e^{\hat{\lambda}(\theta)' g(W_i,\,\theta)}}{\sum_{j=1}^n e^{\hat{\lambda}(\theta)' g(W_j,\,\theta)}}$,

$$\rho_\theta(X,\theta) := \frac{\partial \rho(X,\theta)}{\partial \theta'}, \quad D(Z) := \mathbf{E}^P[\rho_\theta(X,\theta_*)|Z],$$

$$\Sigma(Z) := \mathbf{E}^P[\rho(X,\theta_*)\rho(X,\theta_*)'|Z], \quad \text{and} \quad \rho_{j\theta\theta}(X,\theta_*) := \partial^2 \rho_j(X,\theta_*)/\partial\theta\partial\theta'.$$

TABLE 1 Example 1 ($n = 250$): Volume (prior probability content) of the convex hull (an estimate of the prior support) and posterior summary for K given by 2, $2n^{1/6}$ and 10, 15, 20

	$Vol(H_{n,K})$		Mean	SD	Median	Lower	Upper	Ineff
$K=2$	0.76	θ_1	1.07	0.10	1.07	0.88	1.27	1.10
		θ_2	1.01	0.14	1.01	0.74	1.29	1.14
$K=5$	0.73	θ_1	1.07	0.10	1.07	0.88	1.26	1.15
		θ_2	1.03	0.12	1.03	0.80	1.25	1.08
$K=10$	0.68	θ_1	1.07	0.09	1.07	0.89	1.25	1.14
		θ_2	1.02	0.11	1.02	0.79	1.25	1.13
$K=15$	0.60	θ_1	0.98	0.07	0.98	0.85	1.12	1.14
		θ_2	1.10	0.10	1.10	0.91	1.29	1.13
$K=20$	0.54	θ_1	0.99	0.07	0.99	0.86	1.13	1.11
		θ_2	1.12	0.09	1.12	0.94	1.31	1.11

Notes: Results based on 20,000 Markov chain Monte Carlo (MCMC) draws beyond a burn-in of 1000. 'Lower' and 'Upper' refer to the 0.05 and 0.95 quantiles of the simulated draws, respectively, and 'Ineff' to the inefficiency factor.

For a vector a, $\|a\|$ denotes the Euclidean norm. For a matrix A, $\|A\|$ denotes the operator norm (the largest singular value of the matrix). Finally, let $\mathcal{Z} := \text{supp}(Z)$ denote the support of Z.

The first assumption is a normalization for the second moment matrix of the approximating functions which is standard in the literature, see for example, Newey (1997) and Donald et al. (2003).

Assumption 3.2 For each K there is a constant scalar $\zeta(K)$ such that $\sup_{z\in\mathcal{Z}} \|q^K(z)\| \leq \zeta(K)$, $\mathbf{E}^P[q^K(Z)q^K(Z)']$ has smallest eigenvalue bounded away from zero uniformly in K, and $\sqrt{K} \leq \zeta(K)$.

The bound $\zeta(K)$ is known explicitly in a number of cases depending on the approximating functions we use. Donald et al. (2003) provide a discussion and explicit formulas for $\zeta(K)$ in the case of splines, power series and Fourier series. We also refer to Newey (1997) for primitive conditions for regression splines and power series.

Assumption 3.3 (a) There exists a unique $\theta_* \in \Theta$ that satisfies $\mathbf{E}^P[\rho(X,\theta)|Z] = 0$ for the true P_*; (b) the data $W_i := (X_i, Z_{1i})$, $i = 1, \ldots, n$ are i.i.d. according to P_*; (c) $\mathbf{E}^P[\sup_{\theta\in\Theta} \|\rho(X,\theta)\|^2|Z]$ is bounded.

This assumption is the same as Donald et al. (2003 Assumption 3). The following three assumptions are also the same as the ones required by Donald et al. (2003) to establish asymptotic normality of the generalized empirical likelihood (GEL) estimator.

Assumption 3.4 (a) $\theta_* \in int(\Theta)$; (b) $\rho(X,\theta)$ is twice continuously differentiable in a neighbourhood $\mathcal{U}$ of θ_*, $\mathbf{E}^P[\sup_{\theta\in\mathcal{U}} \|\rho_\theta(X,\theta)\|^2|Z]$ and $\mathbf{E}^P[\sup_{\theta\in\mathcal{U}} \|\rho_{j\theta\theta}(X,\theta_*)\|^2|Z]$, $j = 1, \ldots d$, are bounded on $\mathcal{Z}$; (c) $\mathbf{E}^P[D(X)D(X)']$ is nonsingular.

Assumption 3.5 (a) $\Sigma(Z)$ has smallest eigenvalue bounded away from zero; (b) for a neighbourhood $\mathcal{U}$ of θ_*, $\mathbf{E}^P[\sup_{\theta\in\mathcal{U}} \|\rho(X,\theta)\|^4|z]$ is bounded, and for all $\theta \in \mathcal{U}$, $\|\rho(X,\theta) - \rho(X,\theta_*)\| \leq \delta(X)\|\theta - \theta_*\|$ and $\mathbf{E}^P[\delta(X)^2|Z]$ is bounded.

Assumption 3.6 There is $\gamma > 2$ such that $\mathbf{E}^P[\sup_{\theta\in\Theta} \|\rho(X,\theta)\|^\gamma] < \infty$ and $\zeta(K)^2K/n^{1-2/\gamma} \to 0$.

Part (b) of Assumption 3.5 imposes a Lipschitz condition which allows application of uniform convergence results. The last assumption is about the prior distribution of θ and is standard in the Bayesian literature on frequentist asymptotic properties of Bayes procedures.

Assumption 3.7 (a) π is a continuous probability measure that admits a density with respect to the Lebesgue measure; (b) π is positive on a neighbourhood of θ_*.

We are now able to state our first major result in which we establish the asymptotic normality and efficiency of the posterior distribution of the local parameter $h := \sqrt{n}(\theta - \theta_*)$.

Theorem 3.1 (Bernstein-von Mises). *Under Assumptions* 3.1–3.7, *if* $K \to \infty$, $\zeta(K)K^2/\sqrt{n} \to 0$, *and if for any* $\delta > 0$, $\exists \varepsilon > 0$ *such that as* $n \to \infty$

$$P\left(\sup_{\|\theta - \theta_*\| > \delta} \frac{1}{n} \sum_{i=1}^{n} \left(\ell_{n,\theta}(W_i) - \ell_{n,\theta_*}(W_i) \right) \leq -\varepsilon \right) \to 1, \tag{12}$$

then the posterior distribution $\pi(\sqrt{n}(\theta - \theta_*)|W_{1:n})$ *converges in total variation towards a random normal distribution, that is*,

$$\sup_B \left| \pi(\sqrt{n}(\theta - \theta_*) \in B | W_{1:n}, K) - \mathcal{N}_{\Delta_{n,\theta_*}, V_{\theta_*}}(B) \right| \xrightarrow{p} 0, \tag{13}$$

where $B \subseteq \Theta$ *is any Borel set*, $\Delta_{n,\theta_*} := -\frac{1}{\sqrt{n}} \sum_{i=1}^{n} V_{\theta_*} D(Z_i)' \Sigma(Z_i)^{-1} \rho(X_i, \theta_*)$ *is bounded in probability and* $V_{\theta_*} := (\mathbf{E}^p[D(Z)'\Sigma(Z)^{-1}D(Z)])^{-1}$.

