Journal of Multivariate Analysis 145 (2016) 132–147

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

journal homepage: www.elsevier.com/locate/jmva

On oracle property and asymptotic validity of Bayesian generalized method of moments

Cheng Li [a,*], Wenxin Jiang [b]

[a] Department of Statistical Science, Duke University, Box 90251, Durham NC 27708, United States

[b] Department of Statistics, Northwestern University, 2006 Sheridan Road, Evanston IL 60208, United States

ARTICLE INFO

Article history:
Received 30 January 2015
Available online 29 December 2015

AMS subject classifications:
62F15
62F12

Keywords:
Bayesian
GEE (generalized estimating equations)
GMM (generalized method of moments)
MCMC
Model selection
Moment condition
Oracle property
Posterior validity

ABSTRACT

Statistical inference based on moment conditions and estimating equations is of substantial interest when it is difficult to specify a full probabilistic model. We propose a Bayesian flavored model selection framework based on (quasi-)posterior probabilities from the Bayesian Generalized Method of Moments (BGMM), which allows us to incorporate two important advantages of a Bayesian approach: the expressiveness of posterior distributions and the convenient computational method of Markov Chain Monte Carlo (MCMC). Theoretically we show that BGMM can achieve the posterior consistency for selecting the unknown true model, and that it possesses a Bayesian version of the oracle property, i.e. the posterior distribution for the parameter of interest is asymptotically normal and is as informative as if the true model were known. In addition, we show that the proposed quasi-posterior is valid to be interpreted as an approximate posterior distribution given a data summary. Our applications include modeling of correlated data, quantile regression, and graphical models based on partial correlations. We demonstrate the implementation of the BGMM model selection through numerical examples.

1. Introduction

We consider the estimation problem based on the following unconditional moment restrictions

$$\mathrm{E}\{g(D,\theta)\} = 0 \tag{1}$$

where D is a set of random variables with domain $\mathcal{D}$, θ is a p-dimensional vector of parameters to be estimated, and g is a m-dimensional mapping from $\mathcal{D} \times \mathbb{R}^p$ to $\mathbb{R}^m$. Typically it is necessary to have $m \geq p$ for the point identification of θ. Given an i.i.d. or stationary realization $\boldsymbol{D} = \{D_1, \ldots, D_n\}$ of D, one can estimate θ directly from such a set of m moment functions, without needing to fully specify the underlying data generating process of D. In this paper, we consider the case where in (1), the true parameter θ_0 could possibly lie in a lower dimensional subspace. Our goal is to consistently select the relevant variables and estimate their effects, namely the nonzero components of θ_0, when the specification of full probabilistic model is unavailable but a sufficient number of moment conditions are present.

We consider a Bayesian-flavored approach, where a quasi-posterior can be derived from a prior distribution and a quadratic form of moment restrictions. This enables us to accommodate two important advantages of the Bayesian approach: the expressiveness of the posterior distributions and the convenient computational method of MCMC. These are particularly

* Corresponding author.
E-mail addresses: cl332@stat.duke.edu (C. Li), wjiang@northwestern.edu (W. Jiang).

http://dx.doi.org/10.1016/j.jmva.2015.12.009
0047-259X/

useful for the model selection problem that we study. We are able to report the most probable model, the second most probable model and so on, together with their quasi-posterior probabilities, which are shown to be asymptotically valid in large samples. We can also use the reversible jump MCMC algorithm [15,9] to traverse the space of different models and simulate the quasi-posterior probabilities.

For this framework of moment-based Bayesian method of model selection and model averaging, our paper will prove several appealing fundamental theorems. They will address model selection consistency, oracle property, and valid interpretation of the quasi-posterior distribution. In the following, we will first review the related works and then describe in detail the contributions of our current paper.

1.1. GMM and BGMM

The moment based estimation problem (1) is important and has been extensively studied in econometrics and statistics. Well known methods include the generalized method of moments (GMM, [17,18,41]), the empirical likelihood (EL, [43,44]), the exponential tilting (ET, [30]), the exponential tilted empirical likelihood (ETEL, [46,47]) and the generalized empirical likelihood (GEL, [42]). Essentially they all share the same first order efficiency of optimally weighted GMM estimator, and have been applied to independent data, time series data and panel data in econometrics. On the other hand, researchers in statistics also use the moment based methods for constructing efficient estimators, especially for clustered and correlated longitudinal data. For example, Qu et al. [45] proposed a GMM type estimator to avoid the inefficiency from misspecified working correlation matrices in generalized estimating equations (GEE) for longitudinal data. Wang et al. [50] considered the EL approach to address the within-subject correlation structure. Recently frequentist penalization methods have been proposed to accommodate increasing dimension p. See for example [51,34,8,4], etc. In general, the moment based estimation methods only require information on the low order moments of D and are therefore more flexible, efficient and robust to model misspecification, as long as the moment conditions are correctly specified.

Our work focuses on the Bayesian inference of θ under the moment constraint (1). Compared to the abundance of frequentist literature, the development of Bayesian methods on this problem still remains limited. One difficulty that hinders the fully probabilistic Bayesian modeling is that some prior distribution on both the distribution of D (denoted as P_D) and the parameter θ needs to be specified, such that the pair (P_D, θ) satisfies the set of restrictions (1). Recent progress in this direction includes Kitamura and Otsu [29] and Florens and Simoni [13]. Kitamura and Otsu [29] tried to minimize the Kullback–Leibler divergence of P_D to a Dirichlet process, which leads to an ET type likelihood function that computationally requires optimizations within each MCMC iteration step. Florens and Simoni [13] exploited the Gaussian process prior and required a functional transformation of the data that is only asymptotically Gaussian, which still leads to a misspecified likelihood function in finite samples. Besides, both methods have only been tested on simple examples that involve a few parameters and moments. Instead, another analytically simpler Bayesian way of modeling (1) is *the Bayesian generalized method of moments* (BGMM), first proposed and studied by Kim [27] and Chernozhukov and Hong [6], which constructs the simple quasi-likelihood function

$$q(\boldsymbol{D}|\theta) = \frac{1}{\det\left(2\pi \boldsymbol{V}_n/n\right)^{\frac{1}{2}}} \exp\left\{-\frac{n}{2}\bar{g}(\boldsymbol{D},\theta)^\top \boldsymbol{V}_n^{-1}\bar{g}(\boldsymbol{D},\theta)\right\}, \tag{2}$$

where $\bar{g}(\boldsymbol{D},\theta)$ is the sample average of $g(D_i,\theta)$, $i = 1,\ldots,n$, $\boldsymbol{V}_n$ is a $m \times m$ positive definite matrix that could possibly depend on the data $\boldsymbol{D}$, and $\det(\boldsymbol{A})$ denotes the determinant of a matrix $\boldsymbol{A}$. Hereafter we use the symbol "q" to denote the quasi-likelihood function and the quasi-posterior. This quasi-likelihood function has been studied under a Bayesian framework in [27] and is named *the limited information likelihood (LIL)*, which minimizes the Kullback–Leibler divergence of the true data generating process P_D to the set of all distributions satisfying the less restrictive asymptotic constraint $\lim_{n\to\infty} \mathrm{E}\left\{n\bar{g}(\boldsymbol{D},\theta_0)^\top \boldsymbol{V}_n^{-1}\bar{g}(\boldsymbol{D},\theta_0)\right\}/m = 1$. This relation holds when we choose $\boldsymbol{V}_n$ to be a consistent estimator of the covariance matrix $\mathrm{Var}(g(D,\theta_0))$. Given a prior distribution $\pi(\theta)$, the quasi-posterior takes the form

$$q(\theta|\boldsymbol{D}) \propto \frac{1}{\det\left(2\pi \boldsymbol{V}_n/n\right)^{\frac{1}{2}}} \exp\left\{-\frac{n}{2}\bar{g}(\boldsymbol{D},\theta)^\top \boldsymbol{V}_n^{-1}\bar{g}(\boldsymbol{D},\theta)\right\}\pi(\theta). \tag{3}$$

By using $q(\boldsymbol{D}|\theta)$ in the Bayesian model, we only need to specify a prior on θ and thus circumvent the difficulty of directly assigning a prior on the pair (P_D,θ) with constraints (1). In the computational aspect, $q(\boldsymbol{D}|\theta)$ takes an explicit analytical form that allows straightforward MCMC updating for the corresponding Bayesian posterior without any iterative optimization steps [6]. Furthermore, when $\boldsymbol{V}_n$ is chosen as a consistent estimator of $\mathrm{Var}(g(D,\theta_0))$, the exponential part of $q(\boldsymbol{D}|\theta)$ resembles the optimally weighted GMM criterion function [17], which in large samples can be viewed as a second order approximation to the true negative log-likelihood function that follows a chi-square distribution with p degrees of freedom if $m = p$ and both are fixed [52].

The theoretical properties of BGMM have been investigated extensively in Chernozhukov and Hong [6] and Belloni and Chernozhukov [2], who show that a Bernstein–von Mises theorem holds, i.e. the posterior distribution converges asymptotically to normal. The computational aspects of BGMM with no model selection have been investigated in [52,53]. Kim [28] has established the pairwise consistency theoretically when each candidate model is compared to the true model separately, and has used MCMC in simulations for such model comparison. Hong and Preston [19] have discussed a more

general Bayesian model selection framework including BGMM as well as Bayesian GEL, and have studied the consistency of Bayes factors and Bayesian information criterion (BIC) under both nested and nonnested scenarios (see a more detailed comparison later in Remark 3 in Section 2). Other applications of BGMM include the moment inequality models [38] and the nonparametric instrumental regression [39,26]. However, theoretical properties of BGMM, such as the limiting distribution and the posterior interpretation, have not been systematically studied in the context of model selection with increasing dimensionality.

1.2. Contributions of current paper

We study theoretical properties of BGMM in the context of model selection. The detailed contributions of the current paper include the following:

1. We prove that BGMM automatically achieves the "global model selection consistency" (see, e.g., [25]) under some regularity conditions on the moment function $g(D, \theta)$ and the prior. This is to say that the BGMM posterior probability of the true model converges to 1 with high probability.
2. We derive an oracle property for the BGMM procedure, which states that the BGMM posterior distribution converges in total variation norm to a normal distribution concentrated on the true model space with an efficient variance, as if the true model were known. This oracle property is the Bayesian analog of the frequentist post-model-selection oracle property of Fan and Li [11], and is comparable to the Bayesian oracle property proposed by Ishwaran and Rao [21]. While Ishwaran and Rao [21] showed this oracle property only for the posterior mean estimator in the normal linear model, our version of Bayesian oracle property studies the global asymptotic concentration behavior of the whole posterior for the general form of moment conditions. We apply BGMM to our motivating examples in Section 1.3 and show that the model selection consistency and oracle property hold under mild regularity conditions on the data and the moments.
3. Our theory for BGMM allows the number of parameters p to increase with the sample size n. This is technically challenging because the number of candidate models 2^p will increase exponentially fast with n. Although Hong and Preston [19] and Kim [28] have established model selection consistency for BGMM with a fixed number of models, their techniques based on pairwise model comparison are not sufficient for showing the global model selection consistency under our increasing dimensional setup. Our theoretical results accommodate an increasing dimension p that satisfies $p^4/n \to 0$ up to some logarithm factors, which is the same as the growth rate in Belloni and Chernozhukov [2], who studied BGMM without model selection.
4. We present a novel interpretation of the BGMM quasi-posterior, as an approximate posterior conditional on a data summary that is equivalent to the GMM estimator. Particularly for model selection, we derive the convergence rates of Bayes factors for the BGMM method and the fully Bayesian method given the GMM estimator, and show that they have similar asymptotic behavior. Therefore, the model posterior probabilities from BGMM are asymptotically valid and can be used directly for comparing different models.
5. Our numerical experiments provide practical guidance on the MCMC computation in a complicated setup with 2^p candidate models. The previous works on BGMM computation either have not considered the model selection problem [53], or have only considered pairwise model comparison using MCMC (e.g. [28]). We implement the reversible jump MCMC algorithm and demonstrate BGMM as a practically feasible and efficient alternative to the frequentist regularization methods.

Below we provide some motivating examples that involve the moment condition (1) and can be easily incorporated into the BGMM framework.

