Marginal Likelihood Estimation from the Metropolis Output: Tips and Tricks for Efficient Implementation in Generalised Linear Latent Variable Models

Silia Vitoratou,
Athens University of Economics and Business
Department of Statistics
76, Patision str, P.C. 10434, Greece
email: silia@aueb.gr

Ioannis Ntzoufras*
Athens University of Economics and Business
Department of Statistics
76, Patision str, P.C. 10434, Greece
email: ntzoufra@aueb.gr, phone: +30 210 8203968

Irini Moustaki
London School of Economics
Department of Statistics
Houghton Street,London, WC2A 2AE, United Kingdom
email: i.moustaki@lse.ac.uk, phone: +44 (0)20 7107 5172

March 24, 2013

*Corresponding author

Abstract

The marginal likelihood, can be notoriously difficult to compute, and particularly so in high dimensional problems. Chib and Jeliazkov employed the local reversibility of the Metropolis-Hastings algorithm to construct an estimator in models where full conditional densities are not available analytically. The estimator is free of distributional assumptions and is directly linked to the simulation algorithm. However, it generally requires a sequence of reduced Markov chain Monte Carlo (MCMC) runs which makes the method computationally demanding especially in cases when the parameter space is large. In this article, we study the implementation of this estimator on latent variable models which embed independence of the responses to the observables given the latent variables (conditional or local independence). This property is employed in the construction of a multi-block Metropolis-within-Gibbs algorithm that allows to compute the estimator in a single run, regardless of the dimensionality of the parameter space. The counterpart one-block algorithm is also considered here, by pointing out the difference between the two approaches. The paper closes with the illustration of the estimator in simulated and real life data sets.

KEYWORDS: Generalised linear latent variable models, Laplace-Metropolis estimator, MCMC, Bayes Factor

1 Introduction

Latent variable models are a broad family of models that can be used to capture abstract concepts by means of multiple indicators. Social scientists know them best in the form of factor analysis and structural equation models, in which continuous latent variables are captured by continuous or categorical observed variables also known as items or indicators. Social surveys and many other applications often yield observed variables that are categorical instead of continuous. That gave rise to the generalised linear latent variable model (GLLVM) framework [1] that can handle both continuous, discrete and categorical variables. Latent variable models are widely used in Social Sciences. Specifically, in educational testing where scores can be discrete but also binary indicating a correct/incorrect response (see [2], for real data examples). In applied psychometrics, where constructs such as stress, attitude, behavior are measured through ordinal or discrete indicators [3]. In demography, where fertility preferences and family planing behavior are usually measured with nixed categorical and survival indicators [4]. In archaeology, where classification of subjects is based on chemical analysis that produce continuous and binary indicators [5].

Various estimation methods have been proposed for estimating the parameters of latent variable models (see [6]). They are mainly divided into maximum likelihood estimation (likelihood of observed variables obtained by integrating out the latent variables) and MCMC estimation methods [7]. Model selection criteria such AIC and BIC are often used for selecting among models with different number of factors or between constrained and unconstrained models.

Within the Bayesian framework, model comparisons via the Bayes factor, posterior model probabilities and odds [8] require the computation of the marginal likelihood (integrated likelihood)

$$f(\mathbf{Y}|m) = \int f(\mathbf{Y}|\boldsymbol{\theta}, m)\pi(\boldsymbol{\theta}|m)d\boldsymbol{\theta}, \tag{1}$$

where $\mathbf{Y}$ denotes a vector of observed variables, m stands for the hypothesized model, and $\pi(\boldsymbol{\theta}|m)$ is the density of the model specific parameter vector $\boldsymbol{\theta}$ (m will be dropped hereafter for simplicity). The marginal likelihood often involves high dimensional integrals making the analytic computation infeasible, except in some special cases. Several approximating methods have been proposed in the literature for estimating the marginal likelihood, including Chib's estimator [9], the Bridge sampling estimator [10], the Laplace-Metropolis estimator [11], Chib and Jeliazkov estimator [12], and, lately, the power posterior estimator [13].

One of the most popular marginal likelihood estimators is the one proposed by [12] that extends the Chib's [9] original estimator, by allowing intractable full conditional densities. It is based on the estimation of the posterior ordinate evaluated at a high density point $\boldsymbol{\theta}^*$, using output from sequential Metropolis-Hastings algorithms, one for each element of $\boldsymbol{\theta}$. The sequential Markov chain Monte Carlo (MCMC) runs, appear to be computationally demanding when the parameter space is large. However, the method is favored by the fact that the posterior ordinate is directly obtained by the M-H kernel, used to produce the posterior output, while no additional assumptions are imposed during the marginal likelihood estimation. For instance, the estimators of the importance [14], or bridge family [10], even though very efficient, require to sample from a carefully constructed and well tuned envelope function. Quick approximation techniques, such as the Metropolised Laplace [11] or Gaussian copula

[15] estimators, can be also used but they impose distributional restrictions for the posterior, such as normality or symmetry. On the contrary, the Chib and Jeliazkov's [12] approach is based on the M-H kernel per ce, without any additional restrictions or assumptions.

In this article, the Chib and Jeliazkov [CJ, 12] estimator is studied in latent variables models that have a large number of parameters and together with the latent variables, estimation and goodness-of-fit testing can involve heavy integrations. A wide range of models with latent variables, such as the GLLVM, embed independence of the responses to the observed variables given the latent variables (conditional or local independence). The local independence is employed in the construction of a multi-block Metropolis-within-Gibbs (M-G) algorithm that allows to compute the CJ estimator in a single MCMC run, regardless of the dimensionality of the parameter space. This is achieved simultaneously by marginalizing out the latent vector directly from the M-G kernel and estimating the posterior ordinate via the CJ method. The alternative one-block algorithm is also considered here, by pointing out the difference between the two approaches. In the absence of reduced MCMC runs, the CJ estimator is considerably simplified, minimizing the computational burden. Regarding the models where local independence is not assumed, it is described how the latent variables can be marginalized out, when none of the conditional posterior ordinates is fully available and therefore Rao-Blackwellization is not applicable ([12], [16], [17]).

The rest of the article is organized as follows. Section 2 gives the general model framework for fitting models with latent variables. In Section 3, we describe the CJ [12] estimator. In Section 4, we describe how the method can be simplified using the local independence assumption of the likelihood and we compare it with other single-run versions of the method. The section closes with a discussion on how the estimator can be implemented when none of the conditional posterior ordinates is analytically available. Section 5 describes the implementation in generalised latent variable models including illustrations on simulated and real data sets. Concluding remarks are provided in the closing section of this article.

2 Framework and model formulation

Let us first define the general model structure and the corresponding notation. Here, we study models which can be defined with a likelihood of the following structure

$$f\Big(\boldsymbol{Y}\,|\,\boldsymbol{\Theta} = (\boldsymbol{\theta}_0, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_p), \boldsymbol{L}\Big) = f\Big(\boldsymbol{Y}\,|\,\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_p), \mathbf{Z} = (\boldsymbol{\theta}_0, \boldsymbol{L})\Big), \qquad (2)$$

where

- $\boldsymbol{Y} = (\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_p)$ is a $n \times p$ data array of n observations and p observed variables (items),
- $\mathbf{Y}_j$ is the $n \times 1$ vector with the data values for item j,
- $\boldsymbol{L}$ is the $k \times n$ matrix of the latent variables,
- $\boldsymbol{\Theta}$ is the whole parameters $(k+1) \times p$ vector,
- $\boldsymbol{\theta}_0$ is the set of parameters which is common across different items,

- $\boldsymbol{\theta}_j$ for $j = 1, \ldots, p$ are the item specific parameters (linked to $\mathbf{Y}_j$ only).