We note that the centring Δ_{n,θ_*} of the limiting normal distribution satisfies $\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{d \log \hat{p}_i(\theta_*)}{d\theta} - V_{\theta_*}^{-1} \Delta_{n,\theta_*} \xrightarrow{p} 0$. We also note that the condition $\zeta(K)K^2/\sqrt{n} \to 0$ in the theorem implies $K/n \to 0$, which is a classical condition in the sieve literature. This condition is required to establish a stochastic local asymptotic normality (LAN) expansion, which is an intermediate step to prove the BvM result, as we explain below. The LAN expansion is not required to establish asymptotic normality of the GEL estimators, which explains why our condition is slightly stronger than the condition $\zeta(K)K/\sqrt{n} \to 0$ required by Donald et al. (2003). On the other hand, our condition is weaker than the condition $\zeta(K)^2K^2/\sqrt{n} \to 0$ required by Donald et al. (2009) to establish the mean square error of the GEL estimators. The asymptotic covariance of the posterior distribution coincides with the semiparametric efficiency bound given in Chamberlain (1987) for conditional moment condition models. This means that, for every $\alpha \in (0, 1)$, $(1-\alpha)$-credible regions constructed from the posterior of θ are $(1 - \alpha)$-confidence sets asymptotically.

The proof of this theorem is given in the supplementary appendix and consists of three steps. In the first step we show consistency of the posterior distribution of θ, namely:

$$\pi(\sqrt{n}\|\theta - \theta_*\| > M_n | W_{1:n}, K) \xrightarrow{p} 0 \tag{14}$$

for any $M_n \to \infty$, as $n \to \infty$. To show this, the identification assumption (12) is used. In the second step we show that the ETEL function satisfies a stochastic LAN expansion:

$$\sup_{h \in \mathcal{H}} \left| \sum_{i=1}^{n} \ell_{n,\theta_* + h/\sqrt{n}}(W_i) - \sum_{i=1}^{n} \ell_{n,\theta_*}(W_i) - h' V_{\theta_*}^{-1} \Delta_{n,\theta_*} + \frac{1}{2} h' V_{\theta_*}^{-1} h \right| = o_p(1), \tag{15}$$

where $\mathcal{H}$ denotes a compact subset of $\mathbb{R}^p$ and $V_{\theta_*}^{-1}\Delta_{n,\theta_*} \xrightarrow{d} \mathcal{N}(0, V_{\theta_*}^{-1})$. As the ETEL function is an integrated likelihood, expansion (15) is better known as integral LAN in the semiparametric Bayesian literature, see for example Bickel and Kleijn (2012 section 4). In the third step of the proof we use arguments as in the proof of Van der Vaart (1998 Theorem 10.1) to show that (14) and (15) imply asymptotic normality of $\pi(\sqrt{n}(\theta - \theta_*) \in B|W_{1:n}, K)$. While these three steps are classical in proving the Bernstein–von Mises phenomenon, establishing (15) raises challenges that are otherwise absent. This is because the ETEL function is a nonstandard likelihood that involves estimated parameters $\hat{\lambda}(\theta_*)$ whose dimension is dK, which increases with n. While $\|\hat{\lambda}(\theta_*)\|$ and $\|\frac{1}{n}\sum_{i=1}^n g(W_i, \theta_*)\|$ are expected to converge to zero in the correctly specified case, the rate of convergence is slower than $n^{-1/2}$. In the supplementary appendix we show that this rate is $\sqrt{K/n}$ under the previous assumptions.

3.4 | Misspecified model

We now generalize the preceding BvM result for the important class of misspecified conditional moment models.

Definition 3.1 (Misspecified model). We say that the conditional moment conditions model $\mathbf{E}^P[\rho(X, \theta)|Z] = 0$ is misspecified if the set of probability measures implied by the moment restrictions does not contain the true data generating process P_* for any $\theta \in \Theta$, that is, $P_* \notin \mathcal{P}$ where $\mathcal{P} := \bigcup_{\theta\in\Theta} \tilde{\mathcal{P}}_\theta$ and $\tilde{\mathcal{P}}_\theta = \{Q \in \mathbb{M}_{X|Z}; \mathbf{E}^Q[\rho(X, \theta)|Z] = 0 \ a.s.\}$ with $\mathbb{M}_{X|Z}$ the set of all conditional probability measures of $X|Z$.

In essence, if (3) is misspecified then there is no $\theta \in \Theta$ such that $\mathbf{E}^P[\rho(X, \theta) \otimes q^K(Z)] = 0$ almost surely for every K large enough. Now, for every $\theta \in \Theta$ define $Q^*(\theta)$ as the minimizer of the Kullback–Leibler divergence of P_* to the model $\mathcal{P}_\theta := \{Q \in \mathbb{M}; \mathbf{E}^Q[g(W, \theta)] = 0\}$, where $\mathbb{M}$ denotes the set of all the probability measures on $\mathbb{R}^{d_w}$. That is, $Q^*(\theta) := \text{arginf}_{Q\in\mathcal{P}_\theta} \mathbb{K}(Q\|P_*)$, where $\mathbb{K}(Q\|P_*) := \int \log(dQ/dP_*)dQ$. If we suppose that the dual representation of the Kullback–Leibler minimization problem holds, then the P_*-density of $Q^*(\theta)$ has the closed form: $[dQ^*(\theta)/dP_*](W_i) = \frac{e^{\lambda_\circ' g(W_i,\theta)}}{\mathbf{E}^P[e^{\lambda_\circ' g(W_j,\theta)}]}$, where $\lambda_\circ$ denotes the tilting parameter and is defined in the same way as in the correctly specified case:

$$\lambda_\circ := \lambda_\circ(\theta) := \arg\min_{\lambda\in\mathbb{R}^{dK}} \mathbf{E}^P[e^{\lambda' g(W_i,\theta)}]. \tag{16}$$

We also impose a condition to ensure that the probability measures $\mathcal{P} := \bigcup_{\theta\in\theta} \mathcal{P}_\theta$, which are implied by the model, are dominated by the true probability measure P_*. This is required for the validity of the dual theorem. Therefore, following Sueishi (2013 Theorem 3.1), we replace Assumption 3.3 (a) by the following.

Assumption 3.8 For every $\theta \in \Theta$, there exists $Q \in \mathcal{P}_\theta$ such that Q is mutually absolutely continuous with respect to P_*, where $\mathcal{P}_\theta := \{Q \in \mathbb{M}; \mathbf{E}^Q[g(W, \theta)] = 0\}$ and $\mathbb{M}$ denotes the set of all the probability measures on $\mathbb{R}^{d_w}$.

This assumption implies that $\mathcal{P}_\theta$ is non-empty. A similar assumption is also made by Kleijn and van der Vaart (2012) and Chib et al. (2018) to establish the BvM under misspecification. The

pseudo-true value of the parameter $\theta \in \Theta$ is denoted by $\theta_\circ$ and is defined as the minimizer of the Kullback–Leibler divergence between the true P_* and $Q^*(\theta)$:

$$\theta_\circ := \text{arginf}_{\theta\in\Theta}\mathbb{K}(P_*\|Q^*(\theta)), \tag{17}$$

where $\mathbb{K}(P_*\|Q^*(\theta)) := \int \log(dP_*/dQ^*(\theta))dP_*$. Under the preceding absolute continuity assumption, the pseudo-true value $\theta_\circ$ is available as

$$\theta_\circ = \text{argmax}_{\theta\in\Theta}\mathbf{E}^P \log\left(\frac{e^{\lambda_\circ' g(W_i,\theta)}}{\mathbf{E}^P[e^{\lambda_\circ' g(W_j,\theta)}]}\right). \tag{18}$$

Note that $\lambda_\circ(\theta_\circ)$, the value of the tilting parameter at the pseudo-true value $\theta_\circ$, is nonzero because the moment conditions do not hold.

Assumption 3.8 implies that $\mathbb{K}(Q^*(\theta_\circ)\|P_*) < \infty$. We supplement this with the assumption that $\mathbb{K}(P_*\|Q^*(\theta)) < \infty$, $\forall\ \theta \in \Theta$ (so that $\mathbb{K}(P_*\|Q^*(\theta_\circ)) < \infty$). Because consistency in misspecified models is defined with respect to the pseudo-true value $\theta_\circ$, we need to replace Assumption 3.7 (b) by the following Assumption 3.9 (b) which, together with Assumption 3.9 (a), requires the prior to put enough mass to balls around $\theta_\circ$.