1.3. Three motivating examples

The moment condition model (1) is much more general than probabilistic models such as the normal linear model and generalized linear models, since one could set the moment function to be $g(D, \theta) = \partial_\theta \ln p(D|\theta)$, where $p(D|\theta)$ is the probability density of D. For example, in the Poisson regression model $D = (Y, X^\top)^\top$ and $Y|X \sim \mathcal{P}(e^{X^\top\theta})$ for some covariates X, we can use the moment function $g(D, \theta) = X\left(Y - e^{X^\top\theta}\right)$ for quasi-posterior based inference from (3), although the likelihood based inference would be more straightforward in this case. In fact, the proposed moment based method has more flexibility when only the lower order moments or quantiles are specified rather than the complete probabilistic model, as described in the following examples.

Example 1 (*Correlated Longitudinal Data*)**.** In longitudinal studies, suppose the jth observation for the ith subject is a scalar response variable Y_{ij} and a p-dimensional covariate vector X_{ij}. For simplicity, we assume that each subject has the same number of observations, i.e. $j = 1, \ldots, s$ and $i = 1, \ldots, n$. Let $Y_i = (Y_{i1}, \ldots, Y_{is})^\top$, $\boldsymbol{X}_i = (X_{i1}, \ldots, X_{is})^\top$, and $\mathrm{E}(Y_i|\boldsymbol{X}_i) = \mu_i(\theta)$, where $\mu_i(\theta) = (\mu(X_{i1}^\top\theta), \ldots, \mu(X_{is}^\top\theta))^\top$ and $\mu(\cdot)$ is a monotone link function. To account for the heteroscedasticity, we assume the conditional variance of Y_{ij} given X_{ij} is a function of the single index $X_{ij}^\top\theta$, i.e. $\mathrm{Var}(Y_{ij}|X_{ij}) = \phi(X_{ij}^\top\theta)$. Then the

frequentist GEE method estimates θ by solving equation

$$n^{-1}\sum_{i=1}^{n}\frac{\partial\mu_i(\theta)^\top}{\partial\theta}\boldsymbol{S}_i^{-1}(Y_i-\mu_i(\theta))=0 \tag{4}$$

where $\boldsymbol{S}_i=\boldsymbol{A}_i^{1/2}\boldsymbol{R}\boldsymbol{A}_i^{1/2}$, $\boldsymbol{A}_i=\boldsymbol{A}_i(\theta)=\operatorname{diag}\{\phi(X_{i1}^\top\theta),\ldots,\phi(X_{is}^\top\theta)\}$ is the diagonal matrix with the conditional variance of Y given X and $\boldsymbol{R}$ is a working correlation matrix. If we denote the data as $D_i=(Y_i,\boldsymbol{X}_i)^\top$, then the moment function is defined by

$$g(D_i,\theta)=\frac{\partial\mu_i(\theta)^\top}{\partial\theta}\boldsymbol{S}_i^{-1}(Y_i-\mu_i(\theta)). \tag{5}$$

And the moment condition (1) is satisfied.

Example 2 (*Quantile Regression*)**.** Suppose that Y is a continuously distributed response variable, and X is a p-dimensional predictor vector for the τth quantile ($\tau\in(0,1)$) of Y. The conditional quantile function of Y given X is specified by $F_{Y|X}^{-1}(\tau)=X^\top\theta$, where $F_{Y|X}^{-1}$ is the generalized inverse of conditional distribution function of Y given X. Then let $D=(Y,X^\top)^\top$ and we can construct p moment functions as

$$g(D,\theta)=X\left\{1(Y-X^\top\theta\le 0)-\tau\right\}, \tag{6}$$

where $1(\cdot)$ is the indicator function.

Example 3 (*Partial Correlation Selection*)**.** The partial correlation structure of a s-dimensional random vector Y is specified by its precision matrix $\boldsymbol{\Omega}=\boldsymbol{\Sigma}^{-1}$, where $\boldsymbol{\Sigma}=\mathrm{E}\left\{(Y-\mathrm{E}\,Y)(Y-\mathrm{E}\,Y)^\top\right\}$ is the covariance matrix of Y. Hereafter without loss of generality, we assume that Y is centered such that $\mathrm{E}\,Y=0$. The partial correlation between the ith and the jth components of Y is defined by $\rho_{ij}=-\omega_{ij}/\sqrt{\omega_{ii}\omega_{jj}}$, where ω_{ij} denotes the (i,j)th entry of $\boldsymbol{\Omega}$. $\omega_{ij}=0$ implies zero partial correlation between the ith and the jth components of Y given all the other components. For multivariate Gaussian random vector, there is an equivalence between the conditional independence and the zero partial correlation. In the general case where multivariate Gaussian assumption is not satisfied, we can still use the second moment of Y to identify the zero entries in $\boldsymbol{\Omega}$. Let θ be the vectorized upper triangle part of $\boldsymbol{\Omega}$. Then we can define the moment function

$$g_{ij}(Y,\theta)=Y_iY_j-(\boldsymbol{\Omega}^{-1})_{ij}, \tag{7}$$

for $1\le i\le j\le s$, and the stacked moment vector $g(Y,\theta)$ satisfies (1). We have $\dim(\theta)=\dim(g)=s(s+1)/2=:p$ where θ is just identifiable. The model selection problem for partial correlation has been studied in, for example, [10,24], etc.

1.4. Organization of the paper

The rest of the paper is organized as follows. In Section 2.2, we derive the oracle properties for BGMM model selection based on a set of high level assumptions. In Section 2.3, we discuss the validity of the proposed BGMM quasi-posterior. Section 3 provides the algorithm we use for BGMM and numerical experiments to illustrate the empirical performance of BGMM model selection. Section 4 includes further discussions. We check these assumptions for the three motivating examples in Section 1.3 and include the technical proofs of all theorems in the supplementary material (see Appendix A). A real data application can be found in the online technical report by Li and Jiang [35].

1.5. Some useful notation

We define some useful notation. Let $|\cdot|_k$ denote the L_k norm for $k\in[0,\infty]$ and $\|\cdot\|$ be the Euclidean norm (L_2 norm). For any generic square matrix $\boldsymbol{C}$, let $\underline{\lambda}(\boldsymbol{C})$, $\bar{\lambda}(\boldsymbol{C})$ denote the smallest and the largest eigenvalues of a square matrix $\boldsymbol{C}$. Let $\|\boldsymbol{C}\|=\sqrt{\bar{\lambda}(\boldsymbol{C}^\top\boldsymbol{C})}$ be the matrix operator norm. For two stochastic sequences $\{a_n\}$ and $\{b_n\}$, let $a_n\prec b_n$, $a_n\succ b_n$ and $a_n\asymp b_n$ denote $a_n=o(b_n)$, $b_n=o(a_n)$ and a_n, b_n having the same order as $n\to\infty$. $a\vee b=\max(a,b)$ and $a\wedge b=\min(a,b)$. The notations o_p and O_p always refer to the probability measure $P_{\boldsymbol{D}}$ of the sample $\boldsymbol{D}$. We use "C" to denote any generic constant whose value can change in different places. We use the statement "the event A happens w.p.a.1 as $n\to\infty$" as an abbreviation for the statement "the event A happens with $P_{\boldsymbol{D}}$ probability approaching 1 as $n\to\infty$", i.e. $\lim_{n\to\infty}P_{\boldsymbol{D}}(A)=1$.

2. Theoretical properties of Bayesian GMM model selection

The Bayesian model selection problem has been extensively studied, but mostly for normal linear regression models and generalized linear models. See for example, [7,48,20,22,36,25,37], etc. Our Bayesian model selection is substantially different from all these papers. Instead of having a probabilistic model such as the simple normal linear model, we work with the moment conditions (1) and do model selection using BGMM. Our true parameter θ_0 is the unique solution of (1)

and possibly lies in a lower dimensional subspace of the whole parameter space $\Theta \subseteq \mathbb{R}^p$. We restrict Θ to be a compact and connected set in $\mathbb{R}^p$, with finite L_2 radius $R = \sup_{\theta\in\Theta} \|\theta\|$ for some large constant $R > 0$.

Without loss of generality, in the following we will consider models generated by all the possible coordinate subspaces of $\mathbb{R}^p$, which leads to a total of 2^p different models $\mathcal{M}$ and the parameter space partition $\Theta = \bigcup_{|\mathcal{M}|\le p} \Theta(\mathcal{M})$. Let $k = |\mathcal{M}|$ $(0 \le k \le p)$ be the size of a generic model $\mathcal{M}$, which is the number of nonzero components in any $\theta \in \mathcal{M}$. Suppose $\mathcal{M}_0$ is the true model space that contains θ_0, and $k_0 = |\mathcal{M}_0|$ is the dimension of θ_0. For a given model $\mathcal{M}$ and a generic θ, let $\theta = (\theta_1^\top, \theta_2^\top)^\top$ where $\theta_1 \in \mathbb{R}^k$ and $\theta_2 \in \mathbb{R}^{p-k}$ correspond to the components that lie in and outside $\Theta(\mathcal{M})$, respectively. So $\theta_2 = 0$ if $\theta \in \Theta(\mathcal{M})$. We emphasize that the meaning of subscripts "1" and "2" can change with the model index $\mathcal{M}$.

For such a model selection setup, the prior distribution can be written in the hierarchical structure $\pi(\theta) = \sum_M \pi(\theta|\mathcal{M})\pi(\mathcal{M}) = \sum_{\mathcal{M},k} \pi(\theta|\mathcal{M})\pi(\mathcal{M}|\,|\mathcal{M}| = k)\pi(k)$ for $k = 0, 1, \ldots, p$. If a model $\mathcal{M}$ does not contain all the nonzero components for a given θ, then $\pi(\theta|\mathcal{M}) = 0$. We assume that each $\pi(\theta|\mathcal{M})$ has a density function. For two different models $\mathcal{M}_1$ and $\mathcal{M}_2$, the (quasi-) Bayes factor of $\mathcal{M}_1$ with respect to $\mathcal{M}_2$ is defined as

$$\mathrm{BF_q}[\mathcal{M}_1 : \mathcal{M}_2] = \frac{q(\boldsymbol{D}|\mathcal{M}_1)}{q(\boldsymbol{D}|\mathcal{M}_2)} = \frac{\int_{\Theta(\mathcal{M}_1)} q(\boldsymbol{D}|\theta, \mathcal{M}_1)\pi(\theta|\mathcal{M}_1)\mathrm{d}\theta}{\int_{\Theta(\mathcal{M}_2)} q(\boldsymbol{D}|\theta, \mathcal{M}_2)\pi(\theta|\mathcal{M}_2)\mathrm{d}\theta} \tag{8}$$

and accordingly the (quasi-) posterior odds are the product of the Bayes factor and the prior odds

$$\mathrm{PO_q}[\mathcal{M}_1 : \mathcal{M}_2] = \frac{q(\boldsymbol{D}|\mathcal{M}_1)}{q(\boldsymbol{D}|\mathcal{M}_2)} \cdot \frac{\pi(\mathcal{M}_1)}{\pi(\mathcal{M}_2)} = \frac{\int_{\Theta(\mathcal{M}_1)} q(\boldsymbol{D}|\theta, \mathcal{M}_1)\pi(\theta|\mathcal{M}_1)\mathrm{d}\theta}{\int_{\Theta(\mathcal{M}_2)} q(\boldsymbol{D}|\theta, \mathcal{M}_2)\pi(\theta|\mathcal{M}_2)\mathrm{d}\theta} \cdot \frac{\pi(\mathcal{M}_1)}{\pi(\mathcal{M}_2)}. \tag{9}$$

The model selection consistency we are going to establish is the global model selection consistency [25], in the sense that asymptotically the true model $\mathcal{M}_0$ will not only be the MAP model (*maximum a posteriori*) but also have posterior probability tending to 1. Equivalently, we will show that the sum of all posterior odds $\mathrm{PO_q}[\mathcal{M} : \mathcal{M}_0]$ with $\mathcal{M} \neq \mathcal{M}_0$ converges to zero in probability. This strongest mode of consistency implies that the posterior mass will be concentrated around the true model and most of the 2^p models receive negligible probabilities. This is a desirable property in practice for interpretation, since commonly used Bayesian estimation procedures such as model averaging will then involve only a few models instead of many candidate models.

2.1. Assumptions

The set of assumptions below follows closely the set of conditions for Z-estimation in [2]. They are high level assumptions imposed on the data generation process, the model parameters, the moment conditions and the priors. For a specific model, these assumptions are not necessarily in the most general form, but they do cover a wide class of moment condition models in practice and are sufficient for illustrating the theoretical properties of BGMM.

For the data generation process and the true parameter θ_0, we make the following assumptions.

Assumption 1 (*Data Generation Process*)**.** $\{D_i, i = 1, \ldots, n\}$ is an i.i.d. sequence. $\mathrm{E}\, g(D, \theta_0) = 0$ for some $\theta_0 \in \Theta$. Θ is a compact and connected set with L_2 radius R for some large constant $R > 0$, and it contains an open neighborhood of θ_0.