The above setting includes a variety of models, such as random effect models and the generalised linear latent variable models [1, 18]. In the remaining we focus on the latter model formulation. More specifically, the GLLVM [1] consist of three components: (i) the multivariate random component $\mathbf{Y}$ of the observed variables, (ii) the linear predictor denoted by η_j and (iii) the link function $\upsilon(\cdot)$, which connects the previous two components. Hence, a GLLVM can be summarized as:

$$Y_j|\mathbf{Z} \sim ExpF, \quad \eta_j = \alpha_j + \sum_{\ell=1}^{k} \beta_{j\ell} Z_\ell, \quad \text{and} \quad \upsilon_j\Big(\mu_j(\mathbf{Z})\Big) = \eta_j \tag{3}$$

for $j = 1, \ldots p$; where $ExpF$ is a member of the exponential family and $\mu_j(\mathbf{Z}) = \mathrm{E}(\mathrm{Y}_j|\mathbf{Z})$. Finally, a multivariate distribution $\pi(\mathbf{Z})$ needs to be specified for the latent variables, which is usually assumed to be a multivariate *standard normal* distribution. It is assumed that the responses to the observed variables are independent given the latent vector $\mathbf{Z}$ (within subjects independence), and the item specific parameters $\boldsymbol{\theta}$ (between subjects independence) resulting in,

$$f(\boldsymbol{Y}|\boldsymbol{\theta}, \mathbf{Z}) = \prod_{i=1}^{n} \prod_{j=1}^{p} f(Y_{ij}|\boldsymbol{\theta}_j, \mathbf{Z}_i)\,, \tag{4}$$

where $\boldsymbol{Y}$ is the observed data matrix with elements Y_{ij} denoting the response of subject i to item j, $\mathbf{Z}_i$ are the subject specific values of the latent variables $\mathbf{Z}$, $\boldsymbol{\theta} = (\boldsymbol{\alpha}, \boldsymbol{\beta})$, $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_p)$, $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_p)$, $\boldsymbol{\beta}_j = (\beta_{j1}, \ldots, \beta_{jk})$ and $\boldsymbol{\theta}_j = (\alpha_j, \boldsymbol{\beta}_j)$.

Note that in the model formulation in (2), the pair of parameters and the latent variables $(\boldsymbol{\Theta}, \boldsymbol{L})$ correspond to the pair $(\boldsymbol{\theta}, \mathbf{Z})$ with $\boldsymbol{\theta}$ being the item specific parameters and $\mathbf{Z}$ being the set of parameters and/or latent variables which are common and shared across different items. In GLLVMs, parameters shared across different items do not exist unless equality constraints are imposed. Hence $\mathbf{Z}$ solely refers to latent variables $\boldsymbol{L}$.

Finally, model identification, which is crucial for the parameter estimation, can be obtained if the loading matrix is constrained to be a full rank lower triangular matrix (see also [19], [20] and [21]). Here we follow this approach by setting $\beta_{j\ell} = 0$ for all $j < \ell$ and $\beta_{jj} > 0$.

3 The Chib and Jeliazkov marginal likelihood estimator

Both Chib's [9] and Chib and Jeliazkov [12] estimators, are based on the *candidate's identity* [22]:

$$f(\mathbf{Y}) = \frac{f(\mathbf{Y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\pi(\boldsymbol{\theta}|\mathbf{Y})} \Leftrightarrow \log f(\mathbf{Y}) = \log f(\mathbf{Y}|\boldsymbol{\theta}) + \log \pi(\boldsymbol{\theta}) - \log \pi(\boldsymbol{\theta}|\mathbf{Y}). \tag{5}$$

From (5), the marginal likelihood depends on the posterior density of the model parameters $\pi(\boldsymbol{\theta}|\mathbf{Y})$. Since (5) holds for every point $\boldsymbol{\theta}$ of the parameter space, the posterior density can be estimated using a specific point $\boldsymbol{\theta}^*$. Following [9], let us suppose that the parameter space is divided into p blocks of parameters. Then the posterior ordinate can be decomposed to

$$\pi(\boldsymbol{\theta^*}|\mathbf{Y}) = \pi(\theta_1^*, \boldsymbol{\theta}_2^*, \cdots, \boldsymbol{\theta}_p^*|\mathbf{Y}) = \pi(\boldsymbol{\theta}_1^*|\mathbf{Y})\pi(\boldsymbol{\theta}_2^*|\mathbf{Y}, \boldsymbol{\theta}_1^*) \cdots \pi(\boldsymbol{\theta}_p^*|\mathbf{Y}, \boldsymbol{\theta}_1^*, \boldsymbol{\theta}_2^*, \cdots, \boldsymbol{\theta}_{p-1}^*). \tag{6}$$

The marginal likelihood is calculated in a straightforward manner when (6) is analytically available. In the case when the full conditionals are known, [9] presented an algorithm that uses the output from the Gibbs sampler to estimate them by Rao-Blackwellization. In addition, [12] extended the method to deal with cases where the full conditional posterior distributions are not available and, therefore, a Metropolis–Hastings (M-H) algorithm is used to generate posterior samples. The authors implement for that purpose the kernel of the M-H algorithm, which denotes the transition probability of sampling $\boldsymbol{\theta}_j^*$ given that $\boldsymbol{\theta}_j$ has been already generated

$$K(\boldsymbol{\theta}_j, \boldsymbol{\theta}_j^*|\mathbf{Y}, \boldsymbol{\theta}_{\backslash j}) = a(\boldsymbol{\theta}_j, \boldsymbol{\theta}_j^*|\mathbf{Y}, \boldsymbol{\theta}_{\backslash j})q(\boldsymbol{\theta}_j, \boldsymbol{\theta}_j^*|\mathbf{Y}, \boldsymbol{\theta}_{\backslash j}), \quad j = 1, \cdots, p, \tag{7}$$

where $\boldsymbol{\theta}_{\backslash j}$ is the parameter vector $\boldsymbol{\theta}$ without $\boldsymbol{\theta}_j$, $a(\boldsymbol{\theta}_j, \boldsymbol{\theta}_j^*|\mathbf{Y}, \boldsymbol{\theta}_{\backslash j})$ is the M-H acceptance probability and $q(\boldsymbol{\theta}_j, \boldsymbol{\theta}_j^*|\mathbf{Y}, \boldsymbol{\theta}_{\backslash j})$ is the proposal density. Employing the local reversibility condition, each of the posterior ordinate appearing in (6) can be written as