Assumption 3.9 (a) π is a continuous probability measure that admits a density with respect to the Lebesgue measure; (b) The prior distribution π is positive on a neighbourhood of $\theta_\circ$, where $\theta_\circ$ is as defined in Equation (18).

Hereafter, we use the sub/super index $Q^*(\theta_\circ)$ to denote an expectation, a variance or covariance taken with respect to the probability $Q^*(\theta_\circ)$. The following assumption is analogous to the second part of Assumption 3.2 for the P_*-density of $Q_*(\theta)$ replacing P.

Assumption 3.10 For each K the matrix $\mathbf{E}^{Q^*(\theta_\circ)}[q^K(Z)q^K(Z)']$ has smallest eigenvalue bounded away from zero uniformly in K.

In the next assumption we denote by $int(\Theta)$ the interior of Θ and by $\mathcal{U}$ a ball centred at $\theta_\circ$ with radius $h/\sqrt{n}$ for some $h \in \mathcal{H}$ and $\mathcal{H}$ a compact subset of $\mathbb{R}^p$.

Assumption 3.11

(a) The data $W_i := (X_i, Z_i)$, $i = 1, \ldots, n$ are i.i.d. according to P_* and
(b) The pseudo-true value $\theta_\circ \in int(\Theta)$ is the unique maximizer of

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

where $\Theta \subset \mathbb{R}^p$;
(c) $\rho(X,\theta)$ is continuous at each $\theta \in \Theta$ with probability one;
(d) $\rho(X,\theta)$ is twice continuously differentiable in a neighbourhood $\mathcal{U}$ of $\theta_\circ$ with probability one and for $\kappa = 0, 1$, $\mathbf{E}^P\left[\sup_{\theta\in\mathcal{U}}\left|\frac{dQ^*(\theta)}{dP_*}(W)\right|^\kappa \|\rho_{j\theta\theta}(X,\theta)\|^2\|q^K(Z)\|^2\right] = \mathcal{O}(K)$, $j = 1, \ldots, d$;
(e) for a neighbourhood $\mathcal{U}$ of $\theta_\circ$ and for $\kappa = 0, 1, 2, j = 2, 4$ it holds that

$$\mathbf{E}^P\left[\sup_{\theta\in\mathcal{U}}\left|\frac{dQ^*(\theta)}{dP_*}(W_i)\right|^\kappa \|\rho(X,\theta)\|^j\|q^K(Z)\|^j\right] = \mathcal{O}(\zeta(K)^{j-2}K),$$

where $\zeta(K)$ is as defined in Assumption 3.2;

(f) for a neighbourhood $\mathcal{U}$ of $\theta_\circ$ and for $\kappa = 0, 1, 2, j = 1, 2, 4$ it holds that

$$\mathbf{E}^P\left[\sup_{\theta\in\mathcal{U}}\left|\frac{dQ^*(\theta)}{dP_*}(W_i)\right|^\kappa \|\rho_\theta(X_i,\theta)\|^j \|q^K(Z)\|^j\right] = \mathcal{O}(\zeta(K)^{\max\{j-2,0\}}K),$$

where $\zeta(K)$ is as defined in Assumption 3.2;

(g) the matrix $\mathbf{E}^{Q^*(\theta_\circ)}[\rho(X,\theta_\circ)\rho(X,\theta_\circ)'|Z]$ has smallest (resp. largest) eigenvalue bounded away from zero (resp. infinity);

(h) for a neighbourhood $\mathcal{U}$ of $\theta_\circ$ it holds that $\mathbf{E}^P[\sup_{\theta\in\mathcal{U}}|\frac{dQ^*(\theta)}{dP_*}(W_i)|^2]$ is bounded.

Assumption 3.11 (b) guarantees uniqueness of the pseudo-true value and is a standard assumption in the literature on misspecified models (see e.g. White, 1982). Assumptions 3.11 (d)–(f) are the counterparts of Assumptions 3.4 (b) and 3.4 (b), respectively, for the misspecified case. It is important to notice that they implicitly contain the first part of Assumption 3.2. The reason why we cannot separate the part involving the moment function $\rho(X,\theta)$ (or its derivative) and the one involving $q^K(Z)$ in the assumption, as we do for the correctly specified model, is that the P_*-density of $Q_*(\theta)$ cannot be factorized in a conditional density of X given (Z,θ) and a marginal density of Z independent of θ. In particular, in the misspecified case the pseudo-true value of the tilting parameter $\lambda_\circ(\theta_\circ)$ is not equal to zero as it is the tilting parameter in the correctly specified case. Assumption 3.11 (g) is the counterpart of Assumption 3.5 (b) for the misspecified case.

The BvM theorem for misspecified models now follows. Let $G_i(\theta_\circ) := \rho_\theta(X_i,\theta_\circ)\otimes q^K(Z_i)$, $D_\circ^\dagger(Z) := \mathbf{E}^{Q^*(\theta_\circ)}[\rho_\theta(X,\theta_\circ)|Z]$ and $\Sigma_\circ(Z) := \mathbf{E}^{Q^*(\theta_\circ)}[\rho(X,\theta_\circ)\rho(X,\theta_\circ)'|Z]$. Moreover, let $\mathcal{H}$ denote a compact subset of $\mathbb{R}^p$ and $\theta_h := \theta_\circ + h/\sqrt{n}$, with $h\in\mathcal{H}$.

Theorem 3.2 (Bernstein-von Mises (misspecified)). *Let Assumptions* 3.1, 3.2, 3.8–3.11 *hold. Assume that there exists a constant $C>0$ such that for any sequence $M_n\to\infty$,*

$$P_*\left(\sup_{\|\theta-\theta_\circ\|>M_n/\sqrt{n}}\frac{1}{n}\sum_{i=1}^n\left(\ell_{n,\theta}(W_i)-\ell_{n,\theta_\circ}(W_i)\right)\le -CM_n^2/n\right)\to 1, \tag{19}$$

as $n\to\infty$. If $K\to\infty$ and $\sup_{\theta\in\mathcal{U}}\|\lambda_\circ(\theta)\|^2\max\{\zeta(K),K\}K\sqrt{K/n}\to 0$ then, the posteriors converge in total variation towards a normal distribution, that is,

$$\sup_B\left|\pi(\sqrt{n}(\theta-\theta_\circ)\in B|W_{1:n},K)-\mathcal{N}_{\Delta_{n,\theta_\circ},V_{\theta_\circ}}(B)\right|\xrightarrow{p}0, \tag{20}$$

where $B\subseteq\Theta$ is any Borel set, $\Delta_{n,\theta_\circ}$ is a random vector bounded in probability and $V_{\theta_\circ}$ is a positive definite matrix equal to the inverse of:

$$\begin{aligned}
V_{\theta_\circ}^{-1} &= \mathbf{E}^P\left[D_\circ^\dagger(Z)\Sigma_\circ(Z)^{-1}D_\circ(Z)\right] + \mathbf{E}^{Q^*(\theta_\circ)}[G_i(\theta_\circ)'\lambda_\circ(\theta_\circ)g(W_i,\theta_\circ)']\Omega_\circ^{-1}\mathbf{E}^P[G_i(\theta_\circ)]\\
&\quad - \frac{d\lambda_\circ(\theta_\circ)'}{d\theta}\left(\mathbf{E}^P[G_i(\theta_\circ)]-\mathbf{E}^{Q^*(\theta_\circ)}[G_i(\theta_\circ)]\right) - \sum_{j=1}^d\frac{d^2\lambda_{\circ,j}(\theta_\circ)'}{d\theta d\theta'}\mathbf{E}^P[\rho_j(X_i,\theta_\circ)q^K(Z)]\\
&\quad - \sum_{j=1}^d\left(\mathbf{E}^P\left[\rho_{j\theta\theta}(X_i,\theta_\circ)q^K(Z_i)\right]-\mathbf{E}^{Q^*(\theta_\circ)}\left[\rho_{j\theta\theta}(X_i,\theta_\circ)q^K(Z_i)\right]\right)\lambda_{\circ,j}(\theta_\circ)\\
&\quad + \mathrm{Var}_{Q^*(\theta_\circ)}[G_i(\theta_\circ)'\lambda_\circ(\theta_\circ)] + \frac{d\lambda_\circ(\theta_\circ)'}{d\theta}\mathrm{Cov}_{Q^*(\theta_\circ)}\left(g(W_i,\theta_\circ),G_i(\theta_\circ)'\lambda_\circ(\theta_\circ)\right).
\end{aligned}$$