Assumption 2 (*Dimension*)**.** Let $\dim(\theta) = p$ and $\dim(g) = m$. Assume that $p \le m$, $p \asymp m$, $p^4 \ln^2 n/n \to 0$ and $p^{2+\alpha} \ln n/n^\alpha \to 0$, where α is defined in Assumption 4.

Assumption 3 (*Beta-Min*)**.** Let $\epsilon_n = \sqrt{p/n}$. Assume $1 \succeq \min_{j\in\mathcal{M}_0} |\theta_{0,(j)}| \succ \sqrt{\ln n}\epsilon_n$, where $\theta_{0,(j)}$'s for $j \in \mathcal{M}_0$ denote the nonzero components of the true parameter θ_0.

The i.i.d. assumption in Assumption 1 can be possibly relaxed to a weakly dependent stationary process using more involved techniques. The compactness assumption for the parameter space Θ is standard and mainly for technical convenience, and it can be relaxed to the full space of $\mathbb{R}^p$ if we can control the tail behavior of the prior (see the discussion after Assumptions 7 and 8). Assumption 2 allows increasing dimension p, and the growth rate of p is comparable with those in [2,8,49,34], etc. The beta-min condition in Assumption 3 is commonly used in the frequentist GEE literature (see e.g. [51,34,8]). It gives the minimal magnitude of nonzero coefficients that could be detected by BGMM.

Let $B_0(\epsilon) = \{\theta \in \Theta : \|\theta - \theta_0\| < \epsilon\}$ for any $\epsilon > 0$. We make the following assumptions on the moment conditions.

Assumption 4 (*Moment*)**.**

(i) The moment function $g(D, \theta)$ satisfies the continuity property

$$\sup_{\eta\in\mathbb{R}^m, \|\eta\|=1} \left[\mathrm{E}\left\{(\eta^\top(g(D,\theta) - g(D,\theta_0)))^2\right\}\right]^{1/2} \le O\left((\sqrt{p}\|\theta - \theta_0\|)^\alpha\right),$$

uniformly in $\theta \in \Theta$ for some constant $\alpha \in (0, 1]$.

(ii) The class of functions $\mathcal{F} = \{\eta^\top(g(D,\theta) - g(D,\theta_0)), \theta \in \Theta, \eta \in \mathbb{R}^m, \|\eta\| = 1\}$ has an envelope function F almost surely bounded in L_2 norm $\|\cdot\|_{P_D,2}$ as order $O(\sqrt{p})$. The L_2 uniform covering number $N(\epsilon\|F\|_{P_D,2}, \mathcal{F}, L_2(P_D))$ satisfies that for any small $\epsilon > 0$,

$$\ln N(\epsilon\|F\|_{P_D,2}, \mathcal{F}, L_2(P_D)) = O\Big(p \ln\Big(\frac{n}{\epsilon}\Big)\Big).$$

Assumption 5 (*Linearization*)**.**

(i) $\|\mathrm{E}\, g(D,\theta)\| \geq \delta_0 \wedge (\delta_1\|\theta - \theta_0\|)$ uniformly on Θ for some positive constants δ_0, δ_1.
(ii) $\boldsymbol{G} := \nabla_\theta \mathrm{E}\, g(D,\theta_0)$ exists, and the eigenvalues of $\boldsymbol{G}^\top\boldsymbol{G}$ are bounded from below and above as $n \to \infty$.
(iii) $\boldsymbol{H}(\theta) := \nabla^2_{\theta\theta^\top} \mathrm{E}\, g(D,\theta)$ exists for $\theta \in B_0(C\epsilon_n)$, and uniformly over $\theta \in B_0(C\epsilon_n)$ for any fixed $C > 0$, $\sup_{\|u\|=1,\|v\|=1,u,v\in\mathbb{R}^p} \|\boldsymbol{H}(\theta)(u,v)\| = O(\sqrt{p})$.

Assumptions 4 and 5 on moment function $g(D,\theta)$ parallel the conditions ZE.1 and ZE.2 in [2] respectively. The continuity index α in Assumption 4(i) satisfies $\alpha = 1$ for the mean regression, such as the examples of correlated longitudinal data and partial correlation selection, and $\alpha = 1/2$ for the quantile regression model. The entropy condition in Assumption 4(ii) controls the complexity of the class of moment functions $g(D,\theta)$. Assumption 5(i) guarantees the point identification of the true parameter θ_0, and part (ii) and (iii) impose mild assumptions on the first and second derivatives of $\mathrm{E}\, g(D,\theta)$ around θ_0. These regularity conditions are mainly used to derive large deviation bounds via empirical process results, and they will be verified later for our motivating examples. Note that unlike Wang et al. [51], Leng and Tang [34] and Cho and Qu [8], we do not require the moment function $g(D,\theta)$ itself to be differentiable. This allows more general applications to discontinuous $g(D,\theta)$, such as in the case of quantile regression.

Assumption 6 (*Variance*)**.** $\boldsymbol{V}_n$ is a positive definite matrix for all n, and converges in the matrix operator norm to $\boldsymbol{V} = \mathrm{Var}\{g(D,\theta_0)\}$. The eigenvalues of $\boldsymbol{V}_n$ and $\boldsymbol{V}$ are bounded below and above for some positive constants $\underline{\lambda}$ and $\bar{\lambda}$ w.p.a.1 as $n \to \infty$.

Assumption 6 assumes that the positive definite weighting matrix $\boldsymbol{V}_n$ is a consistent estimator of the covariance matrix of $g(D,\theta)$ at θ_0, similar to the preliminary estimator of the optimal weighting matrix used in the two step GMM estimation. Although this consistency of $\boldsymbol{V}_n$ to $\boldsymbol{V}$ is not required for the model selection consistency, it is necessary for the valid posterior inference such as posterior credible sets. Essentially $\boldsymbol{V}_n$ needs to satisfy the generalized information inequality [6], such that the LIL asymptotically satisfies the second Bartlett identity as a true likelihood function does. Such consistent estimator $\boldsymbol{V}_n$ usually exists for all our motivating examples.

Finally we impose the following assumptions on the prior.

Assumption 7 (*Prior on* θ)**.**

(i) $\pi(\theta|\mathcal{M})$ has a density function restricted to Θ, and is bounded above by a constant c_π uniformly over all model spaces $\mathcal{M}$.
(ii) Suppose $\theta = (\theta_1^\top, \theta_2^\top)^\top$ is decomposed according to the model $\mathcal{M}$. Then uniformly over all models $\mathcal{M} \supseteq \mathcal{M}_0$, for any given $C > 0$, $|\ln\pi(\theta_1|\mathcal{M}) - \ln\pi(\theta_{0,\mathcal{M},1}|\mathcal{M})| = o(1)$ as $n \to \infty$ if $\theta = (\theta_1^\top, 0^\top)^\top \in \Theta(\mathcal{M}) \cap B_0(C\epsilon_n)$, where $\theta_{0,\mathcal{M},1}$ is the subvector of the true parameter θ_0 restricted to $\Theta(\mathcal{M})$.
(iii) Uniformly for all models $\mathcal{M} \supseteq \mathcal{M}_0$, there exist constants $c_0, c_1 > 0$, such that $\pi(\theta_0|\mathcal{M}) \geq e^{-c_0|\mathcal{M}|}$ and $|\ln\pi(\theta_{0,\mathcal{M},1}|\mathcal{M}) - \ln\pi(\theta_0|\mathcal{M}_0)| \leq c_1(|\mathcal{M}| - |\mathcal{M}_0|)$.

Assumption 8 (*Prior on Models*)**.** The model prior $\pi(\mathcal{M})$ satisfies:

(i) $\lim_{n\to\infty} \sup_{\{\mathcal{M}:\mathcal{M}\supset\mathcal{M}_0\}} (\pi(\mathcal{M})/\pi(\mathcal{M}_0)) \big(p/\sqrt{n}\big)^{|\mathcal{M}|-|\mathcal{M}_0|} = 0$.
(ii) $\sup_{\{\mathcal{M}:\ \mathcal{M}_0\setminus\mathcal{M}\neq\emptyset\}} \pi(\mathcal{M})/\pi(\mathcal{M}_0) \preceq e^{r_1 p \ln n}$ for some constant $r_1 > 0$ as $n \to \infty$.

Our Assumption 7(i) that the prior is restricted on a compact set Θ is mainly for technical simplification, which can be relaxed as long as the tail probability of $\pi(\theta|\mathcal{M})$ decays sufficiently fast on each model $\mathcal{M}$. The idea is that the one can divide the possibly noncompact support into a compact set Θ_n with radius increasing sufficiently slowly with n, and let the prior mass outside Θ_n be negligible as in the large n asymptotics [23]. Other than this, Assumption 7(i) is mild and encompasses most of the commonly used priors truncated on Θ. Assumption 7(ii) and (iii) for the priors on parameters $\pi(\theta|\mathcal{M})$ are satisfied by, for example, a uniform prior on the model space $\Theta(\mathcal{M})$, or a truncated multivariate normal prior on $\Theta(\mathcal{M})$.

Assumption 8 requires that the models larger than the true model $\mathcal{M}_0$ do not receive overly large prior mass, and the prior on the true model cannot be exponentially small compared to any other models. This is automatically true when p does not increase with n, provided that $\pi(\mathcal{M}_0)$ is a positive constant. With increasing dimensions, these requirements can be satisfied by, for example, a prior where each coordinate enters the model independently with a fixed probability $\nu \in (0,1)$, which includes the uniform prior as a special case if $\nu = 0.5$. Other examples include priors that propose a model size $|\mathcal{M}|$ according to Poisson or geometric distributions upper truncated at p, while all models of the same size are equally likely. A detailed verification of Assumption 8 for these priors can be found in Section 4 of the supplementary material (see Appendix A).

2.2. Oracle properties of BGMM

With all these assumptions, we now state the main results as follows. The proof of Theorem 1 is given in the supplementary material (see Appendix A).

Theorem 1. *Suppose Assumptions 1–8 hold. Then*

(i) *(Model Selection Consistency)*

$$q(\mathcal{M}_0|\boldsymbol{D}) \to 1, \quad \text{w.p.a.1 as } n \to \infty$$

that is, the quasi-posterior probability of the true model converges to 1, *w.p.a.*1 *as* $n \to \infty$.

(ii) *(Posterior Asymptotic Normality) Given a model* $\mathcal{M}$, *let* $\boldsymbol{G}_{\mathcal{M}}$ *be the submatrix of the derivative matrix* $\boldsymbol{G}$ *with respect to the subvector* θ_1 *in* $\theta = (\theta_1^\top, \theta_2^\top)^\top$. *Let* $\bar{\theta}_{\mathcal{M}_0,1} = \theta_{0,\mathcal{M}_0,1} - (\boldsymbol{G}_{\mathcal{M}_0}^\top \boldsymbol{V}_n^{-1} \boldsymbol{G}_{\mathcal{M}_0})^{-1} \boldsymbol{G}_{\mathcal{M}_0}^\top \boldsymbol{V}_n^{-1} \bar{g}(\boldsymbol{D}, \theta_0)$, *where* $\theta_{0,\mathcal{M}_0,1}$ *is the subvector of* θ_0 *restricted to* $\Theta(\mathcal{M}_0)$. *Then w.p.a.*1 *as* $n \to \infty$,

$$\sup_{A \subseteq \Theta} \left| \int_A q(\theta|\boldsymbol{D}) \mathrm{d}\theta - \int_{A \cap \Theta(\mathcal{M}_0)} \phi\left(\theta_1; \bar{\theta}_{\mathcal{M}_0,1}, (\boldsymbol{G}_{\mathcal{M}_0}^\top \boldsymbol{V}_n^{-1} \boldsymbol{G}_{\mathcal{M}_0})^{-1}/n\right) \mathrm{d}\theta_1 \right| \to 0,$$

where $\phi(\cdot; \mu, \boldsymbol{\Sigma})$ *is the normal density with mean* μ *and covariance matrix* $\boldsymbol{\Sigma}$, *and* $\theta = (\theta_1^\top, \theta_2^\top)^\top$ *is decomposed according to the true model* $\mathcal{M}_0$.

Part (i) of Theorem 1 establishes the global model selection consistency of BGMM, similar to previous Bayesian results from [25,37] for the normal linear model and the generalized linear models. Based on the BGMM posterior, the zero components of the true parameter θ_0 are estimated to be zero with $P_{\boldsymbol{D}}$ probability approaching 1. It also implies that asymptotically the MAP model $\hat{\mathcal{M}}$ converges to the true model $\mathcal{M}_0$ in $P_{\boldsymbol{D}}$-probability. This parallels the frequentist model selection results via penalization for moment based models and estimating equations [51,34,8], etc.