$$\pi(\boldsymbol{\theta}_j^*|\mathbf{Y}, \boldsymbol{\theta}_1^*, \cdots, \boldsymbol{\theta}_{j-1}^*) = \frac{E_1\Big\{a\big(\boldsymbol{\theta}_j, \boldsymbol{\theta}_j^*|\mathbf{Y}, \psi_{j-1}^*, \psi^{j+1}\big)\, q\big(\boldsymbol{\theta}_j, \boldsymbol{\theta}_j^*|\mathbf{Y}, \psi_{j-1}^*, \psi^{j+1}\big)\Big\}}{E_2\Big\{a\big(\boldsymbol{\theta}_j^*, \boldsymbol{\theta}_j|\mathbf{Y}, \psi_{j-1}^*, \psi^{j+1}\big)\Big\}}, \tag{8}$$

where $\psi_{j-1} = (\boldsymbol{\theta}_1, \cdots, \boldsymbol{\theta}_{j-1})$ and $\psi^{j+1} = (\boldsymbol{\theta}_{j+1}, \cdots, \boldsymbol{\theta}_p)$ for $j = 1, \ldots, p$ with ψ_0 and ψ^{p+1} referring to the empty sets. The expectations in the numerator and the denominator are with respect to $\pi\big(\boldsymbol{\theta}_j, \psi^{j+1}|\mathbf{Y}, \psi_{j-1}^*\big)$ and $\pi\big(\psi^{j+1}|\mathbf{Y}, \psi_j^*\big)q\big(\boldsymbol{\theta}_j, \boldsymbol{\theta}_j^*|\psi_{j-1}^*, \psi^{j+1}\big)$ accordingly.

A Monte Carlo estimator for each ordinate can be obtained by replacing the expectations in (8) with their corresponding sample means from simulated samples. The final posterior estimator $(\widehat{CJ})$ is given by multiplying the estimators for each block. Since the expectations in (8) are conditional on specific parameter points $\psi_{j-1}^* = (\boldsymbol{\theta}_1^*, \cdots, \boldsymbol{\theta}_{j-1}^*)$, the corresponding Monte Carlo estimates cannot be obtained by the initial (full) MCMC run. Hence, for a parameter space that consists of p blocks, $p-1$ reduced runs are needed to compute the CJ estimator. For models with latent variables, whose number of parameters when including the latent variables exceeds several hundreds, estimating the posterior ordinate requires a marginalization step. This is discussed in detail in Section 4.

In the case of the GLLVM, the posterior ordinate required to calculate the marginal likelihood includes all parameters, that is $\pi(\boldsymbol{\theta}^*, \mathbf{Z}^*|\boldsymbol{Y})$. Usually, the number of blocks employed for $\boldsymbol{\theta}^*$ is reasonable creating no problem in the computation of CJ. On the contrary, the latent vector $\mathbf{Z}$ is highly dimensional and direct application of the CJ method requires a large number of reduced MCMC runs. Chib and Jeliazkov [12] address the issue of multiple latent variable blocks and suggest to overcome the problem by marginalizing out the latent vector. Specifically, the first $p-1$ ordinates are estimated via (8), while the last one is calculated via a Rao-Blackwellization step as the average of $\pi(\boldsymbol{\theta}_p^*|\mathbf{Y}, \psi_{p-1}^*, \mathbf{Z})$ with respect to $\pi(\mathbf{Z}|\mathbf{Y}, \psi_{p-1}^*)$. This straightforward solution occurs when at least one conditional density is analytically available. The procedure is discussed in detail in Chib and Jeliazkov [12], along with examples, as well as within the longitudinal data setting considered in [23].

In the next section we describe how the Metropolis kernel can be used to marginalize out the latent vector, when the Rao-Blackwellization step is not applicable. Different scenarios are considered for models under the setting given in equation (2).

4 Efficient estimation of the posterior ordinate in latent variable models

Within the framework given in equation (2), a multi-block CJ estimator using a single-run of the Metropolis algorithm is described, based on local independence properties of models with latent vectors. The one-block approach that also leads to single-run CJ estimators is discussed along with practical solutions when the local independence assumptions are not met.

4.1 Models with local independence

As mentioned in Section 2, the GLLVM framework embraces the within subjects independence that is typical also in various models with latent vector and/or random effects. This property is met in the literature as the local (conditional) independence assumption.

Definition 4.1 *The **local independence** refers to the independence of the data ($\boldsymbol{Y}$) conditional on the latent vector (within subjects independence). That is, under the assumption of local independence, it holds that*

$$f(\boldsymbol{Y}|\,\boldsymbol{\theta}, \boldsymbol{Z}) = \prod_{j=1}^{p} f(\,\boldsymbol{Y}_j|\,\boldsymbol{\theta}_j, \boldsymbol{Z})\,, \tag{9}$$

The local independence implies also that the association among the observed variables for the ith individual is induced solely by the individual's latent position $\boldsymbol{Z}_i$, $i \in \{1, 2, ..., n\}$.

The key observation here is that the local independence can be extended to the posterior distribution of the parameters provided that *prior local independence* exists, that is introduced in Definition 4.2 which follows.

Definition 4.2 *For any model with likelihood given by equation (2), a set of parameters $\boldsymbol{\theta}$ is defined as **a-priori locally independent** if they are a-priori independent conditionally on $\boldsymbol{Z}$. Therefore, the prior will satisfy the following equation*

$$f(\boldsymbol{\theta}|\boldsymbol{Z}) = \prod_{j=1}^{p} f(\boldsymbol{\theta}_j|\boldsymbol{Z})\,. \tag{10}$$

Similarly we can introduce the *posterior local independence* using Definition 4.3.

Definition 4.3 *For any model with likelihood (2), a set of parameters $\boldsymbol{\theta}$ is defined as* ***a-posteriori locally independent*** *if they are a-posteriori independent conditionally on $\boldsymbol{Z}$. Therefore the posterior distribution will satisfy the following equation*

$$f(\boldsymbol{\theta}|\, \boldsymbol{Y}, \boldsymbol{Z}) = \prod_{j=1}^{p} f(\boldsymbol{\theta}_j|\, \boldsymbol{Y}_j, \boldsymbol{Z})\,. \tag{11}$$

For any model where the assumptions of local and prior local independence hold, it is trivial to show that the posterior local independence holds as well. These properties naturally affect the acceptance probability of the sampling algorithm and consequently the implementation of the CJ estimator in either multi-block or one-block designs.