Just as in Kleijn and van der Vaart (2012), this theorem establishes that the posterior distribution of the centred and scaled parameter $\sqrt{n}(\theta - \theta_\circ)$ converges to a Normal distribution with a random mean that is bounded in probability. The rate restriction $\sup_{\theta \in \mathcal{U}} \|\lambda_\circ(\theta)\|^2 \max\{\zeta(K), K\} K\sqrt{K/n} \to 0$ is much stronger than the one in Theorem 3.1. The slower rate condition on K is intuitive. When the conditional moment conditions are misspecified, limiting the number of implied unconditional moments serves to limit the magnification of the misspecification. In the (completely) hypothetical situation where one knew that the conditional moment conditions are misspecified, one would either discard the misspecified moment conditions or take a small and fixed number of expanded moment conditions (for instance with $K = 1$). In practice, of course, this strategy cannot be implemented (because one does not know whether the model is correctly specified or misspecified) and, therefore, K must always go to ∞, but slower than under correct specification.

An additional remark is that, because of misspecification, the covariance matrix $V_{\theta_\circ}$ appearing in Theorem 3.2 is expected to be different from the asymptotic covariance matrix of the corresponding frequentist point estimator in Ai and Chen (2007 Theorem 4.1). Therefore, while correctly centred, the $(1 - \alpha)$ posterior credible sets are not in general $(1 - \alpha)$ confidence sets. The coverage rate can be less, or more, than the nominal level depending on the true data generating process and the extent of misspecification.

The strategy of the proof of this result is generally similar to the proof of Theorem 3.1 with θ_* replaced by the pseudo-true value $\theta_\circ$. However, proving that the ETEL function satisfies a stochastic LAN expansion is more complex, for the following reasons. First, the limit of $\hat{\lambda}(\theta_\circ)$ is $\lambda_\circ(\theta_\circ)$, which is not zero. Therefore, several terms that were equal to zero in the LAN expansion under correct specification, are non-zero in the misspecified case. The limit in distribution of these terms has to be derived. This explains our stronger assumptions with respect to the correctly specified case. Second, the quantity $\frac{1}{\sqrt{n}}\sum_{i=1}^{n} g(W_i, \theta_\circ)$ is no longer centred on zero, which leads to an additional bias term.

Example 1 (continued). Consider now the impact of K in misspecified models. For simplicity, we fix θ_0 to a certain value and we only estimate θ_1. We generate the sample data as above with $(\theta_0 = 1, \theta_1 = 1)$ and then estimate θ_1, fixing θ_0 at 0.5. Therefore, the moment condition $\mathbf{E}^P[\varepsilon|Z] = 0$ is misspecified. The pseudo-true value of θ_1 is obtained from a side calculation. More specifically, we generate 5 million observations from the true model. We assume that this sample size is large enough to represent the $n \to \infty$ situation. Then, we misspecify the CM conditions by setting $\theta_0 = 0.5$, construct the expanded moments with these 5 millions observations setting $K = 26$, and numerically find the value of θ_1 that maximizes the BETEL posterior distribution. The pseudo-true value is this maximized value. It is 1.004. The adverse impact of increasing K (for a fixed n) on the Bayesian bias and posterior standard deviation is reported in Table 2. As in the case of the correctly specified models, relatively small values of K (around $K = 5$ for $n = 250$) lead to the best results and, in addition, the value of the Bayesian bias increase when K increases beyond the recommended value. The posterior sd declines more sharply as K increases. As pointed out before, these effects are due to the reduction in the support of the prior, now magnified by moment misspecification. Finally, we calculate frequentist coverage rate of the equal-tailed 90% credible set based on 100 repetitions for both the correctly specified model (θ_0 now set equal to 1) and misspecified model with $K \approx 2n^{1/6}$. Consistent with our theory the BETEL credible set is different from the frequentist confidence

TABLE 2 Bayesian bias and posterior SD for different values of K under correct and incorrect conditional moments

	Correctly specified model		Misspecified model	
$n = 250$	|Bias|	SD	|Bias|	SD
$K = 2$	0.277	0.147	0.273	0.154
$K = 5$	0.048	0.110	0.044	0.102
$K = 9$	0.053	0.105	0.049	0.098
$K = 12$	0.059	0.105	0.055	0.097
$K = 20$	0.064	0.102	0.060	0.084

set when conditional moment conditions are misspecified. When $n = 250$ the coverage rates are 91% for the correctly specified case and 86% for the misspecified case. When the number of observations increases to 1,000, the coverage rate for the misspecified case further moves down to 80% while the coverage rate for the correctly specified case remains around 90%.

4 | MODEL COMPARISONS

In practice, we can be unsure about elements of the conditional moment model. For instance, we can be faced with a large number of variables in Z, but only some of which are relevant. In such cases, any specific model may be considered to be misspecified, and the goal is to find the best model given the data.

Let M_ℓ denote the ℓth model in the model space. Each model is characterized by a parameter θ^ℓ and an extended set of moment functions given by $g^\ell(W, \theta^\ell)$. In addition, each model M_ℓ is described by a prior distribution for θ^ℓ. The posterior distribution is obtained based on (11). The aim is to compare these models by marginal likelihoods, denoted by $m(W_{1:n}|M_\ell, K)$. These are each calculated by the marginal likelihood identity of Chib (1995) (where we explicit the dependence on M_ℓ in the notation):

$$\log m(W_{1:n}|M_\ell, K) = \log \pi(\tilde{\theta}^\ell|M_\ell) + \log p(W_{1:n}|\tilde{\theta}^\ell, M_\ell, K) - \log \pi(\tilde{\theta}^\ell|W_{1:n}, M_\ell, K), \quad (21)$$

and by the method of Chib and Jeliazkov (2001). In this expression, $\tilde{\theta}^\ell$ is any point in the support of the posterior (such as the posterior mean).

Remark 4.1 Comparison of CM condition models differs in one important aspect from the framework for comparing unconditional moment condition models that was established in Chib et al. (2018), where it is shown that to make the unconditional moment condition models comparable it is necessary to linearly transform the moment functions so that all the transformed moments are included in each model. This linear transformation consists of adding an extra parameter different from zero to the components of the vector $g(\theta, W)$ that correspond to the restrictions not included in a specific model. When comparing conditional moment models, however, this transformation is not necessary because the convex hulls associated with different expanded models have the same dimension asymptotically.

4.1 | Model selection consistency

Suppose that there are J contending models. Suppose also that at least $J-1$ of these models are misspecified and the remaining one can be either misspecified or correctly specified. Moreover, suppose that a model M_ℓ is selected by the size of the marginal likelihoods. Then, in Theorem 4.1 we show that this criterion in the limit picks the model M_ℓ with the smallest Kullback–Leibler divergence between P_* and the corresponding $Q^*(\theta^\ell)$, where $Q^*(\theta^\ell) = \text{arginf}_{Q\in\mathcal{P}_{\theta^\ell}} \mathbb{K}(Q||P_*)$ and $\mathcal{P}_{\theta^\ell}$ is defined in Section 3.4.