Part (ii) of Theorem 1 establishes an asymptotic normality result, in the sense that the total variation difference between the BGMM posterior measure and a k_0-dimensional normal distribution concentrated on the true model converges to zero in probability as the sample size increases. This is a direct extension of the Bayesian CLT result in [6,2] from a single full model space to the joint of all submodel spaces. Because the BGMM posterior is a mixture distribution on 2^p model spaces $\Theta(\mathcal{M})$ with different dimensions, we do not present result using the L_1 distance between two densities $q(\theta|\boldsymbol{D})$ and $\phi(\theta_1; \bar{\theta}_{\mathcal{M}_0,1}, (\boldsymbol{G}_{\mathcal{M}_0}^\top \boldsymbol{V}_n^{-1} \boldsymbol{G}_{\mathcal{M}_0})^{-1}/n)$. The asymptotic mean of the normal distribution $\bar{\theta}_{\mathcal{M}_0,1}$ is the first order approximation to $\hat{\theta}_{\mathcal{M}_0,1} = \arg\min_{\theta \in \Theta(\mathcal{M}_0)} \bar{g}(\boldsymbol{D}, \theta)^\top \boldsymbol{V}_n^{-1} \bar{g}(\boldsymbol{D}, \theta)$, i.e. the GMM estimator restricted to the subspace $\Theta(\mathcal{M}_0)$. Furthermore, given Assumption 6, the generalized information equality is satisfied [6], and the asymptotic variance of the limiting normal distribution is the same as the corresponding frequentist variance of the GMM estimator $\hat{\theta}_{\mathcal{M}_0,1}$.

Remark 1 (*Bayesian Oracle Property*)**.** The conclusion of Theorem 1 can be written heuristically as follows: Let $\theta = (\theta_1^\top, \theta_2^\top)^\top$ be decomposed according to the true model $\mathcal{M}_0$, then

(i) $\theta_2|\boldsymbol{D} \approx 0$ w.p.a.1 as $n \to \infty$;
(ii) $q(\theta_1|\boldsymbol{D}) \approx \mathcal{N}\left(\bar{\theta}_{\mathcal{M}_0,1}, (\boldsymbol{G}_{\mathcal{M}_0}^\top \boldsymbol{V}^{-1} \boldsymbol{G}_{\mathcal{M}_0})^{-1}/n\right)$ w.p.a.1 as $n \to \infty$.

The zero components in θ_0 are estimated to be zero given the data $\boldsymbol{D}$ with large probability, and the nonzero components in θ_0 almost follow a normal distribution centered at the first order approximation to the GMM estimator under the model $\mathcal{M}_0$, with the same optimal GMM asymptotic variance matrix, as if the true model $\mathcal{M}_0$ were known. We call this the *Bayesian oracle property* for model selection, which resembles the frequentist oracle property for penalized likelihood in [11]. Theorem 1 guarantees that the BGMM posterior will automatically identify the unknown true model, and automatically converge to an asymptotic normal distribution centered around the unknown true parameter with the optimal GMM variance, as if the true model were known. Compared to the oracle property in [21], our version is much stronger in two aspects: 1. Our model assumptions are based on the general form of moment conditions (1) and are therefore more general than the normal linear regression model in [21]; 2. Our oracle property characterizes the overall shrinkage of posterior distribution to an asymptotic normal distribution on the true model, while [21] only considered the asymptotics of the posterior mean estimator.

Remark 2. We explain why the oracle center $\bar{\theta}_{\mathcal{M}_0,1}$ (of the asymptotic normal approximation to the quasi posterior) is a desirable result. Roughly speaking, this oracle center will be often close to the unknown nonzero components of the true parameter θ_0 in large samples, since their difference has the order $O_p\left(\sqrt{p/n}\right)$. This oracle center is also similar to the center of Bayesian CLT in [2] and it applies to all our motivating examples in Section 1.3. Furthermore, in many cases we have the higher order approximation from $\bar{\theta}_{\mathcal{M}_0,1}$ to the GMM estimator $\hat{\theta}_{\mathcal{M}_0,1}$, with the difference $\|\bar{\theta}_{\mathcal{M}_0,1} - \hat{\theta}_{\mathcal{M}_0,1}\| = O_p(p/n)$, following the stochastic expansion of GMM estimator in [42]. When this high order approximation holds, the oracle center $\bar{\theta}_{\mathcal{M}_0,1}$ in Theorem 1(ii) is equivalent to and can be replaced by the oracle GMM estimator $\hat{\theta}_{\mathcal{M}_0,1}$.

Remark 3. Our work in model selection of BGMM may be regarded as a more detailed study of a special case of Hong and Preston [19]. They have considered model selection in a more general framework, which allows general objective functions, including the GMM and GEL criterion functions. In addition, they allow multiplicity in the set of "best models" which could be mutually nonnested (see their Section 4.2.2). Their results indicate that model selection consistency can hold in the nested case but fail in the nonnested case. Regarding such opposite conclusions in these two cases, we have benefited from an anonymous referee on clarifying this point, who noted that consistent model selection has two meanings in [19]. The first meaning is that a consistent model selection procedure selects the set of "best" models w.p.a.1. The second meaning is that if there is multiplicity in the set of best models, a consistent model selection procedure should pick the most parsimonious model among the best models w.p.a.1. Our result in model selection consistency of BGMM is obtained in the nested case, since we have considered all 2^p coordinate subspaces of $\mathbb{R}^p$. Therefore, the oracle property of BGMM in Theorem 1 fulfills both two meanings of consistent model selection described in [19].

Although the nested case we have considered is not as general as Hong and Preston [19], we have allowed the dimension p to increase with n, which is new for BGMM model selection and also technically challenging. Because the number of candidate models is 2^p, which increases exponentially fast in p and hence in n, the previously studied pairwise model comparison using posterior odds or Bayes factors between one candidate model and the true model (such as Hong and Preston [19], Kim [28]) is insufficient to show the global model selection consistency. In addition to the increasing dimensionality and the more detailed study on the limiting distribution, we will also discuss below the asymptotic validity and interpretation of the BGMM quasi-posterior, which is new in the literature.

2.3. Asymptotic validity of the BGMM posterior

As shown in [27], the limited information likelihood we have used for BGMM provides a large sample approximation to the true likelihood function of θ given the moment restrictions $\mathrm{E}\, g(D,\theta) = 0$. One may ask about how well this approximation could be. For the validity of usual Bayesian inference, such as constructing the Bayesian credible sets, it is necessary and sufficient to impose Assumption 6 that $\boldsymbol{V}_n$ consistently estimates $\boldsymbol{V}$, i.e. $\boldsymbol{V}_n$ satisfies the generalized information equality as in [27,6]. However, due to the limited information contained in $\mathrm{E}\, g(D,\theta) = 0$, in general one cannot expect the LIL $q(\boldsymbol{D}|\theta)$ to coincide with the true likelihood function $p(\boldsymbol{D}|\theta)$. Instead the quasi-posterior $q(\theta|\boldsymbol{D})$ can be used to approximate the posterior of θ given some summary statistic from the sample. Let $\hat{\theta}$ be the minimizer of the GMM criterion function $\bar{g}(\boldsymbol{D},\theta)^\top \boldsymbol{V}^{-1}\bar{g}(\boldsymbol{D},\theta)$ over the full p-dimensional model space. So $\hat{\theta}$ is implicitly a statistic of the sample $\boldsymbol{D}$, and it does not depend on θ_0 and the unknown true model $\mathcal{M}_0$. Since the asymptotic center of the BGMM posterior is the first order approximation to the GMM estimator, one can expect that the LIL $q(\boldsymbol{D}|\theta)$ approximates the density $p(\hat{\theta}|\theta)$ of $\hat{\theta}$. Accordingly, the BGMM posterior $q(\theta|\boldsymbol{D})$ approximates the posterior $p(\theta|\hat{\theta})$ of θ given $\hat{\theta}$, at least asymptotically. In the following, we formalize this idea and show more general results under the model selection setup.

For two generic models M_1 and M_2, we define the Bayes factor based on $p(\hat{\theta}|\theta)$ as

$$\mathrm{BF}_{\hat{\theta}}[\mathcal{M}_1 : \mathcal{M}_2] = \frac{p(\hat{\theta}|\mathcal{M}_1)}{p(\hat{\theta}|\mathcal{M}_2)} = \frac{\int_{\Theta(\mathcal{M}_1)} p(\hat{\theta}|\theta)\pi(\theta|\mathcal{M}_1)\mathrm{d}\theta}{\int_{\Theta(\mathcal{M}_2)} p(\hat{\theta}|\theta)\pi(\theta|\mathcal{M}_2)\mathrm{d}\theta}.$$

For theory development, in this section we focus on the situation with a nonincreasing dimension p. We make the following extra assumption.

Assumption 9.

(i) $\dim(\theta) = p$ and $1 \le p \le \bar{p}$, for some large fixed integer $\bar{p}$.
(ii) $\min_{j\in\mathcal{M}_0} |\theta_{0,(j)}| \ge \underline{\theta}$ for some small constant $\underline{\theta} > 0$.
(iii) Let $\boldsymbol{V}(\theta) = \mathrm{Var}\{\bar{g}(D,\theta)\}$ and $\boldsymbol{G}(\theta) = \nabla_\theta \mathrm{E}\, \bar{g}(D,\theta)$. Then the elements of $\boldsymbol{V}(\theta)$ and $\boldsymbol{G}(\theta)$ are continuous functions of θ, and the eigenvalues of $\boldsymbol{G}(\theta)^\top\boldsymbol{G}(\theta)$ and $\boldsymbol{V}(\theta)$ are uniformly bounded below and above for all $\theta \in \Theta$.
(iv) For any two models $\mathcal{M}_1$ and $\mathcal{M}_2$, there exists a constant $r > 0$ such that $\pi(\mathcal{M}_2)/\pi(\mathcal{M}_1) \le r$.
(v) $\left\|\hat{\theta} - \bar{\theta}\right\| = O_p(1/n)$, where $\hat{\theta}$ is the GMM estimator on the full model space, and $\bar{\theta} = \theta_0 - (\boldsymbol{G}^\top\boldsymbol{V}^{-1}\boldsymbol{G})^{-1}\boldsymbol{G}^\top\boldsymbol{V}^{-1}\bar{g}(\boldsymbol{D},\theta_0)$.

The strengthened beta-min condition in (ii) is to emphasize the difference between the models that make the type I error and the type II error. According to theorems we are going to present below, the models in the former group have an exponentially small $\mathrm{BF}_{\mathrm{q}}[\mathcal{M} : \mathcal{M}_0]$, while the models in the latter group have a polynomially small $\mathrm{BF}_{\mathrm{q}}[\mathcal{M} : \mathcal{M}_0]$. This is also the essential behavior from the Bayesian hypothesis test, which favors the true alternative hypothesis more. We will show that similar behavior is also shared by $\mathrm{BF}_{\hat{\theta}}[\mathcal{M} : \mathcal{M}_0]$, and hereby establish a correspondence between the BGMM method and the exact Bayesian method given $\hat{\theta}$.

Part (iii) assumes the continuity of the matrices in θ and also the uniform bound for eigenvalues. This is a mild assumption given the compactness of Θ. Part (iv) has strengthened Assumption 8 and required that no model should be assigned extremely large or small prior. Part (v) is about the high order approximation of $\bar{\theta}$ to the GMM estimator $\hat{\theta}$ on the full model space, similar to the discussion in Remark 2, which usually holds when the moment condition $g(D,\theta)$ is continuously differentiable in θ [42] and hence may not apply to the example of quantile regression.

Let $\boldsymbol{F}(\theta)$ be a $p\times p$ matrix such that $\boldsymbol{F}(\theta)^\top\boldsymbol{F}(\theta)=\boldsymbol{G}(\theta)^\top\boldsymbol{V}(\theta)^{-1}\boldsymbol{G}(\theta)$. Define $Z=\sqrt{n}\boldsymbol{F}(\theta)(\hat\theta-\theta)$. Then Z is asymptotically p-dimensional standard normal if the true parameter is $\theta_0=\theta$. We impose the following high level assumption on the difference between the exact density function $p_Z(z)$ of Z and the normal density.