4.1.1 CJ estimator from a single run using multi-block MCMC

In this section we introduce a simplification of the original CJ estimator that occurs in models with local (conditional) independence, denoted hereafter as the *independence* CJ estimator ($\widehat{CJ}^{\mathrm{I}}$). The estimator occurs under the Metropolis-within-Gibbs algorithm described by the following steps:

1. Sample $\mathbf{Z}$ from $f(\mathbf{Z}|\mathbf{Y}, \boldsymbol{\theta})$ using any sampling scheme.
2. for $j = 1, \ldots, p$
   (a) When $\boldsymbol{\theta}_j$ is the current parameter value, propose $\boldsymbol{\theta}'_j$ from a proposal with density $q(\boldsymbol{\theta}_j, \boldsymbol{\theta}'_j|\mathbf{Y}, \mathbf{Z})$.
   (b) Accept the proposed move with probability

$$\begin{aligned} a(\boldsymbol{\theta}_j, \boldsymbol{\theta}'_j|\mathbf{Y}, \boldsymbol{\theta}_{\setminus j}, \mathbf{Z}) &= \min\left\{1, \frac{f(\mathbf{Y}|\boldsymbol{\theta}_{\setminus j}, \boldsymbol{\theta}'_j, \mathbf{Z})\pi(\boldsymbol{\theta}_{\setminus j}, \boldsymbol{\theta}'_j|\mathbf{Z})\pi(\mathbf{Z})q(\boldsymbol{\theta}_j, \boldsymbol{\theta}'_j|\mathbf{Y}, \mathbf{Z})}{f(\mathbf{Y}|\boldsymbol{\theta}_{\setminus j}, \boldsymbol{\theta}_j, \mathbf{Z})\pi(\boldsymbol{\theta}_{\setminus j}, \boldsymbol{\theta}_j|\mathbf{Z})\pi(\mathbf{Z})q(\boldsymbol{\theta}_j, \boldsymbol{\theta}'_j|\mathbf{Y}, \mathbf{Z})}\right\} \\ &= \min\left\{1, \frac{f(\mathbf{Y}_j|\boldsymbol{\theta}'_j, \mathbf{Z})\pi(\boldsymbol{\theta}'_j|\mathbf{Z})\, q(\boldsymbol{\theta}'_j, \boldsymbol{\theta}_j|\mathbf{Y}, \mathbf{Z})}{f(\mathbf{Y}_j|\boldsymbol{\theta}_j, \mathbf{Z},)\pi(\boldsymbol{\theta}_j|\mathbf{Z})\, q(\boldsymbol{\theta}_j, \boldsymbol{\theta}'_j|\mathbf{Y}, \mathbf{Z})}\right\} = a(\boldsymbol{\theta}_j, \boldsymbol{\theta}'_j|\mathbf{Y}, \mathbf{Z}), \end{aligned} \tag{12}$$

due to local and prior local independence defined in (9) and (10). Therefore the acceptance rate given in (12) depends only on the current and new (proposed) values of component $\boldsymbol{\theta}_j$ and the latent vector $\mathbf{Z}$. This assumption is common when implementing Metropolis-within-Gibbs algorithms, with the simpler case described by a simple random walk algorithm. Moreover, since the components of $\boldsymbol{\theta}$ are independent for given values of $\mathbf{Z}$ it is reasonable to adopt proposals that take into account only the current status of $\boldsymbol{\theta}_j$.

The simplification of the acceptance probability achieved due to the local independence directly affects the Metropolis kernel given in (7). Following similar arguments as in [12], we can exploit the the local reversibility condition at any point $\boldsymbol{\theta}^*_j$:

$$K(\boldsymbol{\theta}_j, \boldsymbol{\theta}^*_j|\mathbf{Y}, \mathbf{Z}, \boldsymbol{\theta}_{\setminus j})\, \pi(\boldsymbol{\theta}_j|\mathbf{Y}, \mathbf{Z}, \boldsymbol{\theta}_{\setminus j}) = K(\boldsymbol{\theta}^*_j, \boldsymbol{\theta}_j|\mathbf{Y}, \mathbf{Z}, \boldsymbol{\theta}_{\setminus j})\, \pi(\boldsymbol{\theta}^*_j|\mathbf{Y}, \mathbf{Z}, \boldsymbol{\theta}_{\setminus j}),$$

taking under consideration the posterior local independence given in equation (11)

$$K(\boldsymbol{\theta}_j, \boldsymbol{\theta}^*_j|\mathbf{Y}, \mathbf{Z})\, \pi(\boldsymbol{\theta}_j|\mathbf{Y}, \mathbf{Z}) = K(\boldsymbol{\theta}^*_j, \boldsymbol{\theta}_j|\mathbf{Y}, \mathbf{Z})\, \pi(\boldsymbol{\theta}^*_j|\mathbf{Y}, \mathbf{Z}).$$

By integrating both sides of the equation over $\boldsymbol{\theta}_j$, we obtain

$$\int K(\boldsymbol{\theta}_j, \boldsymbol{\theta}_j^* | \mathbf{Y}, \mathbf{Z})\, \pi(\boldsymbol{\theta}_j | \mathbf{Y}, \mathbf{Z}) d\boldsymbol{\theta}_j = \int K(\boldsymbol{\theta}_j^*, \boldsymbol{\theta}_j | \mathbf{Y}, \mathbf{Z})\, \pi(\boldsymbol{\theta}_j^* | \mathbf{Y}, \mathbf{Z}) d\boldsymbol{\theta}_j,$$

resulting in

$$CJ_j^I = \pi(\boldsymbol{\theta}_j^* | \mathbf{Y}, \mathbf{Z}) = \frac{\int K(\boldsymbol{\theta}_j, \boldsymbol{\theta}_j^* | \mathbf{Y}, \mathbf{Z})\, \pi(\boldsymbol{\theta}_j | \mathbf{Y}, \mathbf{Z}) d\boldsymbol{\theta}_j}{\int K(\boldsymbol{\theta}_j^*, \boldsymbol{\theta}_j | \mathbf{Y}, \mathbf{Z}) d\boldsymbol{\theta}_j}\ , \tag{13}$$

if we solve with respect to $\pi(\boldsymbol{\theta}_j^* | \mathbf{Y}, \mathbf{Z})$.

The expression for the posterior $\pi(\boldsymbol{\theta}|\mathbf{Y})$ is then given by multiplying CJ_j^I over all p blocks and integrate out the latent variables directly from the kernel. Therefore, we have that

$$\begin{aligned}
\pi(\boldsymbol{\theta}^*|\mathbf{Y}) &= \int \prod_{j=1}^{p} \pi(\boldsymbol{\theta}_j^* | \mathbf{Y}_j, \mathbf{Z}) \pi(\mathbf{Z}|\mathbf{Y}) d\mathbf{Z} \\
&= \int \prod_{j=1}^{p} \left[\frac{\int K(\boldsymbol{\theta}_j, \boldsymbol{\theta}_j^* | \mathbf{Y}, \mathbf{Z})\, \pi(\boldsymbol{\theta}_j | \mathbf{Y}, \mathbf{Z}) d\boldsymbol{\theta}_j}{\int K(\boldsymbol{\theta}_j^*, \boldsymbol{\theta}_j | \mathbf{Y}, \mathbf{Z}) d\boldsymbol{\theta}_j} \right] \pi(\mathbf{Z}|\mathbf{Y})\, d\mathbf{Z} \\
&= \int \left[\frac{\prod_{j=1}^{p} K(\boldsymbol{\theta}_j, \boldsymbol{\theta}_j^* | \mathbf{Y}, \mathbf{Z})}{\prod_{j=1}^{p} \int K(\boldsymbol{\theta}_j^*, \boldsymbol{\theta}_j | \mathbf{Y}, \mathbf{Z}) d\boldsymbol{\theta}_j} \right] \pi(\boldsymbol{\theta}, \mathbf{Z}|\mathbf{Y})\, d(\boldsymbol{\theta}, \mathbf{Z}) \\
&= E_{\boldsymbol{\theta}, \mathbf{Z}|\mathbf{Y}} \left[\frac{\prod_{j=1}^{p} a\left(\boldsymbol{\theta}_j, \boldsymbol{\theta}_j^* | \mathbf{Y}, \mathbf{Z}\right) q\left(\boldsymbol{\theta}_j, \boldsymbol{\theta}_j^* | \mathbf{Y}, \mathbf{Z}\right)}{\prod_{j=1}^{p} E_{q_j}\left[a\left(\boldsymbol{\theta}_j^*, \boldsymbol{\theta}_j | \mathbf{Y}, \mathbf{Z}\right) \right]} \right],
\end{aligned} \tag{14}$$