Theorem 4.1 *Let the assumptions of Theorem* 3.2 *hold. Let us consider the comparison of $J<\infty$ models M_ℓ, $\ell = 1, \ldots, J$, such that $J-1$ of these models each has at least one misspecified moment condition and model M_j can be either correctly specified or contain some misspecified moment condition, that is, M_ℓ does not satisfy Assumption* 3.3 (a), $\forall\, \ell \neq j$. *Then,*

$$\lim_{n\to\infty} P_*\left(\log m(W_{1:n}|M_j, K) > \max_{\ell\neq j} \log m(W_{1:n}|M_\ell.K)\right) = 1$$

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

Note that if one model in the contending set of models is correctly specified, then this model will have zero Kullback–Leibler divergence and, therefore, according to Theorem 4.1, that model will have the largest marginal likelihood and will be selected by our procedure.

To understand the ramifications of the preceding result, suppose that we are interested in comparing models with the same moment conditions but different conditioning variables:

$$\text{Model 1: } \mathbf{E}^P[\rho(X,\,\theta)|Z_1] = 0, \quad \text{Model 2: } \mathbf{E}^P[\rho(X,\,\theta)|Z_2] = 0, \tag{22}$$

where Z_1 and Z_2 may have some elements in common, in particular Z_2 might be a subvector of Z_1 (or vice versa). A situation of this type, where we are unsure about the validity of instrumental variables, is the following.

Example 2 (Comparing IV models). Consider the following model with three instruments $(Z_1,\, Z_2,\, Z_3)$:

$$\begin{aligned} Y &= \theta_0 + \theta_1 X + e_1, \\ X &= f(Z_1, Z_2, Z_3) + e_2, \\ Z_1 &\sim U[0,1], \quad Z_2 \sim U[0,1], \quad \text{and} \quad Z_3 \sim \mathcal{B}(0.4), \end{aligned}$$

where $(e_1, e_2)'$ are non-Gaussian and correlated. Thus, X in the outcome model is correlated with the error e_1. Let true $\theta = (\theta_0, \theta_1)$ equal (1, 1). Moreover, suppose that the Z_j's are relevant instruments, that is, $\text{cov}(X, Z_j) \neq 0$ for $j \leq 3$, and

$$f(Z_1,\, Z_2,\, Z_3) = 6(\sqrt{0.3}Z_1 + \sqrt{0.7}Z_2)^3(1 - \sqrt{0.3}Z_1 - \sqrt{0.7}Z_2)Z_3 + Z_1Z_2(1 - Z_3). \tag{23}$$

In practice, some instruments can be valid and some not, and the goal is to select the valid instruments. To this end, we generate $(e_1,\, e_2,\, Z_1)$ from a Gaussian copula whose covariance matrix is $\Sigma = [1, 0.7, 0.7; 0.7, 1, 0; 0.7, 0,1]$ such that the marginal distribution of e_1 is the skewed mixture of two normal distributions $0.5\mathcal{N}(0.5,\, 0.5^2) + 0.5\mathcal{N}(-0.5,\, 1.118^2)$

TABLE 3 Model comparison: IV regression example

	$\mathcal{M}_1$	$\mathcal{M}_2$	$\mathcal{M}_3$
$n = 100$	0%	52%	48%
$n = 250$	0%	40%	60%
$n = 1000$	0%	2%	98%

Notes: Each entry in the table presents the model selection frequency in 100 repetitions; ($n = 100$, $K = 4$), ($n = 250$, $K = 5$), and ($n = 1000$, $K = 6$), where K is based on $K = 2n^{1/6}$. Each result from 10,000 Markov chain Monte Carlo (MCMC) draws beyond a burn-in of 1000.

and the marginal distribution of e_2 is $\mathcal{N}(0,\ 1)$. Under this setup, Z_1 is an invalid instrument. Consider the following three models

$$\mathcal{M}_1 \ : \ \mathbf{E}^P[(Y - \theta_0 - \theta_1 X)|Z_1,\ Z_2, Z_3] = 0, \tag{24}$$

$$\mathcal{M}_2 \ : \ \mathbf{E}^P[(Y - \theta_0 - \theta_1 X)|Z_1,\ Z_3] = 0, \tag{25}$$

$$\mathcal{M}_3 \ : \ \mathbf{E}^P[(Y - \theta_0 - \theta_1 X)|Z_2,\ Z_3] = 0. \tag{26}$$

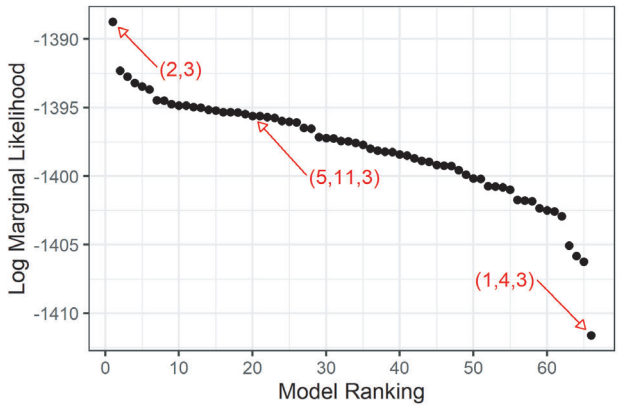
Because Z_1 is an invalid instrument, models $\mathcal{M}_1$ and $\mathcal{M}_2$ are misspecified.

In $\mathcal{M}_1$, the basis matrix B is made from the variables ($\boldsymbol{z}_1$, $\boldsymbol{z}_2$, $\boldsymbol{z}_1 \odot \boldsymbol{z}_2$, $\boldsymbol{z}_1 \odot \boldsymbol{z}_3$, $\boldsymbol{z}_2 \odot \boldsymbol{z}_3$), each using K knots, concatenated with the vector $\boldsymbol{z}_3$. In $\mathcal{M}_2$, B is made from the variables ($\boldsymbol{z}_1$, $\boldsymbol{z}_1 \odot \boldsymbol{z}_3$), each using K knots, concatenated with the vector $\boldsymbol{z}_3$. In $\mathcal{M}_3$, B is made from the variables ($\boldsymbol{z}_2$, $\boldsymbol{z}_2 \odot \boldsymbol{z}_3$), each using K knots, concatenated with the vector $\boldsymbol{z}_3$. The number of columns in the B matrix is $5(K - 1) + 2$ for $\mathcal{M}_1$, and $2(K - 1) + 2$ for $\mathcal{M}_2$ and $\mathcal{M}_3$. The prior for θ_0 and θ_1 is the product of student-t distributions with mean zero, dispersion 5, and degrees of freedom equal to 2.5. A repeated sampling experiment is conducted. The marginal likelihood of each model is calculated in 200 repeated samples. Table 3 reports the model selection frequency for ($n = 100$, $K = 4$), ($n = 250$, $K = 5$), and ($n = 1000$, $K = 6$), where K is based on $K = 2n^{1/6}$. Note that the model with the valid instruments, that is, $\mathcal{M}_3$, is selected more frequently as the number of observation gets larger, in conformity with the theory.

5 | ADDITIONAL TOPICS

5.1 | High-dimensional Z

We now consider the case where Z lies in a high-dimensional space. If all the elements of Z are relevant, then the situation can be challenging, but there is an interesting sub-case that is worth discussing. Suppose that the conditional expectation depends only on a few elements of Z or, in other words, most of the elements of Z are redundant. In this case, one can find the relevant elements of Z by estimating and comparing models that condition on different subsets of Z, where the cardinality of these subsets is say 2 or 3. The relevant elements of Z correspond to the model with the largest marginal likelihood. We refer to this procedure as sparsity-based model selection. The next example provides an illustration.