Assumption 10 (*Uniform Bound*)**.** As $n\to\infty$,

$$\sup_{\theta\in\Theta}\sup_{z}\left(1+\|z\|^{p+1}\right)\left|p_Z(z|\theta)-\phi(z;0,\boldsymbol{I}_p)\right|=\tau_n,$$

where $\tau_n=o(1)$ does not depend on z and θ, and $\boldsymbol{I}_p$ is the $p\times p$ identity matrix.

Assumption 10 claims that the difference between the density of the normalized GMM estimator Z and its asymptotic limit of normal density can be uniformly bounded by an integrable function $c(\|z\|)=1/(1+\|z\|^{p+1})$, and the uniformity is for both the value of z and the parameter θ in the compact space Θ. This is a high level condition that originates from the Condition E in [54]. We do not intend to give a full proof of it under low level assumptions, but we explain why it is a reasonable assumption below.

Consider the case where the $(p+1)$th moment of $g(D,\theta)$ exists. To show Assumption 10, we proceed in several steps. First, under similar regularity conditions that make Assumption 9(v) hold, one can see that for a fixed θ, the density of Z is asymptotically uniformly close to the density of the normalized first order approximation $\bar Z=\sqrt{n}\boldsymbol{F}(\theta)(\bar\theta-\theta)$, up to the order $O(1/\sqrt{n})$, where $\bar\theta$ is defined in Theorem 1(ii). See [31,32] for the formal proofs of a general class of nonlinear estimators, which can also be applied to the GMM estimator. Second, due to the sample average form of $\bar\theta$ and hence $\bar Z$, one can use Proposition 1 in [54] and take $c(x)=1/(1+x^{p+1})$. This proposition provides a bound for the difference between the density of $\bar Z$ and its limiting normal density, which holds uniformly for all $\theta\in\Theta$. Its proof involves the techniques in Chapter 19 of Bhattacharya and Ranga [3] about the uniform convergence of continuous characteristic functions in the compact set Θ. Third, one can show that in Proposition 1 of Yuan and Clarke [54], the summation of the Edgeworth series beyond the leading normal density term has the order $o_p(1)$. This is due to the finite moments of $g(D,\theta)$ up to the $(p+1)$th order, as well as the boundedness of multivariate Hermite polynomials. Finally we combine all these pieces and conclude that the uniform deviation in Assumption 10 holds with some $\tau_n=o(1)$.

The next theorem provides a comparison between the convergence rates for the Bayes Factors $\mathrm{BF}_{\hat\theta}[\mathcal{M}:\mathcal{M}_0]$ from the likelihood given the statistic $\hat\theta$ with $\mathrm{BF}_{\mathrm{q}}[\mathcal{M}:\mathcal{M}_0]$ from the BGMM method.

Theorem 2 (*Equivalence of Bayes Factors*)**.** *Suppose Assumptions 1–10 hold, and the true model size is $|\mathcal{M}_0|=k_0$. Then under the same prior $\pi(\theta|\mathcal{M})$ and $\pi(\mathcal{M})$, w.p.a.1 as $n\to\infty$,*

(i) *For any model $\mathcal{M}$ with $\mathcal{M}\supseteq\mathcal{M}_0$,*

$$\frac{\mathrm{BF}_{\mathrm{q}}[\mathcal{M}:\mathcal{M}_0]}{\mathrm{BF}_{\hat\theta}[\mathcal{M}:\mathcal{M}_0]}\to 1;$$

$$\mathrm{BF}_{\mathrm{q}}[\mathcal{M}:\mathcal{M}_0]\asymp\mathrm{BF}_{\hat\theta}[\mathcal{M}:\mathcal{M}_0]\asymp n^{-\frac{|\mathcal{M}|-k_0}{2}}\succeq n^{-\frac{p-k_0}{2}};$$

(ii) *For any model with $\mathcal{M}$ with $\mathcal{M}_0\setminus\mathcal{M}\neq\emptyset$, there exists a constant $C>0$, such that*

$$\mathrm{BF}_{\mathrm{q}}[\mathcal{M}:\mathcal{M}_0]\le\exp\left(-Cn\underline{\theta}^2\right)\prec n^{-\frac{p-k_0+1}{2}};$$

$$\mathrm{BF}_{\hat\theta}[\mathcal{M}:\mathcal{M}_0]\le\exp\left(-Cn\underline{\theta}^2\right)\vee\tau_n n^{-\frac{p-k_0+1}{2}}\prec n^{-\frac{p-k_0+1}{2}}.$$

Theorem 2 compares the Bayes factors from BGMM and $p(\hat\theta|\theta)$, for the models that make a type I error (Part ii) and a type II error (Part i). The theorem has at least two direct implications. First, for the models that make a type II error (including more components of θ than necessary), the Bayes factors are asymptotically equal, and both decrease polynomially in the sample size n. The polynomial index reflects the difference in dimensions between $\mathcal{M}$ and $\mathcal{M}_0$. Second, for the models that make a type I error (missing at least one nonzero component in θ_0), the Bayes factor from BGMM decreases exponentially fast in n. For the Bayes factor from $p(\hat\theta|\theta)$, we have obtained an upper bound for its rate, which also depends on the rate τ_n in Assumption 10 besides the usual exponential rate. Because $\tau_n=o(1)$ by Assumption 10, we can see clearly that there exists at least a $n^{-1/2}$ gap between the convergence rates of Bayes factors for the models with type I and type II errors. The threshold rate is $n^{-(p-k_0)/2}$, which depends on the unknown dimension k_0 of the true model $\mathcal{M}_0$. In general, the posterior probabilities of the models with type I errors converge faster to zero than the posterior of the models with type II errors.

This extra part $\tau_n n^{-(p-k_0+1)/2}$ for the Bayes factor in (ii) arises mainly technically from our Assumption 10. Usually, the order $\tau_n=o(1)$ in Assumption 10 is tight and cannot be improved. However, we conjecture that it could be removed by making stronger assumptions on the density function $p(\hat\theta|\theta)$, or the density $p_Z(z|\theta)$ of the normalized statistic Z. For example, one can assume that $p_Z\left(\sqrt{n}\boldsymbol{F}(\theta)(\hat\theta-\theta)\big|\theta\right)$ decreases exponentially fast in n as θ moves away from the true parameter θ_0. However, we note that usually it is difficult to verify such assumptions because $\hat\theta$ does not have an explicit

density, except for a few special cases where $\hat{\theta}$ comes from the exponential family. We also note that such compromised rate also shows up in Lemma 1 of Marin et al. [40], where they studied the convergence rates of Bayes factors given a general statistic. Although typically one cannot obtain the exact form of the density $p(\hat{\theta}|\theta)$ and its posterior $p(\theta|\hat{\theta})$, Theorem 2 provides some evidence that in the asymptotic sense, the Bayes factors from BGMM behave very similarly to the Bayes factors from $p(\hat{\theta}|\theta)$, indicating the validity of using $\mathrm{BF}_{\mathrm{q}}[\mathcal{M} : \mathcal{M}_0]$ for model selection purpose.

Remark 4. In principle, Theorem 2 provides a guideline to interpret the BGMM posterior probabilities of different models. For simplicity, suppose that all models receive the uniform prior $\pi(\mathcal{M}) \propto 1$. Then since $q(\mathcal{M}_0|\boldsymbol{D}) \to 1$, the posterior $q(\mathcal{M}|\boldsymbol{D})$ is roughly the same as $\mathrm{BF}_{\mathrm{q}}[\mathcal{M} : \mathcal{M}_0]$. Because of the gap between the polynomial rate in (i) and the exponential rate in (ii), we can choose any rate in between as a threshold, for example $e^{-\sqrt{n}}$. If a model $\mathcal{M}$ has $q(\mathcal{M}|\boldsymbol{D}) \geq e^{-\sqrt{n}}$, then we can approximately regard $q(\mathcal{M}|\boldsymbol{D})$ as the true posterior probability $p(\mathcal{M}|\hat{\theta})$ and consider $\mathcal{M}$ as a model with nonnegligible posterior. This fits well with the common practice that we rank the models according to their posterior probabilities and only study the models on top of the list.

Based on Theorem 2, we can further show that the BGMM posterior $q(\theta|\boldsymbol{D})$ and the exact posterior $p(\theta|\hat{\theta})$ are close in the total variation distance asymptotically.

Theorem 3. *Suppose Assumptions 1–10 hold. Let the full model be $\mathcal{M}_{\mathrm{full}}$. Then under the same prior $\pi(\theta|\mathcal{M})$ and $\pi(\mathcal{M})$, w.p.a.1 as $n \to \infty$,*

(i) *(Model Selection Convergence Rate) If $\mathcal{M}_0 \neq \mathcal{M}_{\mathrm{full}}$, then*

$$\frac{q(\mathcal{M} : \mathcal{M} \neq \mathcal{M}_0|\boldsymbol{D})}{p(\mathcal{M} : \mathcal{M} \neq \mathcal{M}_0|\hat{\theta})} \to 1;$$

$$q(\mathcal{M} : \mathcal{M} \neq \mathcal{M}_0|\boldsymbol{D}) \asymp p(\mathcal{M} : \mathcal{M} \neq \mathcal{M}_0|\hat{\theta}) \asymp n^{-\frac{1}{2}} \to 0;$$

If $\mathcal{M}_0 = \mathcal{M}_{\mathrm{full}}$, then for some constant $C > 0$,

$$q(\mathcal{M} : \mathcal{M} \neq \mathcal{M}_0|\boldsymbol{D}) \leq \exp\left(-Cn\underline{\theta}^2\right) \to 0;$$

$$p(\mathcal{M} : \mathcal{M} \neq \mathcal{M}_0|\hat{\theta}) \leq \exp\left(-Cn\underline{\theta}^2\right) \vee \tau_n n^{-\frac{p-k_0+1}{2}} \to 0.$$

(ii) *(Asymptotic Posterior Validity)*

$$\sup_{A \subseteq \Theta} \left| \int_A q(\theta|\boldsymbol{D}) \mathrm{d}\theta - \int_A p(\theta|\hat{\theta}) \mathrm{d}\theta \right| \to 0.$$

Part (i) of the theorem is a direct corollary from Theorem 2. It implies that the posterior probability of the true model $\mathcal{M}_0$ converges to 1 at exactly the same rate using either the BGMM or $p(\hat{\theta}|\theta)$, when the true model is a strict submodel of the full model. When the true model is exactly the same as the full model, we have only upper bounds for the model selection convergence rates, as they usually decrease exponentially fast, but again the rate is compromised by τ_n from Assumption 10 when we consider $p(\mathcal{M} : \mathcal{M} \neq \mathcal{M}_0|\hat{\theta})$. In either scenario, we have the global model selection consistency for both the BGMM posterior and the posterior given $\hat{\theta}$.

Part (ii) gives the asymptotic validity of the BGMM posterior, in the sense that it provides the same asymptotic inference as the exact posterior of θ given the statistic $\hat{\theta}$. It has the immediate implication that the posterior credible sets for the parameters constructed from the BGMM posterior are asymptotically valid. It is worth noting that the conclusion of (ii) is only related to the global model selection consistency for both the BGMM posterior and the posterior of $p(\theta|\hat{\theta})$, and does not depend on the exact convergence rates of model selection in Part (i). In fact, Part (ii) also holds for general non-model selection prior $\pi(\theta)$ as long as it has a bounded continuous density on Θ. This can be obtained from combining Theorem 1 in [6] and Theorem 2 in [54] (where $T_n = \hat{\theta}$), under Condition E in [54]. Our proof of Theorem 3 follows a similar route by using Assumption 10, but has accommodated the nature of model selection priors $\pi(\theta, \mathcal{M}) = \pi(\theta|\mathcal{M})\pi(\mathcal{M})$.

Remark 5. We have discussed the asymptotic closeness of the BGMM posterior to the posterior given the GMM estimator $\hat{\theta}$. One can further explore the higher order asymptotics of $q(\theta|\boldsymbol{D})$ and $p(\theta|\hat{\theta})$, for example expanding both posterior densities as Edgeworth series of the asymptotic pivotal quantity $\sqrt{n}\boldsymbol{F}(\theta)(\theta - \hat{\theta})$. In this sense, our result in Part (ii) of Theorem 3 only captures the leading order closeness from $q(\theta|\boldsymbol{D})$ to $p(\theta|\hat{\theta})$. However, we conjecture that in general the higher order terms of $q(\theta|\boldsymbol{D})$ and $p(\theta|\hat{\theta})$ do not match with each other, since the LIL takes a quadratic form of the moment conditions while the true density of $\hat{\theta}$ depends on other features of $P_{\boldsymbol{D}}$, such as the high order moments. Similar work in this direction includes Fang and Mukerjee [12], where they have shown by a simple example of sample mean that the Edgeworth expansions from the empirical likelihood and the density of the sample average do not agree in high order terms.