where $E_{\boldsymbol{\theta}, \mathbf{Z}|\mathbf{Y}}$ is the posterior mean and E_{q_j} are the expectations with respect each of the proposal densities $q(\boldsymbol{\theta}_j^*, \boldsymbol{\theta}_j | \mathbf{Y}, \mathbf{Z})$. Thus equation (14) can be estimated from:

$$\widehat{CJ}^I = \frac{1}{R} \sum_{r=1}^{R} \left[\frac{\prod_{j=1}^{p} a\big(\boldsymbol{\theta}_j^{(r)}, \boldsymbol{\theta}_j^* | \mathbf{Y}, \mathbf{Z}^{(r)}\big)\, q\big(\boldsymbol{\theta}_j^{(r)}, \boldsymbol{\theta}_j^* | \mathbf{Y}, \mathbf{Z}^{(r)}\big)}{\prod_{j=1}^{p} \left[\frac{1}{M} \sum_{m=1}^{M} a\big(\boldsymbol{\theta}_j^*, \boldsymbol{\theta}_j^{(m)} | \mathbf{Y}, \mathbf{Z}^{(r)}\big) \right]} \right]. \tag{15}$$

The sample $\{\boldsymbol{\theta}_1^{(r)}, \boldsymbol{\theta}_2^{(r)}, \cdots, \boldsymbol{\theta}_p^{(r)}, \mathbf{Z}^{(r)}\}_{r=1}^{R}$ comes from the joint posterior of $(\boldsymbol{\theta}, \mathbf{Z})$ which is available from a full MCMC run. For each sampled set of latent and parameter values $\big(\boldsymbol{\theta}^{(r)}, \mathbf{Z}^{(r)}\big)$, $r = 1, ..., R$, additional points $\{\boldsymbol{\theta}_j^{(m)}\}_{m=1}^{M}$ are generated from each proposal density $q(\boldsymbol{\theta}_j^*, \boldsymbol{\theta}_j | \mathbf{Y}, \mathbf{Z}, \boldsymbol{\theta})$. These values are used to compute the expectation in the denominator of (14). From (15), it is straightforward to see that a single MCMC run from the posterior of the model under study is required to compute the independence estimator $\widehat{CJ}^{\mathrm{I}}$.

To sum up, $\widehat{CJ}^{\mathrm{I}}$ is based on the local independence assumption. The prior local independence (10), on its turn, is a reasonable assumption for such models. The above properties lead to the posterior local independence which actually ensures the one run procedure. Most importantly, the $\widehat{CJ}^{\mathrm{I}}$ is based solely on the generation of a posterior sample using a multi-block Metropolis-within-Gibbs algorithm and is applicable when none of the posterior ordinates are analytically available, since the marginalization is directly implemented in the corresponding kernel.

4.1.2 An alternative one-block CJ estimator

An alternative way to obtain a single-run CJ estimator is to consider all parameters $\boldsymbol{\theta}$ as one block jointly proposed by $q(\boldsymbol{\theta}, \boldsymbol{\theta}'|\mathbf{Y}, \mathbf{Z})$. For models with structure described by (2), under local and prior local independence assumptions, the acceptance probability under the one-block design is given by

$$a(\boldsymbol{\theta}, \boldsymbol{\theta}'|\mathbf{Y}, \mathbf{Z}) = \min \left\{ 1, \frac{\prod_{j=1}^{p} \left[f(\mathbf{Y}_j|\boldsymbol{\theta}_j', \mathbf{Z})\pi(\boldsymbol{\theta}_j'|\mathbf{Z}) \right] q(\boldsymbol{\theta}', \boldsymbol{\theta}|\mathbf{Y}, \mathbf{Z})}{\prod_{j=1}^{p} \left[f(\mathbf{Y}_j|\boldsymbol{\theta}_j, \mathbf{Z})\pi(\boldsymbol{\theta}_j|\mathbf{Z}) \right] q(\boldsymbol{\theta}, \boldsymbol{\theta}'|\mathbf{Y}, \mathbf{Z})} \right\}. \tag{16}$$

Even though the properties of the local and prior local independence were also used here, the expression in (16) cannot be simplified further, since it requires the entire parameter vector $\boldsymbol{\theta}$, unlike the acceptance probabilities in (12). This is the major difference between the two sampling schemes and is directly reflected to the corresponding posterior ordinate expressions, under the CJ method. As opposed to (14), the expression of the posterior ordinate under the one-block design is given by

$$\pi(\boldsymbol{\theta}^*|\mathbf{Y}) \quad = \quad E_{\boldsymbol{\theta},\mathbf{Z}|\mathbf{Y}} \left[\frac{a\left(\boldsymbol{\theta}, \boldsymbol{\theta}^*|\mathbf{Y}, \mathbf{Z}\right) q(\boldsymbol{\theta}^*, \boldsymbol{\theta}|\mathbf{Y}, \mathbf{Z})}{E_q \left[a\left(\boldsymbol{\theta}^*, \boldsymbol{\theta}|\mathbf{Y}, \mathbf{Z}\right) \right]} \right], \tag{17}$$

with draws coming from the posterior $\pi(\boldsymbol{\theta}, \mathbf{Z}|\mathbf{Y})$ for the nominator and from the proposal density $q(\boldsymbol{\theta}^*, \boldsymbol{\theta}|\mathbf{Y}, \mathbf{Z})$ for the denominator. The difference between the expressions in (17) and (14) becomes more evident if we assume $q(\boldsymbol{\theta}', \boldsymbol{\theta}|\mathbf{Y}, \mathbf{Z}) = \prod_{j=1}^{p} q(\boldsymbol{\theta}_j', \boldsymbol{\theta}_j|\mathbf{Y}, \mathbf{Z})$ that is reasonable due to the local and prior local independence. By defining the quantity A_j as

$$A_j = \frac{f(\mathbf{Y}_j|\boldsymbol{\theta}_j^*, \mathbf{Z})\pi(\boldsymbol{\theta}_j^*|\mathbf{Z})\, q(\boldsymbol{\theta}_j^*, \boldsymbol{\theta}_j|\mathbf{Y}, \mathbf{Z})}{f(\mathbf{Y}_j|\boldsymbol{\theta}_j, \mathbf{Z})\pi(\boldsymbol{\theta}_j|\mathbf{Z})\, q(\boldsymbol{\theta}_j, \boldsymbol{\theta}_j^*|\mathbf{Y}, \mathbf{Z})},$$

the acceptance probabilities involved in the posterior ordinate expressions (17) and (14) are given by $\min\left\{1, \prod_{j=1}^{p} A_j\right\}$ in the case of the one-block design, and by $\prod_{j=1}^{p} \min\{1, \ A_j\}$ under a multi-block design, respectively.