Ranking	Model	log(ML)	Prob
1	(2,3)	-1388.73	0.903
2	(4,3)	-1392.31	0.025
3	(5,3)	-1392.75	0.016
4	(7,3)	-1393.20	0.010
5	(5,9,3)	-1393.46	0.008

FIGURE 2 Left figure presents log marginal likelihood for each of 66 models in the model space. Arrows point to the best model (Z_2, Z_3), top 20 model (Z_6, Z_{12}, Z_3), and the worst model (Z_1, Z_4, Z_3). Right table presents log marginal likelihood and posterior model probability for top 5 models. Posterior model probabilities are computed with a uniform prior on model space

Example 3 (Sparsity-based model selection). Recall our Example 2, but assume that one has nine additional potential Z's

$$Z_j = \frac{9}{10}Z_1 + \frac{1}{10}\eta_j, \quad \eta_j \sim \text{Unif}([0, 1])$$

for $j = 4, 5, \ldots, 12$. Recall, Z_1 is an invalid instrument. Therefore, Z_j's for $j = 4, \ldots, 12$ are also invalid. Suppose that Z_3 affects the conditional expectation, but that one is unsure about the remaining elements of Z. Suppose one believes that at most three elements of Z affect the conditional expectation (the sparsity assumption). In this situation one can compute marginal likelihoods of the following 66 models:

$$\mathbf{E}^P[(Y - \theta_0 - \theta_1 X)|Z_j, Z_3] = 0, \quad j \in \{1, 2, 4, 5, \ldots, 12\}, \tag{27}$$

and

$$\mathbf{E}^P[(Y - \theta_0 - \theta_1 X)|Z_j, Z_k, Z_3] = 0, \quad j, k \in \{1, 2, 4, 5, \ldots, 12\} \text{ and } k \neq j\,, \tag{28}$$

with the correct model given by

$$\mathbf{E}^P[(Y - \theta_0 - \theta_1 X)|Z_2, Z_3] = 0 \tag{29}$$

Sample data (size $n = 250$) are generated from the design in Example 2. Estimation and marginal likelihood computations are based on expanded moments from $K = 3$ basis functions for each conditioning Z_j. A summary of the marginal likelihood results appears in Figure 2, sorted by the size of the marginal likelihood. The top ranked model is the true model. As shown in the right panel of the same figure, which reports the posterior model probabilities (under a uniform prior on model space), the support for the true model is decisive.

5.2 | High-dimensional θ: TaRB-MH

It is also important to consider the estimation of conditional moment models that contain a high-dimensional θ. While the regularizing role of the Bayesian prior is important, MCMC sampling of the posterior simulation becomes more complicated. From our experience, the one-block M-H algorithm that we have used above tends to be inefficient. A straightforward alternative sampling scheme is offered by the tailored randomized block MH algorithm of Chib and Ramamurthy (2010). This algorithm, which has proved useful in several similarly complex settings, trades more computations for gains in simulation efficiency.

Example 3 (continued). (IV regression with additional exogenous regressors). Consider the previous IV regression model, but now with 18 additional exogenous regressors w

$$y_i = \theta_0 + \theta_1 x_i + w_i'\gamma + e_{1,i}$$

where $w_i = [w_i^{(1)'}, w_i^{(2)'}, w_i^{(3)'}]'$, and each group $w_i^{(j)'}$, $j \leq 3$, are identically and independently drawn from $\mathcal{N}_6(0, \Sigma(\rho))$, where $\Sigma(\rho)$ is a 6×6 matrix in correlation form with each off-diagonal element set equal to 0.97. In addition, γ is a vector of ones. In total, there are 20 unknown parameters. Other elements of the DGP are unchanged. Suppose one has 1500 observations from this DGP, and we estimate $[\theta_0, \theta_1, \gamma']'$ from $\mathbf{E}^P[(Y - \theta_0 - \theta_1 X)|Z_2, Z_3, W] = 0$ with the expanded moment conditions similar to $\mathcal{M}_3$ in Example 1. The basis function matrix is formed with Z_2, Z_3, and $Z_2 Z_3$, concatenated with columns in W. We set $K = 6$, following our recommendation, which leads to 36 expanded moment conditions. The training sample prior is based on the first 10% of the sample, and estimation on the remaining 90%. The prior is a product of independent student-t distributions with 5 degrees of freedom, centred on the two-stage least squares (2SLS) estimate, and scale equal to two times the 2SLS standard error.

The results appear in Table 4. In implementing the TaRB-MH sampling scheme, the probability of starting a new block is set to 0.3, so that the each block, within each MCMC iteration, contains six parameters on average. For comparison, results from the single-block sampling scheme (on the same conditional moments) are also included. It is evident that the two MCMC samplers produce identical posterior moments, but that the TaRB-MH sampler dominates the one-block MH sampler in terms of simulation efficiency as measured by 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 are essentially independent. Therefore, armed with the TaRB-MH sampler, computational efficiency is retained, even in higher dimensional θ problems.

6 | APPLICATIONS

6.1 | Asset pricing

A key question in finance concerns the makeup of the pricing kernel, or the stochastic discount factor (SDF). Factors in the SDF are the risk factors that explain the cross-section of expected

TABLE 4 Posterior summary of IV regression example with additional covariates ($n = 1500$)

	TaRB-MH			One-block-MH		
	Mean	**SD**	**Ineff**	**Mean**	**SD**	**Ineff**
θ_0	1.03	0.03	4.35	1.03	0.03	14.34
θ_1	0.91	0.11	7.57	0.91	0.10	47.38
γ_1	0.99	0.03	6.92	0.99	0.03	13.83
γ_2	0.93	0.15	6.06	0.92	0.15	15.09
γ_3	0.96	0.15	2.75	0.97	0.15	12.41
γ_4	1.14	0.17	2.21	1.15	0.17	12.54
γ_5	1.32	0.16	1.97	1.33	0.16	14.26
γ_6	1.02	0.16	1.57	1.02	0.15	12.79
γ_7	0.99	0.03	7.05	0.99	0.02	10.58
γ_8	1.04	0.14	5.42	1.04	0.14	14.60
γ_9	1.18	0.15	2.98	1.18	0.14	11.73
γ_{10}	1.22	0.16	2.37	1.22	0.15	14.66
γ_{11}	0.94	0.15	1.81	0.94	0.15	11.56
γ_{12}	0.64	0.16	1.69	0.64	0.16	24.87
γ_{13}	1.01	0.03	7.00	1.01	0.03	12.89
γ_{14}	0.86	0.15	5.58	0.86	0.15	14.66
γ_{15}	0.84	0.15	3.34	0.84	0.15	14.86
γ_{16}	1.04	0.16	2.24	1.05	0.15	12.84
γ_{17}	1.18	0.17	2.11	1.18	0.17	20.25
γ_{18}	0.86	0.16	1.44	0.86	0.16	16.31

Notes: The true value of all parameters (θ's and γ's) are set to one. The summaries are based on 50,000 Markov chain Monte Carlo (MCMC) draws beyond a burn-in of 10,000 for the one-block-MH sampler and 3,000 draws beyond a burn-in of 1,000 for the TaRB-MH. The M-H acceptance rate is around 37% for the one-block-MH and 87% for TaRB-MH. 'Ineff' is the inefficiency factor.

equity returns and, for this reason, establishing the identity of these risk factors has been a long-standing quest in finance.