3. Numerical study

3.1. Algorithm

Because the LIL (2) allows any form of moment function $g(D, \theta)$, usually one cannot derive an analytical close form for the BGMM model posterior $q(\mathcal{M}|\boldsymbol{D})$. Therefore, we adopt a reversible jump MCMC algorithm with Metropolis moves both between models and within a model to explore the joint posterior of $q(\theta, \mathcal{M}|\boldsymbol{D})$, similar in spirit to the MCMC algorithm for the Gibbs posterior model selection [5,23], and also the PAC-Bayesian model selection [1,16]. In the ith iteration, the between-model steps either add a new component to the nonzero part of $\theta^{(i)}$, or remove an existing component in the nonzero part of $\theta^{(i)}$, each with probability 0.5. When we add a new component, the parameter value for this new component is sampled from $\mathcal{N}(0, \sigma^2_{\text{add}})$, while the values of the existing components in $\theta^{(i)}$ are retained. Both the "add" and the "remove" operations will be accepted or rejected with a probability based on the ratio of the posteriors evaluated at the new proposed parameter and the current parameter. This between-model step is then followed by a within-model step, in which we draw a new parameter value in the same model as $\theta^{(i)}$ from a proposal distribution. In practice, to efficiently explore each model space, we use a normal distribution as a proposal distribution, with mean zero and a properly chosen variance $c \cdot \boldsymbol{\Xi}_{\mathcal{M}}$. Here $\boldsymbol{\Xi}_{\mathcal{M}}$ is the submatrix of $\boldsymbol{\Xi}$ with rows and columns corresponding to the model $\mathcal{M}$, and $\boldsymbol{\Xi}$ is an estimated covariance matrix for the GMM estimator $\hat{\theta}$, which can be obtained numerically by inverting the Hessian matrix at the preliminary one-step GMM estimator $\tilde{\theta}$ on the full model space. We set $c = 2.4^2$ as suggested in [14] to achieve the ideal acceptance rate for within-model Metropolis moves. We also run pilot chains to tune the value of σ_{add} for better mixing of the Markov chain. As a result, the Markov chain consists of $\theta^{(i)}$ drawn from the full BGMM posterior across different model spaces.

3.2. Example: correlated binary responses

The conditional mean $\mu_{ij}(\theta) = \mathrm{E}(Y_{ij}|X_{ij})$ of the longitudinal binary response Y_{ij} is given by

$$\ln \frac{\mu_{ij}(\theta)}{1-\mu_{ij}(\theta)} = X_{ij}^{\top}\theta, \tag{10}$$

where $i = 1, \ldots, n$ and $j = 1, \ldots, s$. In the following simulations, we first fix the sample size $n = 400$ and the cluster size $s = 10$, in order to compare with the similar simulation setups in [51]. For $X_{ij} = (X_{ij1}, \ldots, X_{ijp})^{\top}$, we consider two situations with $p = 50$ and $p = 100$. $X_{ij1}, \ldots, X_{ijp}$ are generated independently from a uniform distribution on $[-1, 1]$. We also consider two sets of true parameter values,

$$\theta_0 = (1.5, -1.5, 1, -1, 0.5, -0.5, 0, 0, \ldots, 0)$$
$$\theta_0 = (1.5, -1.5, 1.5, -1.5, 1, -1, 1, -1, 0.5, -0.5, 0.5, -0.5, 0, 0, \ldots, 0)$$

with the number of nonzero components $k_0 = 6$ and $k_0 = 12$ respectively. Note that θ_0 contains weak signals 0.5 and -0.5 and more nonzero components in the second setting. Similar to Wang et al. [51] and Cho and Qu [8], we use the R package `mvtBinaryEP` to generate the correlated binary responses $(Y_{i1}, \ldots, Y_{is})^{\top}$ for each $i = 1, \ldots, n$ with an exchangeable correlation structure with correlation coefficient $\rho = 0.3$.

Since this is a special case of the first motivating example in Section 1.3, we examine the performance of BGMM using the moment function $g(D, \theta)$ defined in (5). We compare the BGMM method to the frequentist penalized GEE method (PGEE) proposed by Wang et al. [51] which is used to fit high dimensional longitudinal data. Let $\theta_{(k)}$ be the kth component of θ. The PGEE solves a similar estimating equation to (4)

$$n^{-1}\sum_{i=1}^{n} \frac{\partial \mu_i(\theta)^{\top}}{\partial \theta} \boldsymbol{S}_i^{-1}(Y_i - \mu_i(\theta)) - P_{\lambda_n}(\theta) = 0,$$

with an additional SCAD penalty $P_\lambda(\theta) = (P_\lambda(\theta_{(1)}), \ldots, P_\lambda(\theta_{(p)}))^{\top}$ and for $k = 1, \ldots, p$,

$$P_\lambda(\theta_{(k)}) = \lambda_n \left\{ 1(\theta_{(k)} \le \lambda_n) + 1\left(\lambda_n < \theta_{(k)} \le a\lambda_n\right) \frac{a\lambda_n - \theta_{(k)}}{(a-1)\lambda_n} \right\}.$$

The PGEE can be solved by an iterative Newton–Raphson algorithm as described in [51]. In our simulations, we perform in the same way as Cho and Qu [8], fix $a = 3.7$ and truncate the estimated coefficients to zero if $|\hat{\theta}_{(k)}| \le 10^{-3}$ $(k = 1, \ldots, p)$. λ_n is selected from the grid set $\{0.01, 0.02, \ldots, 0.2\}$ by 5-fold cross validation. We use an estimated correlation matrix for $\boldsymbol{R}$ based on the sample, instead of varying the correlation structures in [51]. In fact, the finite sample estimates of $\boldsymbol{R}$ are quite precise for the true $\boldsymbol{R}$ in our $p < n$ case.

For the BGMM method, the prior on θ given a model $\mathcal{M}$ is the product of independent normal densities

$$\pi(\theta|\mathcal{M}) = \prod_{j\in\mathcal{M}} \frac{1}{\sqrt{2\pi}\sigma_\theta} e^{-\frac{\theta_{(j)}^2}{2\sigma_\theta^2}}, \tag{11}$$

Table 1
Comparison of BGMM with PGEE for Correlated Binary Responses. k_0 is the number of nonzero components in the true parameter θ_0. p is the dimension of θ_0. n is the sample size. Standard errors are shown in the parentheses. EX: exact selection; UN: under selection; OV: over selection; TP: true positives; FP: False positives; MSE: mean square error of θ; pMSE: prediction mean square error of $\mu_{ij}(\theta)$.

	EX	UN	OV	TP	FP	MSE ($\times 10^{-3}$)	pMSE ($\times 10^{-4}$)
$k_0 = 6, p = 50, n = 400$							
Naive	0	0	1	6	44	24.81 (0.57)	21.77 (0.43)
Oracle	1	0	0	6	0	4.15 (0.35)	2.66 (0.16)
PGEE	0.33	0	0.67	6	5.18	7.02 (0.57)	5.05 (0.41)
BGMM1	0.80	0	0.20	6	0.21	5.58 (0.40)	3.85 (0.23)
BGMM2	0.98	0	0.02	6	0.02	5.06 (0.38)	3.33 (0.20)
$k_0 = 12, p = 50, n = 400$							
Naive	0	0	1	12	38	13.09 (0.35)	20.33 (0.40)
Oracle	1	0	0	12	0	5.53 (0.49)	4.82 (0.19)
PGEE	0.33	0	0.67	12	4.08	6.13 (0.58)	6.61 (0.38)
BGMM1	0.88	0	0.12	12	0.13	5.13 (0.29)	5.91 (0.23)
BGMM2	0.94	0	0.06	12	0.07	4.89 (0.27)	5.63 (0.22)
$k_0 = 6, p = 100, n = 400$							
Naive	0	0	1	6	94	49.28 (0.67)	42.53 (0.61)
Oracle	1	0	0	6	0	6.72 (0.68)	3.09 (0.17)
PGEE	0.28	0	0.72	6	5.27	9.16 (1.37)	5.21 (0.51)
BGMM1	0.55	0	0.45	6	0.61	8.39 (0.51)	5.32 (0.27)
BGMM2	0.93	0	0.07	6	0.07	7.32 (0.51)	4.05 (0.21)
$k_0 = 12, p = 100, n = 400$							
Naive	0	0	1	12	88	27.39 (0.55)	42.74 (0.63)
Oracle	1	0	0	12	0	8.09 (0.64)	5.85 (0.22)
PGEE	0.16	0	0.84	12	6.35	10.20 (1.40)	8.62 (0.63)
BGMM1	0.59	0	0.41	12	0.55	8.24 (0.52)	8.82 (0.36)
BGMM2	0.90	0	0.10	12	0.10	7.76 (0.52)	7.68 (0.29)

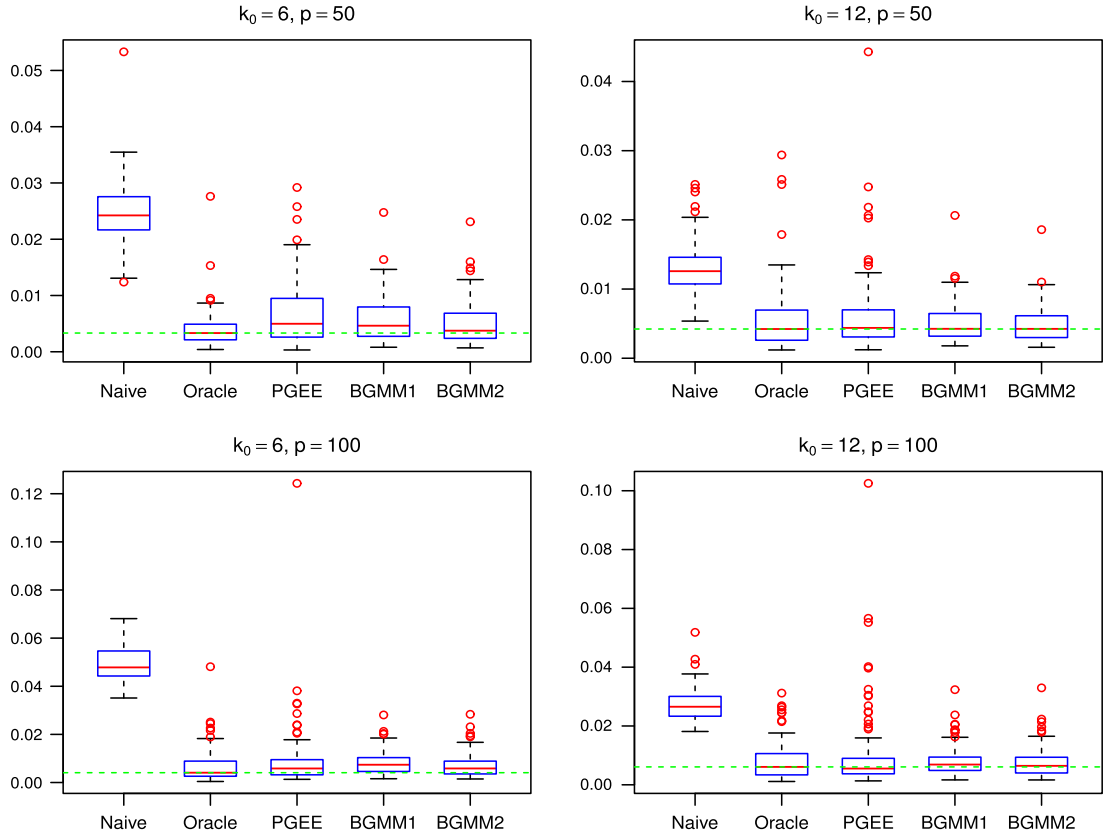

Fig. 1. Boxplots for the MSE of θ over 100 simulated datasets.