Using one-block MCMC for $\boldsymbol{\theta}$ may be beneficial in terms of mixing only when parameters are a-posteriori depended [24, see Section 1.4.2] which is not the case for the models here where local and prior local independence is assumed. Therefore, the single-run multi-block estimator (15) is expected to be more efficient and accurate than the alternative one-block, for the same number of iterations.

4.2 Models without local independence

When local independence cannot be assumed, one of the posterior ordinates in (6) can be exploited in order to marginalize out the latent vector $\mathbf{Z}$. Chib and Jeliazkov [12] suggest to add a Rao-Blackwellization step at the end of the procedure for this purpose, provided that $\pi(\boldsymbol{\theta}_p^*|\mathbf{Y},\mathbf{Z},\psi_{p-1}^*)$ is analytically available. Here, we further describe that if there is not such a conditional ordinate analytically available, then we estimate it by integrating out $\mathbf{Z}$ from (8) and them implement the same strategy as in the CJ method. That is achieved directly from the local reversibility condition of the corresponding sub-kernel:

$$\pi(\boldsymbol{\theta}_p^*|\mathbf{Y},\mathbf{Z},\psi_{p-1}^*) = \frac{\int K(\boldsymbol{\theta}_p\ ,\boldsymbol{\theta}_p^*|\mathbf{Y},\mathbf{Z},\psi_{p-1}^*)\pi(\boldsymbol{\theta}_p|\mathbf{Y},\mathbf{Z},\psi_{p-1}^*)\ d\boldsymbol{\theta}_p}{\int K(\boldsymbol{\theta}_p^*\ ,\boldsymbol{\theta}_p|\mathbf{Y},\mathbf{Z},\psi_{p-1}^*)\ d\boldsymbol{\theta}_p}.$$

The latent vector is then integrated out directly from the kernel

$$\begin{aligned}\pi(\boldsymbol{\theta}_p^*|\mathbf{Y},\psi_{p-1}^*) &= \int\left[\frac{\int K(\boldsymbol{\theta}_p\ ,\boldsymbol{\theta}_p^*|\mathbf{Y},\mathbf{Z},\psi_{p-1}^*)\pi(\boldsymbol{\theta}_p|\mathbf{Y},\mathbf{Z},\psi_{p-1}^*)\ d\boldsymbol{\theta}_p}{\int K(\boldsymbol{\theta}_p^*\ ,\boldsymbol{\theta}_p|\mathbf{Y},\mathbf{Z},\psi_{p-1}^*)\ d\boldsymbol{\theta}_p}\right]\pi(\mathbf{Z}|\mathbf{Y},\psi_{j-1}^*)\ d\mathbf{Z} \\ &= \int\frac{K(\boldsymbol{\theta}_p\ ,\boldsymbol{\theta}_p^*|\mathbf{Y},\mathbf{Z},\psi_{p-1}^*)}{\int K(\boldsymbol{\theta}_p^*\ ,\boldsymbol{\theta}_p|\mathbf{Y},\mathbf{Z},\psi_{p-1}^*)\ d\boldsymbol{\theta}_p}\ \pi(\boldsymbol{\theta}_p\ ,\mathbf{Z}|\mathbf{Y},\psi_{p-1}^*)\ d(\boldsymbol{\theta}_p\ ,\mathbf{Z})\,. \qquad (18)\end{aligned}$$

The corresponding estimator of (18) is identical with (15), for $p = 1$ and conditioning upon $\{\mathbf{Y},\mathbf{Z},\psi_{p-1}^*\}$. Naturally, the first $p-1$ ordinates in (6) are estimated via (8), while the last ordinate is used to marginalize out the latent variables. The M-H output required for the marginalization is already available from the reduced run implemented to assess the denominator of the previous ordinate. Finally, a single-run estimator can be obtained, in a straightforward manner, by sampling all parameters in $\boldsymbol{\theta}$ one block as described in Section 4.1.2.

5 Applications on GLLVM

In this section we illustrate the estimators discussed in Section 4 in simulated and real datasets. Emphasis is given in the estimation of the marginal likelihood and in the computation of the Bayes factor as means of comparing models with different number of factors. All our examples are for binary observed variables but the methodology, as already discussed, can be applied in all GLLVMs.

5.1 Latent trait model and Bayes factor estimation

The latent variable model for binary observed variables is also known in the Psychometric literature as a latent trait model (LTM). The LTM is a a special case of the GLLVM discussed in [1]. The link used here is the logit giving:

$$\text{logit}\Big[E\big(Y_j|\mathbf{Z}\big)\Big] = \log\frac{P(Y_j=1\mid\mathbf{Z})}{1-P(Y_j=1\mid\mathbf{Z})} = \alpha_j + \sum_{\ell=1}^{k}\beta_{j\ell}\mathbf{Z}_\ell,\quad j=1,\ldots,p. \qquad (19)$$

The prior used here is based on the ideas presented by [25] and further explored in the context of generalised linear models by Fouskakis et al. [26, equation 6]. For GLLVMs with binary variables, this prior corresponds to a $N(0,4)$ for all non-constrained loadings and for all α_j. For all the β_{jj} parameters we assume a standardized normal distribution as a prior for each $\log \beta_{jj}$ inducing prior a standard deviation for β_{jj} approximately equal to 2, in analogy with the rest non-zero parameters β_{jl}. To summarize, the prior is given by:

$$\pi(\beta_{j\ell}) = \begin{cases} 0 \text{ with probability } 1 & \text{if} \quad j < \ell \\ LN(0,1) & \text{if} \quad j = \ell \\ N(0,4) & \text{if} \quad j > \ell \end{cases}$$

where $Y \sim LN(\mu, \sigma^2)$ is the log-normal distribution with the mean and the variance of $\log Y$ being equal to μ and σ^2, respectively. Finally, latent variables are assumed to be a-priori distributed as independent standard normal distributions i.e. $\mathbf{Z}_\ell \sim N(0,1)$ for all subjects. These complete the model specification, under the Bayesian paradigm and we may proceed to the marginal likelihood and the Bayes factor (BF) estimation. The BF is computed for pairs of competing models (m_1, m_2) as the ratio of their marginal likelihoods given by (1): $BF_{12} = \frac{f(\mathbf{Y}|m_1)}{f(\mathbf{Y}|m_2)}$, or, alternatively, as the ratio of their posterior model probabilities assuming that they are a-priori equivalent [8]. Kass and Raftery [8] state threshold values for the BF. Specifically, values larger than one provide evidence in favor of m_1, while values higher than two are considered decisive.

The estimate of the log marginal likelihood based on the CJ^I is given by

$$\log \widehat{\mathcal{L}} = \log f(\mathbf{Y}|\boldsymbol{\theta}^*) + \log \pi(\boldsymbol{\theta}^*) - \log \widehat{CJ^I}. \tag{20}$$

Note that for a random sample of size n, the observed likelihood $f(\boldsymbol{Y}|\boldsymbol{\theta})$, is obtained by marginalizing out the latent variables:

$$f(\boldsymbol{Y}|\boldsymbol{\theta}) = \prod_{i=1}^{n} f(\mathbf{Y}_i|\boldsymbol{\theta}) = \prod_{i=1}^{n} \int f(\mathbf{Y}_i|\boldsymbol{\theta}, \mathbf{Z}_i)\, \pi(\mathbf{Z}_i)\, d\mathbf{Z}_i\,. \tag{21}$$

The integrals with respect to the subject specific latent variables $\mathbf{Z}_i$ in (21) can be approximated with fixed Gauss-Hermite quadrature points (used to calculate each $f(\mathbf{Y}_i|\boldsymbol{\theta})$ in equation 21). Other more accurate approximations can be also used, such as the adaptive quadrature points ([27], [28]) or Laplace approximations [29].