Following notation from Chib and Zeng (2020), write the SDF M_t at time (month) t as

$$M_t = 1 - b'(x_t - \mu_x) \tag{30}$$

where x_t is a $(k_x \times 1)$ vector of risk factors (empirically these are the excess returns on portfolios of stocks), and b is the unknown risk-factor premia and $\mu_x = \mathbf{E}^P(x_t)$. The parameters $(b,\ \mu_x)$ are unknown. Suppose that there are other factors (excess returns on other portfolios) that are collected in a $(k_w \times 1)$ vector w_t. Let $f_t := (x_t', w_t')'$ be a $(k_f \times 1)$-vector, where $k_f = k_x + k_w$. If x_t are risk factors, then finance theory dictates that the restriction $\mathbf{E}^P(M_t f_t) = 0$ holds. Given a sample of observations $\{f_t\}_{t=1}^n$, one can estimate $(b,\ \mu_x)$ based on the following moment conditions

$$\mathbf{E}^P[(1 - b'(x_t - \mu_x))f_t] = 0, \quad \mathbf{E}^P[x_t - \mu_x | f_{t-1}] = 0,$$

TABLE 5 Asset pricing data: Summary of the posterior distribution based on 50,000 Markov chain Monte Carlo (MCMC) draws after 1000 burn-in

	Mean	SD	Median	5%	95%
b	2.981	0.730	2.955	1.818	4.211
μ_x	0.006	0.001	0.006	0.004	0.008

where the second conditional moment restriction identifies μ_x.

As an example, consider the data at http://apps.olin.wustl.edu/faculty/chib/rpackages/czfactor/czfactor.pdf on monthly excess returns (January 1974–December 2018) on $k_f = 12$ potential risk factors. Thus, in this situation, there are 12 conditioning variables, an illustration of a modestly high-dimensional Z. Let x_t be the excess return on the market portfolio (denoted Mkt in the data).

Now construct the expanded moment conditions as

$$\mathbf{E}^P \left[(x_t - \mu_x) \otimes [q^K(f_{1,t-1}), \tilde{q}^K(f_{2,t-1}), \ \dots \ , \tilde{q}^K(f_{12,t-1})] \right] = 0, \tag{31}$$

where $q^K(f_{1,t-1})$ consist of $K = 3$ basis functions, and $\tilde{q}^K(f_{j,t-1})$ $(j \geq 2)$ each consist of two basis functions derived from $q^K(f_{j,t-1})$ by subtracting the second and third columns from the first and then dropping the first. Along with these 25 expanded moment conditions, the pricing conditions supply an additional twelve, for a total of 37 moment conditions.

For the prior, one can employ the training sample approach. From the first 80 observations (the training sample) the hyperparameters of the independent student-t distribution of (b, μ_x) with 2.5 degrees of freedom are set as follows. The centre of the prior density is set to the generalized method of moment (GMM) estimate, and the scale to two times the GMM standard error. This black-box prior is particularly convenient if the analysis has to be repeated for different possible variables in the SDF. The remaining 459 observations are used to construct a joint posterior distribution of b and μ_x.

The posterior summary of (b, μ_x) from 50,000 MCMC draws is given in Table 5, and the prior and posterior density of b are presented in Figure 3. This summary, specifically the lower and upper limits of the marginal posterior of b confirm that b is non-zero, and hence that the Mkt variable is a risk factor.

6.2 | ATE under conditional ignorability

For another important application of the methods in this paper, consider the problem of estimating the average treatment effect (ATE) under the assumption of conditional ignorability. In the frequentist literature, the ATE is commonly estimated by propensity score methods and, on the Bayesian side, from models of the potential outcomes. These models generally have a non-parametric mean function, but parametric noise. By adopting the conditional moment perspective, however, one can evade the burden of distributional assumptions.

Data are from the 1997 Child Development Supplement to the Panel Study of Income Dynamics Guo and Fraser (2015 section 5.8.2), where the ATE is calculated by the propensity score. The research question is the effect of childhood welfare dependency on academic achievement. The

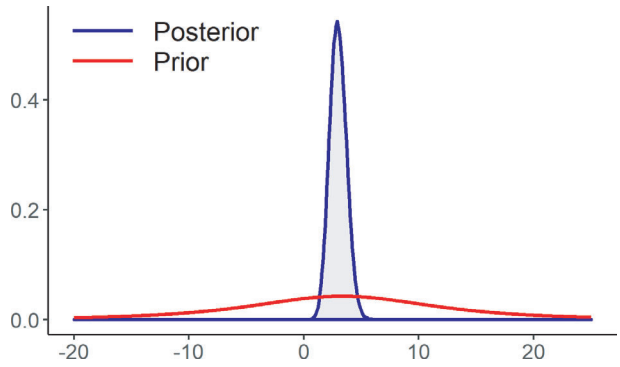


FIGURE 3 Asset pricing data: Prior and posterior density of b. Prior is Student-t density based on the first 80 observation. Posterior density is based on the remaining 459 observations. 50,000 Markov chain Monte Carlo draws after 1000 burn-in

latter, the dependent variable y, is measured by the child's score on the 'letter-word identification' section of the Woodcock-Johnson Revised Tests of Achievement. The treatment variable x equals one if the child received AFDC (Aid to Families with Dependent Children) benefits at any time from birth to 1997 (the survey year) and equals zero if the child never received benefits during that period. It is assumed that the potential outcomes are independent of x, conditioned on z_1, z_2, … , z_6 (the assumption of conditional ignorability), where

- z_1: mratio97, the ratio of family income to the poverty line in 1997
- z_2: pcged97, the caregiver's years of schooling
- z_3: pcg_adc, the number of years in which the caregiver received AFDC in her childhood
- z_4: age97, the child's age in 1997
- z_5: race, one for African-American children and zero for other
- z_6: male, one if the child is male and zero if female.

Two observations from the sample are dropped. These have values of mratio97 larger than 9 standard deviation from the mean of mratio97. Apart from mratio97 and age97, the other variables are categorical. There are $n_0 = 727$ control subjects and $n_1 = 274$ treated subjects. The ATE is expected to be negative, reflecting the hypothesis that welfare dependency has an adverse effect on academic achievement.

To answer the research question, suppose that the potential outcomes for the controls satisfy the conditional moments

$$\mathbf{E}^P((y_{i0} - \beta_{00} - h_{01}(z_1) - \beta_{02}z_2 - \beta_{03}z_3 - h_{04}(z_4) - \beta_{05}z_5 - \beta_{06}z_6)|z_i) = 0$$

and those for the treated satisfy the conditional moments

$$\mathbf{E}^P((y_{i1} - \beta_{10} - h_{11}(z_1) - \beta_{12}z_2 - \beta_{13}z_3 - h_{14}(z_4) - \beta_{15}z_5 - \beta_{16}z_6)|z_i) = 0$$

where $\{h_{01}, h_{04}, h_{11}, h_{14}\}$ are four non-parametric functions. These are each modelled by natural cubic splines with five knots. Thus, the parameters θ_j of the jth potential outcome model consist

of $(\beta_{j0}, \beta_{j2}, \beta_{j3}, \beta_{j5}, \beta_{j6})$ plus the eight spline coefficients. Special cases of this model, mentioned below, are considered and evaluated by marginal likelihoods. For example, models in which the h functions are linear are of interest.

The expanded moments are constructed as follows. The basis matrix has cubic spline basis functions for (z_1, z_4), each with five knots, concatenated with (z_2, z_3, z_5, z_6) (as is) because the latter variables are all essentially categorical. In total, this produces 13 expanded unconditional moments for the estimation of the y_0 and y_1 models. The prior distribution on the parameters is a product of student-t distributions with 2.5 degrees of freedom with mean of the intercept equal to the mean of the first 50 observations, the mean of the slopes equal to 0, and dispersion equal to 5.

Four models are estimated and evaluated. In the baseline model, the h functions are linear. In the second model, only the effect of z_1 is non-parametric. In the third model, only the effect of z_4 is assumed to be non-parametric and, finally, in the fourth model, both z_1 and z_4 are non-parametric. The results given in Table 6 show that the model best supported by these data is the third.