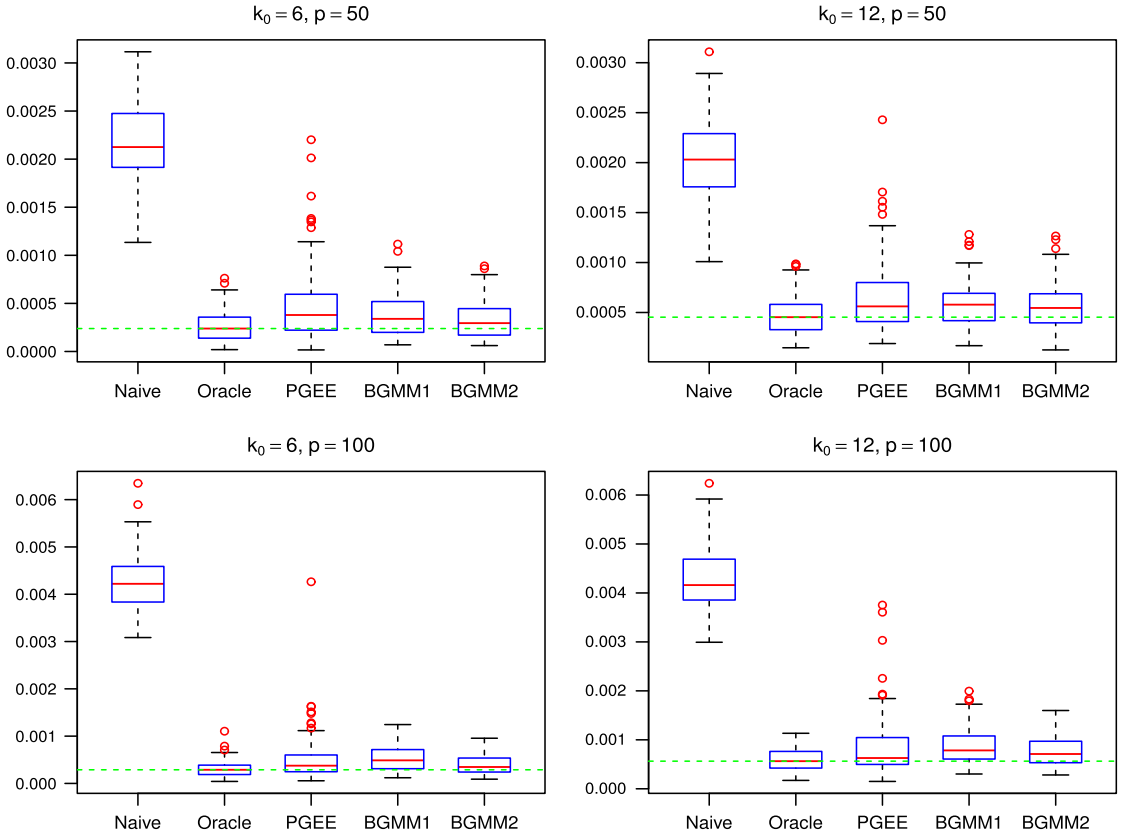

Fig. 2. Boxplots for the MSE of $\mu_{ij}(\theta)$ over 100 simulated datasets.

where we choose $\sigma_\theta = 10$ for a large prior spread. Note that although theoretically this prior is not truncated on a compact set Θ as in Assumption 7, in practice this has no influence on in our experimental results.

The prior on the model $\mathcal{M}$ is specified as follows:

$$\pi(\mathcal{M}) \propto \nu^{|\mathcal{M}|}(1-\nu)^{p-|\mathcal{M}|}, \tag{12}$$

which means that each component of θ independently enters the model $\mathcal{M}$ with probability $\nu \in (0, 1)$. When $\nu = 0.5$, this is the same as the uniform prior over all 2^p models. When ν moves towards zero, the prior gradually induces more sparsity on θ and favors more parsimonious models, which imposes a further penalization on the model size besides the incorporated BIC-type penalization in BGMM. It can be verified (see Section 4 of the supplementary material, Appendix A) that the prior (12) satisfies Assumption 8 when $\nu \in (0, 1)$ is either fixed or $\nu = n^{-c}$ for some $c > 0$, and it satisfies Assumption 9(iv) if ν is fixed.

In our simulation, for each simulated datasets, we run one single Markov chain with the length 3×10^4, and drop the first 10^4 iterations as burnin. We consider two choices of the tuning parameter ν in (12). In the first case (referred to as BGMM1), we fix $\nu = 0.5$ throughout the Markov chain for the next 2×10^4 iterations. In the second case (referred to as BGMM2), we adopt a two-step tuning strategy in an effort to make the value of ν more adaptive to the sparsity level of the true model. We estimate the posterior average model size $\hat{\mathrm{E}}_{\mathcal{M}|\boldsymbol{D}}|\mathcal{M}|$ using the first 10^4 MCMC runs after the burnin, and then reset $\nu = \hat{\mathrm{E}}_{\mathcal{M}|\boldsymbol{D}}|\mathcal{M}|/p$ for the next 10^4 MCMC runs. Finally for both chains of BGMM1 and BGMM2, we keep $N = 10^3$ MCMC samples from the last 10^4 runs for every 10 iteration. The variance of proposal normal density described in Section 3.1 is fixed at $\sigma_{\text{add}} = 0.2$. Our experiments with other values of σ_{add} (such as 0.05, 0.1, 0.15, 0.25) show that $\sigma_{\text{add}} = 0.2$ is sufficient for exploring the full posterior of θ, and the MCMC results such as the MAP models and posterior distributions of parameters are not sensitive to different values of σ_{add}.

As a benchmark, the PGEE method and the BGMM method are compared together with the naive method and the oracle method. The naive method estimates θ by usual GEE without doing model selection, while for the oracle method, the true model is pretendedly known and θ is estimated only on the nonzero components. We apply each method to the same dataset and repeat this process for 100 Monte Carlo replications. We compare three aspects of these methods: the model selection, the parameter estimation, and the prediction.

To evaluate the model selection performance, we consider the model selected by PGEE and the MAP model from BGMM, and report the proportion of times the method exact selecting (EX), underselecting (UN) and overselecting (OV) the nonzero

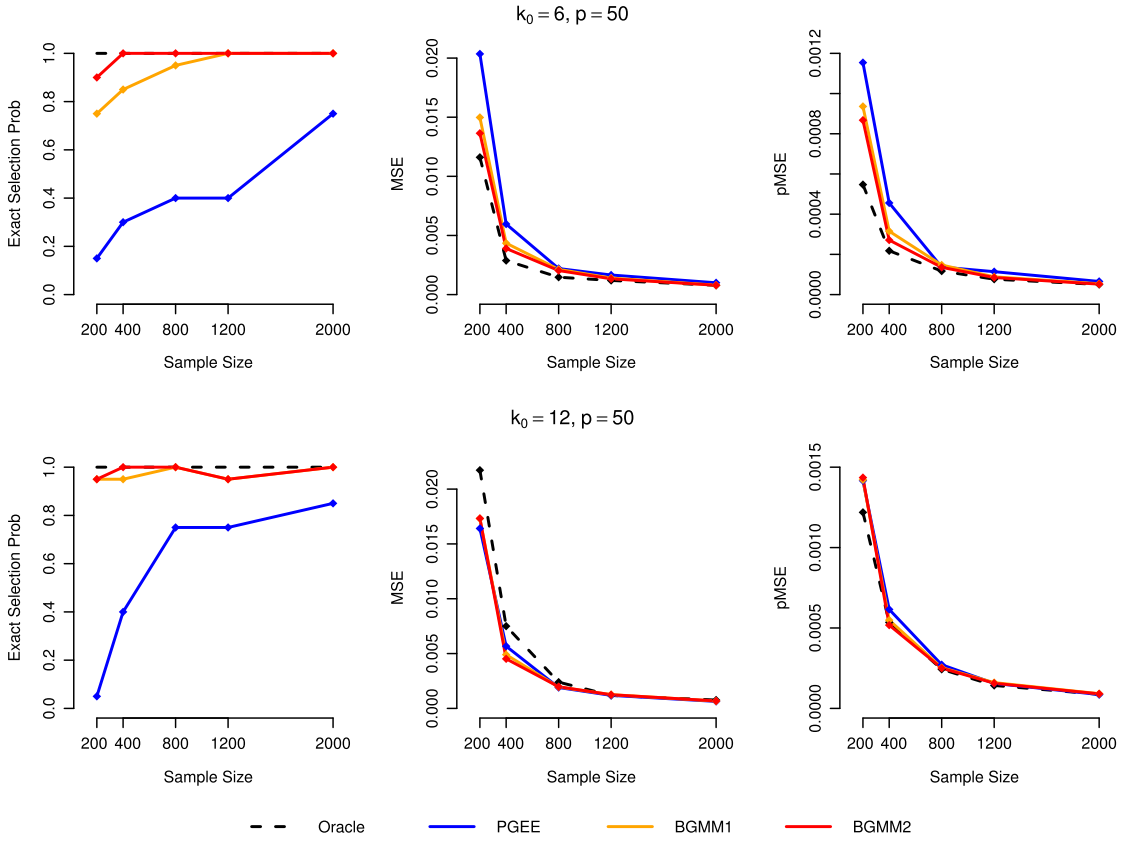

Fig. 3. Exact selection probability, MSE, and prediction MSE for $p = 50$ over 20 simulated datasets.

components of θ_0. We also report the true positives (TP, the average number of correctly selected nonzero components in θ_0), and the false positives (FP, the average number of selected nonzero components that are actually zero in θ_0).

For the estimation accuracy, similar to Cho and Qu [8], we report the estimated mean square error (MSE) $\sum_{m=1}^{100} \|\hat{\theta}_m - \theta_0\|^2/(100k_0)$, where $\hat{\theta}_m$ is the mth estimated parameter vector. This MSE is calculated for the naive method, the oracle method, the PGEE method, and the posterior mean of θ from the BGMM method.

For the prediction accuracy, we calculate the average MSE for the conditional mean μ_{ij} (denoted by pMSE), defined as $\sum_{i=1}^{n}\sum_{j=1}^{s}\left(\mu_{ij}(\hat{\theta}) - \mu_{ij}(\theta_0)\right)^2/(ns)$ for the naive, the oracle, and the PGEE method. For the BGMM method, we use the pMSE averaged over the posterior sample $\sum_{i=1}^{n}\sum_{j=1}^{s}\sum_{k=1}^{N}\left(\mu_{ij}(\theta^{(k)}) - \mu_{ij}(\theta_0)\right)^2/(Nns)$, where $\theta^{(1)}, \ldots, \theta^{(N)}$ are the MCMC draws of θ.

As Table 1 indicates, both the frequentist PGEE method and our BGMM method have always successfully identified the nonzero components of θ_0 with no underselection. However, the PGEE performs much more conservatively and has a serious overselection problem in all the simulations settings, which is consistent with the findings in [8]. It selects the true model for 33% of all time when $p = 50$, and only 16% of all time when $p = 100$ and $k_0 = 12$. Meanwhile PGEE overselects about 4–6 extra redundant variables on average. In contrast, the BGMM MAP models have much higher probability of exactly selecting the true model, and have much smaller false positives. We also note that the extra two-step tuning of ν in BGMM2 has brought significant advantage over the uniform model prior with fixed $\nu = 0.5$ in BGMM1. When $p = 100$, the performance of BGMM1 deteriorates as the probability of exact selection drops to about 50%, but BGMM2 still maintains a high accuracy with over 90% of exact model selection. This is because that in BGMM2, the first step of 10^4 runs has consistently estimated the true proportion of nonzero components in θ, and then the second step of 10^4 runs can learn the sparsity of the model space better with ν roughly equal to the true average marginal inclusion probability.