The Monte Carlo error (MCE) of the $\log \widehat{\mathcal{L}}$ was estimated using the method of batch means ([30], [31]). The simulated sample was divided into 30 batches and the marginal log-likelihood was estimated via (21) at each batch. The mean over all batches, denoted by $\overline{\log \mathcal{L}}$, is referred to as the batch mean estimator, while the the standard deviation of the log-marginal likelihood estimator over the different batches is considered as its MCE estimate. The same procedure was repeated using three alternative measures of central location of the posterior distribution (the componentwise posterior mean, median and mode) as $\boldsymbol{\theta}^*$.

Moreover, the Laplace-Metropolis estimator (LM) proposed by [11] was used as benchmark method. The Laplace-Metropolis method was implemented on the posterior $\pi(\boldsymbol{\theta}|\mathbf{Y})$, therefore, the vector of the latent variables $\mathbf{Z}$ was marginalized out. The normal approximation used in the Laplace method was applied to the original parameters for all α_j and $\beta_{j\ell}$, with $j < \ell$, and on the $\log \beta_{jj}$ for $j = 1, \ldots, k$ for the diagonal loadings. For the latter, we have used the logarithms instead of the original parameters in order to avoid asymmetries caused by their positivity constraint and, by this way, to achieve a well behaved approximation of the marginal likelihood.

5.2 Tuning M and R

We initially use a dataset generated from a one-factor model with 4 binary items and 400 individuals ($p = 4$, $N = 400$ and $k = 1$ respectively, that is 408 unknown parameters). We use this rather restricted example in order to examine the convergence of the estimator as a function of the number of M and R values generated from the proposal and the posterior densities, respectively. Specifically, 300,000 posterior observations were generated after discarding additional 10,000 iterations as a burn in period from a Metropolis-Hastings, within a Gibbs, algorithm. A thinning interval of 10 iterations was additionally considered in order to diminish autocorrelations, leaving a total of 30,000 values available for posterior analysis. All simulations were conducted using R version 2.12 on a quad core i5 Central Processor Unit (CPU), at 3.2GHz and with 4GB of RAM.

Before dividing the simulated sample into batches, we have graphically examined the convergence of the estimator by changing

a) M, that is, the number of points generated from the proposal density $q(\boldsymbol{\theta}, \boldsymbol{\theta}^* | \mathbf{Y}, \mathbf{Z})$ used for the estimation of the denominator in (15),

b) R, that is, the number of points generated from the posterior $\pi(\boldsymbol{\theta}, \mathbf{Z}|Y)$ that are required for the computation of $\widehat{\mathcal{L}}$ within each batch.

We initially focused on (a), with M ranging from 100 to 2000, and kept R fixed at 1000 iterations. Figure 1(a) illustrates that all versions of $\log \widehat{\mathcal{L}}$ were stabilized up to a decimal point, even for $M \geq 40$. Time increased linearly, with M varying from 0.5 to 4.7 mins, which is approximately one minute increment per 25 generated values.

Regarding (b), the ergodic estimator was computed with R taking values from 100 to 2000 and $M = 50$, which seem more than sufficient according to Figure 1(a). The ergodic estimators of all versions of $\log \widehat{\mathcal{L}}$ for each selected R are depicted in Figure 1(b). The estimates were close and stable for $R \geq 500$. The CPU time was also increased linearly from 0.5 to 9 mins at the cost of half a minute per 100 additional iterations.

Based on Figure 1, we proceeded with thirty batches of size $R = 1000$ and $M = 50$ to ensure convergence of the estimates. Figure 2 presents the marginal likelihood estimates based on CJ and LM using the posterior mean, median and mode as points of central location. When using the posterior mean, LM was found to be equal to -977.76, while $\overline{\log \mathcal{L}}$ was equal to -977.73, with the estimated MCE being equal to 0.026. The estimators are quite robust, regardless of the choice of the posterior point of central location. Specifically, the LM was -977.65 at the median and -977.71 at the mode. Similarly, the $\log \widehat{\mathcal{L}}$ was -977.77 at the median and -977.75 at the mode, with equivalent MCEs (0.020 and 0.022 respectively).

In the next section we proceed with more realistic illustrations, using both simulated and real data sets. In all the examples which follow, the same tuning procedure was followed but it is not reported for brevity.

5.3 Computation of Bayes Factor: simulated examples

Here we demonstrate the performance of the CJ estimator using the output from a single run of a multi-block Metropolis-within-Gibbs algorithm, in three simulated datasets of larger size, allowing, in addition, for the models to be fitted with multiple factors of higher dimension. We consider the datasets with the following settings:

a) $N = 600$ observations with $p = 6$ items generated from a $k = 1$ factor model

b) $N = 600$ observations with $p = 6$ items generated from a $k = 2$ factor model

c) $N = 800$ observations with $p = 7$ items generated from a $k = 3$ factor model

All model parameters were selected randomly from a uniform distribution, $U(-2, 2)$. The number of unknown parameters for the posterior ordinate in (15) is equal to $k(p + N) + p$, corresponding to 606, 1218 and 2428 parameters, respectively, for each of the three situations described above. Models that either overestimate or underestimate k were also considered, this time evaluating the Bayes factor in favour of the true generating model. Using the same procedure as in Section 5.2, we have concluded that it is sufficient to select 30 batches of 1000, 2000 and 3000 iterations for the one, two and three-factor models, respectively. All estimators were evaluated at the componentwise posterior median (that is, $\boldsymbol{\theta}^*$=posterior median).

The LM estimate of the marginal likelihood is reported as a gold standard using an MCMC output of 30,000 iterations, while the $\log \widehat{\mathcal{L}}$ refers to the estimate of the first batch (of 1,000 iterations). The batch mean estimator and the corresponding error were calculated as described in Section 5. The results in Table 1 suggest that estimates based on the independence CJ method, proposed in Section 4.1.1, are similar to the ones of the benchmark method (LM), even from the first batch. Moreover, the Monte Carlo error of the $\log \widehat{\mathcal{L}}$ is fairly small but naturally gets higher as the number of unknown parameters in the posterior ordinate increase for a fixed number of iterations. Nevertheless, this Monte Carlo error can be efficiently reduced by increasing the number of MCMC iterations.