Consider now posterior inference on the ATE. By definition, the sample version of the ATE is

$$\text{ATE} = \frac{1}{n}\sum_{i=}^{n}\left(\mathbf{E}^P\left(y_{i1}|z_i, \theta_1\right) - \mathbf{E}^P\left(y_{i0}|z_i, \theta_0\right)\right),$$

where, in the model selected by the preceding comparison,

$$\mathbf{E}^P\left(y_{ij}|z_i, \theta_j\right) = \beta_{j0} + \beta_{j1}z_1 + \beta_{j2}z_2 + \beta_{j3}z_3 + h_{j4}(z_4) + \beta_{j5}z_5 + \beta_{j6}z_6.$$

Clearly, if we evaluate the latter expression at each posterior draw of (θ_0, θ_1), we produce a sample of the ATE from its posterior distribution. We summarize this sample in Table 7 and Figure 4. One can see that the ATE posterior point estimate is similar in size to

TABLE 6 Academic achievement data: Marginal likelihoods of four competing causal models, based on 20,000 Markov chain Monte Carlo (MCMC) draws beyond a burn-in 1000

	Non-treated	Treated
Linear	−4823.76	−1555.55
z_1 non-parametric	−4829.23	−1556.23
z_4 non-parametric	**−4813.86**	**−1555.78**
z_1 and z_4 non-parametric	−4818.98	−1556.56

TABLE 7 Academic achievement data: Summary of the posterior distribution of the ATE from model in which the effect of z_4 is non-parametric

	Mean	SD	Median	Lower	Upper
Propensity score matching	−5.682	1.976		−9.496	−1.502
Bayesian ATE	−5.257	1.393	−5.267	−7.983	−2.545

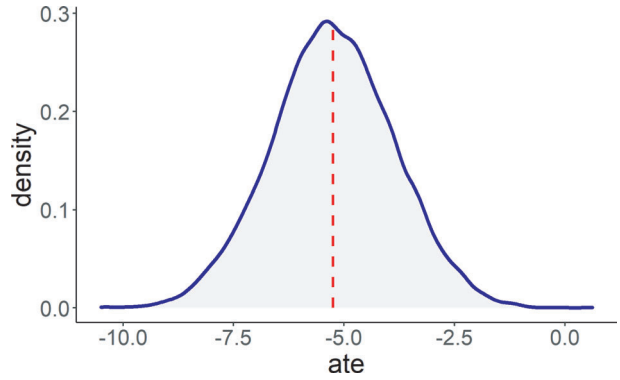


FIGURE 4 Academic achievement data: Posterior density of the average treatment effect from model in which the effect of z_4 is non-parametric. Red dashed line is at -5.251, the posterior mean. Posterior density based on 20,000 posterior draws after 1000 burn-in

the propensity score estimate, but the posterior standard deviation is smaller, leading to a less dispersed interval estimate. As a takeaway, it is striking that the Bayesian analysis of this important problem can be prosecuted under such minimal assumptions.

7 | CONCLUSION

In this paper we have developed perhaps the first Bayesian framework for analysing an important and broad class of semiparametric models in which the distribution of the outcomes is defined only up to a set of conditional moments, some of which may be misspecified. We have derived BvM theorems for the behaviour of the posterior distribution under both correct and incorrect specification of the conditional moments, and developed the theory for comparing different conditional moment models through a comparison of model marginal likelihoods. In addition, we have discussed settings with a high-dimensional Z and θ, the former addressed by a sparsity-based model search procedure, and the latter by the TaRB-MH MCMC algorithm for efficient posterior sampling.

Our theory and various examples, taken together, show that the developments in this paper make possible, for the first time, the formal (and practical) Bayesian analysis of a new, large class of problems that were hitherto difficult, or not possible, to tackle from the Bayesian viewpoint. This research we believe should have numerous positive ramifications for the growth and practice and teaching of Bayesian statistics.

DISCLAIMER

The views expressed in this paper are solely those of the authors and do not necessarily reflect the views of the Federal Reserve Bank of Philadelphia or the Federal Reserve System.

ACKNOWLEDGEMENT

We are enormously grateful to the editor, associate editor and referee for their constructive comments and penetrating questions that spawned many improvements.

ORCID

Siddhartha Chib https://orcid.org/0000-0003-2073-3124

REFERENCES

Ai, C. & Chen, X. (2003) Efficient estimation of models with conditional moment restrictions containing unknown functions. *Econometrica*, 71(6), 1795–1843.

Ai, C. & Chen, X. (2007) Estimation of possibly misspecified semiparametric conditional moment restriction models with different conditioning variables. *Journal of Econometrics*, 141(1), 5–43.

Bickel, P.J. & Kleijn, B.J.K. (2012) The semiparametric Bernstein-von Mises theorem. *Annals of Statistics*, 40(1), 206–237.

Carrasco, M. & Florens, J.-P. (2000) Generalization of GMM to a continuum of moment conditions. *Econometric Theory*, 16(6), 797–834.

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

Chen, X., Christensen, T. & Tamer, E.T. (2018) Monte Carlo confidence sets for identified sets. *Econometrica*, 86(6), 1965–2018.

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

Chib, S. & Greenberg, E. (1995) Understanding the Metropolis-Hastings algorithm. *The American Statistician*, 49(4), 327–335.

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

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

Chib, S. & Ramamurthy, S. (2010) Tailored randomized block MCMC methods with application to DSGE models. *Journal of Econometrics*, 155(1), 19–38.

Chib, S. & Zeng, X. (2020) Which factors are risk factors in asset pricing? A model scan framework. *Journal of Business & Economic Statistics*, 38, 771–783.

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

Csiszar, I. (1984) Sanov property, generalized *i*-projection and a conditional limit theorem. *Annals of Probability*, 12(3), 768–793.

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

Donald, S.G., Imbens, G.W. & Newey, W.K. (2003) Empirical likelihood estimation and consistent tests with conditional moment restrictions. *Journal of Econometrics*, 117(1), 55–93.

Donald, S.G., Imbens, G.W. & Newey, W.K. (2009) Choosing instrumental variables in conditional moment restriction models. *Journal of Econometrics*, 152(1), 28–36.

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

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

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

Guo, S. & Fraser, M.W. (2015) *Propensity score analysis: statistical methods and applications, advanced quantitative techniques in the social sciences*, 2nd edn, CA: Sage, Thousand Oaks.

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

Kitamura, Y. & Otsu, T. (2011) Bayesian analysis of moment condition models using nonparametric priors, Technical report, Yale University.

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

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

Liao, Y. & Jiang, W. (2011) Posterior consistency of nonparametric conditional moment restricted models. *Annals of Statistics*, 39(6), 3003–3031.
Liao, Y. & Simoni, A. (2019) Bayesian inference for partially identified smooth convex models. *Journal of Econometrics*, 211(2), 338–360.
Newey, W.K. (1997) Convergence rates and asymptotic normality for series estimators. *Journal of Econometrics*, 79(1), 147–168.
Schennach, S.M. (2005) Bayesian exponentially tilted empirical likelihood. *Biometrika*, 92(1), 31–46.
Shin, M. (2014) Bayesian GMM, Technical report, University of Pennsylvania.
Sueishi, N. (2013) Identification problem of the exponential tilting estimator under misspecification. *Economics Letters*, 118(3), 509–511.
Van der Vaart, A.W. (1998) Asymptotic Statistics, Cambridge Series in Statistical and Probabilistic Mathematics.
White, H. (1982) Maximum likelihood estimation of misspecified models. *Econometrica*, 50(1), 1–25.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

How to cite this article: Chib, S., Shin, M. & Simoni, A. (2021) Bayesian estimation and comparison of conditional moment models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 1–25. Available from: https://doi.org/10.1111/rssb.12484