For the estimation and prediction, it is clear that the naive GEE estimator with no model selection performs poorly in MSE and pMSE compared to the oracle estimator. Figs. 1 and 2 show that the MSE and pMSE for the BGMM method are comparable to those from the oracle and the PGEE method, as their boxplots largely overlap with each other. Also it seems that BGMM tends to have smaller variation across difference simulations than PGEE. The averaged levels of three MSEs from both BGMM methods are also slightly smaller than those from the PGEE estimator in most of the cases (Table 1), and they are all close to the MSE and pMSE from the oracle estimator. Overall, BGMM2 seems to be the best of all these methods

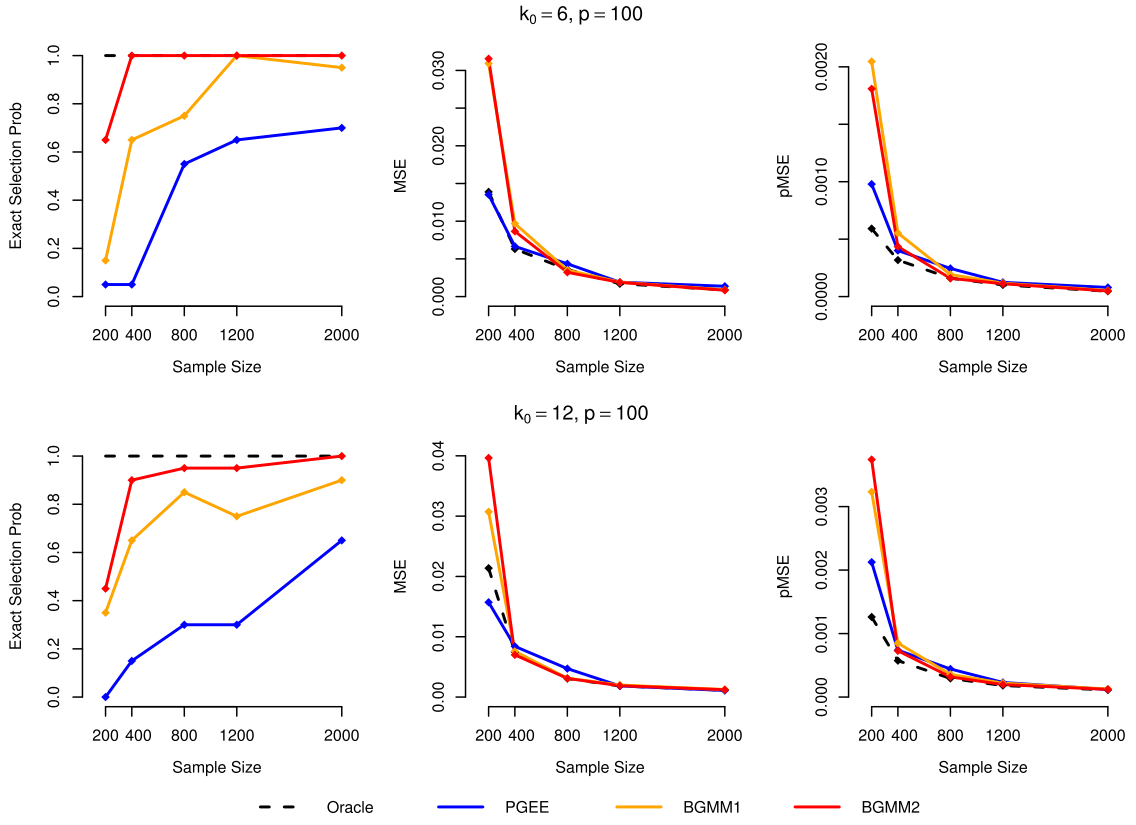

Fig. 4. Exact selection probability, MSE, and prediction MSE for $p = 100$ over 20 simulated datasets.

besides the oracle. This has partly supported our theoretical results about the oracle properties of the BGMM method, in the sense that the posterior variance of BGMM is asymptotically the same as the variance of the oracle GMM estimator.

Finally, we vary the sample size n among 200, 400, 800, 1200, 2000 and compare the performance of PGEE, BGMM1 and BGMM2 for the model (10) averaged over 20 simulated datasets. Figs. 3 and 4 plot their exact model selection probabilities (the same as the EX in Table 1), MSEs and pMSEs. Overall, BGMM2 has the best performance of exact model selection, and BGMM1 tends to perform better as n increases. All three methods have poor model selection accuracy for $p = 100$, $n = 200$ due to the relative high dimension and the small sample size. As the sample size n increases, the differences between their MSEs and pMSEs become negligible, as they all perform similarly to the oracle estimator.

4. Discussions

In this paper, we have studied some theoretical properties and applications of a Bayesian moment based model selection method. As we have commented, this method combines advantages of a Bayesian approach, such as the expressiveness of the posterior distribution and convenient MCMC algorithms for computation, with the model robustness of the moment based methods. We have formulated and proved the Bayesian oracle property of the proposed model selection method, which guarantees efficient posterior inference as if we knew which variables are truly relevant. We have studied the meaning of the quasi-posterior probabilities used in BGMM, which can be interpreted as the leading order large sample approximation to the true posterior probabilities conditional on the observed GMM estimator. The empirical performance of BGMM has been demonstrated by numerical experiments.

We have only considered quasi-posterior constructed from the GMM based quasi-likelihood function. Many other alternatives, such as EL, GEL, and ETEL, can be formulated under a similar Bayesian framework, with possible interpretations of the induced quasi-Bayesian posterior. See for example, [6,33,46], etc. We conjecture that similar Bayesian asymptotic properties for model selection can be derived for these quasi-likelihoods.

Acknowledgments

We thank the Associate Editor and the two anonymous Referees for their helpful suggestions on improving the paper.

Appendix A. Supplementary material

Supplementary material related to this article can be found online at http://dx.doi.org/10.1016/j.jmva.2015.12.009.

References

[1] P. Alquier, G. Biau, Sparse single-index model, J. Mach. Learn. Res. 14 (2013) 243–280.
[2] A. Belloni, V. Chernozhukov, On the computational complexity of MCMC-based estimators in large samples, Ann. Statist. 37 (2009) 2011–2055.
[3] R.N. Bhattacharya, R. Ranga Rao, Normal Approximation and Asymptotic Expansions, Wiley, New York, 1976, Reprinted by Robert E. Krieger, Melbourne, Florida, 1986.
[4] M. Caner, H.H. Zhang, Adaptive elastic net GMM estimator, J. Bus. Econom. Statist. 32 (2013) 30–47.
[5] K. Chen, W. Jiang, M. Tanner, A note on some algorithms for the Gibbs posterior, Statist. Probab. Lett. 80 (2010) 1234–1241.
[6] V. Chernozhukov, H. Hong, An MCMC approach to classical estimation, J. Econometrics 115 (2003) 293–346.
[7] H. Chipman, E.I. George, R.E. McCulloch, The practical implementation of Bayesian model selection, in: Model Selection, in: IMS Lecture Notes—Monograph Series, vol. 38, 2001, pp. 65–116.
[8] H. Cho, A. Qu, Model selection for correlated data with diverging number of parameters, Statist. Sinica 23 (2013) 901–927.
[9] P. Dellaportas, J.J. Forster, I. Ntzoufras, On Bayesian model and variable selection using MCMC, Stat. Comput. 12 (2002) 27–36.
[10] M. Drton, M.D. Perlman, Model selection for Gaussian concentration graphs, Biometrika 91 (2004) 591–602.
[11] J. Fan, R. Li, Variable selection via nonconcave penalized likelihood and its oracle properties, J. Amer. Statist. Assoc. 96 (2001) 1348–1360.
[12] K.T. Fang, R. Mukerjee, Empirical-type likelihoods allowing posterior credible sets with frequentist validity: Higher-order asymptotics, Biometrika 93 (2006) 723–733.
[13] J.P. Florens, A. Simoni, Gaussian processes and Bayesian moment estimation, Manuscript, 2012.
[14] A. Gelman, J. Carlin, H. Stern, D. Dunson, A. Vehtari, D. Rubin, Bayesian Data Analysis, third ed., Chapman and Hall/CRC, 2013.
[15] P. Green, Reversible jump Markov chain Monte Carlo, Biometrika 82 (1995) 711–732.
[16] B. Guedj, P. Alquier, PAC-Bayesian estimation and prediction in sparse additive models, Electron. J. Stat. 7 (2013) 264–291.
[17] L.P. Hansen, Large sample properties of generalized method of moments estimators, Econometrica 50 (1982) 1029–1054.
[18] L.P. Hansen, J. Heaton, A. Yaron, Finite-sample properties of some alternative gmm estimators, J. Bus. Econom. Statist. 14 (1996) 262–280.
[19] H. Hong, B. Preston, Bayesian averaging, prediction and nonnested model selection, J. Econometrics 167 (2012) 358–369.
[20] H. Ishwaran, J.S. Rao, Spike and slab variable selection: frequentist and Bayesian strategies, Ann. Statist. 33 (2005) 730–773.
[21] H. Ishwaran, J.S. Rao, Consistency of spike and slab regression, Statist. Probab. Lett. 81 (2011) 1920–1928.
[22] W. Jiang, Bayesian variable selection for high dimensional generalized linear models: Convergence rates of the fitted densities, Ann. Statist. 35 (2007) 1487–1511.
[23] W. Jiang, M.A. Tanner, Gibbs posterior for variable selection in high dimensional classification and data mining, Ann. Statist. 36 (2008) 2207–2231.
[24] W. Jiang, B. Turnbull, The indirect method: Inference based on intermediate statistics—a synthesis and examples, Statist. Sci. 19 (2004) 239–263.
[25] V.E. Johnson, D. Rossell, Bayesian model selection in high dimensional settings, J. Amer. Statist. Assoc. 107 (2012) 649–660.
[26] K. Kato, Quasi-Bayesian analysis of nonparametric instrumental variables models, Ann. Statist. 41 (2013) 2359–2390.
[27] J.Y. Kim, Limited information likelihood and Bayesian analysis, J. Econometrics 107 (2002) 175–193.
[28] J.Y. Kim, An alternative quasi likelihood approach, Bayesian analysis and data-based inference for model specification, J. Econometrics 178 (2014) 132–145.
[29] Y. Kitamura, T. Otsu, Bayesian analysis of moment condition models using nonparametric priors, Mimeo, 2011.
[30] Y. Kitamura, M. Stutzer, An information-theoretic alternative to generalized method of moments estimation, Econometrica 65 (1997) 861–874.
[31] G. Kundhi, P. Rilstone, Edgeworth expansions for GEL estimators, J. Multivariate Anal. 106 (2012) 118–146.
[32] G. Kundhi, R. Rilstone, Edgeworth and saddlepoint expansions for nonlinear estimators, Econometric Theory 29 (2013) 1057–1078.
[33] N.A. Lazar, Bayesian empirical likelihood, Biometrika 90 (2003) 319–326.
[34] C. Leng, C.Y. Tang, Penalized empirical likelihood and growing dimensional general estimating equations, Biometrika 99 (2012) 703–716.
[35] C. Li, W. Jiang, Model selection for likelihood-free Bayesian methods based on moment conditions: theory and numerical examples, 2014. arXiv:1405.6693.
[36] F. Liang, R. Paulo, G. Molina, M. Clyde, J.O. Berger, Mixture of g-priors for Bayesian variable selection, J. Amer. Statist. Assoc. 103 (2008) 410–423.
[37] F. Liang, Q. Song, K. Yu, Bayesian subset modeling for high-dimensional generalized linear models, J. Amer. Statist. Assoc. 108 (2013) 589–606.
[38] Y. Liao, W. Jiang, Bayesian analysis in moment inequality models, Ann. Statist. 38 (2010) 275–316.
[39] Y. Liao, W. Jiang, Posterior consistency of nonparametric conditional moment restricted models, Ann. Statist. 39 (2011) 3003–3031.
[40] J. Marin, N.S. Pillai, C.P. Robert, J. Rousseau, Relevant statistics for Bayesian model choice, J. R. Stat. Soc. Ser. B 76 (2014) 833–859.
[41] W.K. Newey, Efficient semiparametric estimation via moment restrictions, Econometrica 72 (2004) 1877–1897.
[42] W.K. Newey, R.J. Smith, Higher order properties of GMM and generalized empirical likelihood estimators, Econometrica 72 (2004) 219–255.
[43] A.B. Owen, Empirical likelihood ratio confidence intervals for a single functional, Biometrika 75 (1988) 237–249.
[44] J. Qin, J. Lawless, Empirical likelihood and general estimating equations, Ann. Statist. 22 (1994) 300–325.
[45] A. Qu, B.G. Lindsay, B. Li, Improving generalized estimating equations using quadratic inference functions, Biometrika 87 (2000) 823–836.
[46] S.M. Schennach, Bayesian exponentially tilted empirical likelihood, Biometrika 92 (2005) 31–46.
[47] S.M. Schennach, Point estimation with exponentially tilted empirical likelihood, Ann. Statist. 35 (2007) 634–672.
[48] M. Smith, R. Kohn, Nonparametric regression using Bayesian variable selection, J. Econometrics 75 (1996) 317–343.
[49] L. Wang, GEE analysis of clustered binary data with diverging number of covariates, Ann. Statist. 39 (2011) 389–417.
[50] S. Wang, L. Qian, R.J. Carroll, Generalized empirical likelihood methods for analyzing longitudinal data, Biometrika 97 (2010) 79–93.
[51] L. Wang, J. Zhou, A. Qu, Penalized generalized estimating equations for high-dimensional longitudinal data analysis, Biometrics 68 (2012) 353–360.
[52] G. Yin, Bayesian generalized method of moments, Bayesian Anal. 4 (2009) 191–208.
[53] G. Yin, Y. Ma, F. Liang, Y. Yuan, Stochastic generalized method of moments, J. Comput. Graph. Statist. 20 (2011) 714–727.
[54] A. Yuan, B. Clarke, Asymptotic normality of the posterior given a statistic, Canad. J. Statist. 32 (2004) 119–137.