In addition, the one-block (OB) M-H approach described in Section 4.1.2 was implemented for the second data set (b). The batch mean and the corresponding error were computed over 30 batches, as in the case of the multi-block design. In the case where one factor was assumed, the batch mean was $\overline{\log \mathcal{L}}_{OB} = -2200.68$, with $MCE(\log \widehat{\mathcal{L}}_{OB}) = 1.98$. In the case where two factors were assumed, the batch mean was $\overline{\log \mathcal{L}}_{OB} = -2066.23$, with $MCE(\log \widehat{\mathcal{L}}_{OB}) = 3.11$. In both cases the estimated log-marginal is far away from the corresponding ones reported in Table 1 using the LM and the independence CJ estimator. Moreover, under the one-block design, the estimated MCE was 60 and 47 times as high as the corresponding values under the more efficient multi-block design, presented in Table 1. It is therefore verified that between the two single-run approaches, the independence CJ estimator is more efficient and accurate than the one-block CJ estimator.

With regards to the BF, the estimates (in log scale) reported in Table 2 are based on the marginal likelihood estimates presented in Table 1. In all three simulated datasets, the estimated Bayes factors $\widehat{BF}$ indicated the true model. Moreover, when the independence CJ was used, the true model was suggested by the BF estimator at every batch. Bayes factors for the second and the third dataset clearly indicate the true model, with values ranging from e^{33} to e^{116}. Only in the first dataset is the Bayes factor much lower and equal to $e^3 \approx 20$. In the latter case, or in more extreme cases where two competing models have Bayes factors close to one, the Monte Carlo error should be small enough in order to be able to identify which model is a-posteriori supported. Here we estimated an error equal to 0.25, with 95% of the estimates ranging between $e^{2.5} = 12.2$ and $e^{3.3} = 27.1$. Hence, the independence CJ method infers safely in favour of the true generating mechanism, providing BF estimates similar to the ones obtained from the gold standard of the LM, in all cases.

5.4 Illustration on real data

We proceed with two real-data examples also analyzed in Bartholomew et al. [2, chapter 8]. In all examples the marginal likelihood was estimated via CJ^{I} and LM methods at the median point, over samples of 10 thousand iterations (after discarding 1000 iterations as a burn in period and keeping 1 every 10 iterations to reduce autocorrelations).

The first data set is originally provided by [32] and is part of the Law School Admission Test (LSAT) completed by $N = 1005$ individuals. The test consists of five items and was designed to measure one latent factor which is also supported by the computed Bayes factor (≈ 0.22 and 0.24 for LM and CJ^I based estimators, respectively; posterior weight of one-factor model 0.802 and 0.817 respectively) reported in the first row of Table 3. In particular, the BF of the one-factor versus the two-factor model was less than 0.5 and therefore according to [8] the evidence against the unidimensional model "do not worth more than a bare mention".

The second data set is part of the 1990 Workplace Industrial Relations Survey (WIRS, [33]). The Bayes factor of the two versus the one-factor model clearly supports the latter ($\log BF_{21} \approx 69$); see second line of Table 3. As further analysis, Bartholomew et al. [2] suggested to omit the most poorly fitted item (here item 1) of the scale in order to improve the fit of the one-factor model. The analysis was replicated for the remaining 5 items to suggest again the two-factor model as the preferred model ($BF_{21} = 40$, that corresponds to "decisive evidence" against the one-factor model, [8]). To summarize, simulations and real-data analysis suggest that the independence CJ estimator succeeds to detect the true model, provides similar estimates to the benchmark method (LM) and has an acceptable MCMC error.

6 Closing remarks

The paper focused on the CJ [12] marginal likelihood estimator for latent variable models. In the popular case where the likelihood expression embodies local independence, conditional on the latent vector, it was illustrated that the CJ estimator can be computed in a single run of a Metropolis-within-Gibbs algorithm. This approach drastically reduces the computational effort required for the marginal likelihood estimate. Under conditional independence, the dimensionality of the model is no longer an aspect of the CJ [12] estimator. Hence, this strategy can be implemented to reduce the computational time even in models with no latent variables. That is in models where the likelihood can be augmented using auxiliary variables ([17],[34]) to introduce likelihood local independence.

Two more additional points are discussed: (a) the differences of the proposed simplified CJ estimator from the (trivial) single-run CJ estimator obtained from one-block Metropolis-Hastings samplers and (b) how we can use the Metropolis kernel to integrate out the latent variables when no posterior ordinate is analytically available.

The points outlined in this article simplify the implementation of the CJ method on specific cases making a method, which is accurate and already established in bibliography, easier to use and more efficient in practice.

Acknowledgements

This research was funded by the Research Centre of the Athens University of Economics and Business.

References

[1] I. Moustaki and M. Knott. Generalized latent trait models. *Psychometrika*, 65:391–411, 2000.

[2] D. J. Bartholomew, F. Steele, I. Moustaki, and J. Galbraith. *Analysis of Multivariate Social Science Data*. Chapman & Hall/CRC, 2nd edition, 2008.

[3] S. Vitoratou, I. Ntzoufras, N. Smyrnis, and C. N. Stefanis. Factorial composition of the aggression questionnaire: a multi-sample study in greek adults. *Psychiatry Research*, 168(410):32–39, 2009.

[4] I. Moustaki and F. Steele. Latent variable models for mixed categorical and survival responses with an application to fertility preferences and family planning in bangladesh. *Statistical modelling*, 5(4):327–342, December 2005.

[5] I. Moustaki and I. Papageorgiou. Latent class models for mixed variables with applications in archaeometry. *Computational Statistics and Data Analysis*, 48(3):659–675, 2005.

[6] D.J. Bartholomew, M. Knott, and I. Moustaki. *Latent variable models and factor analysis: a unified approach*. Wiley Series on Probability and Statistics. John Wiley and Sons, Ltd, London, 3rd edition, 2011.

[7] R. Patz and B. A. Junker. A straightforward approach to Markov chain Monte Carlo methods for item response models. *Journal of Educational and Behavioral Statistics*, 24: 146–178, 1999.

[8] R.E. Kass and A.E Raftery. Bayes factors. *Journal of the American Statistical Association*, 90:773–795, 1995.

[9] S. Chib. Marginal likelihood from the Gibbs output. *Journal of the American Statistical Association*, 90:1313–1321, 1995.

[10] X.L. Meng and W.H. Wong. Simulating ratios of normalizing constants via a simple identity: A theoretical exploration. *Statistica Sinica*, 6:831–860, 1996.

[11] S.M. Lewis and A.E. Raftery. Estimating Bayes factors via posterior simulation with the Laplace Metropolis estimator. *Journal of the American Statistical Association*, 92: 648–655, 1997.

[12] S. Chib and I. Jeliazkov. Marginal likelihood from the Metropolis-Hastings output. *Journal of the American Statistical Association*, 96:270–281, 2001.

[13] N. Friel and A.N. Pettit. Marginal likelihood estimation via power posteriors. *Journal of Royal Statistical Society*, 770:589–607, 2008.

[14] M.A. Newton and A.E. Raftery. Approximate Bayesian inference with the weighted likelihood bootstrap. *Journal of the Royal Statistical Society*, 56:3–48, 1994.

[15] D.J. Nott, R. Kohn, and M. Fielding. Approximating the marginal likelihood using copula. *arXiv:0810.5474v1*, 2008. Available at `http://arxiv.org/abs/0810.5474v1` .

[16] A. E. Gelfand and A. F. M. Smith. Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85(410):398–409, 1990.

[17] M. A. Tanner and W. H. Wong. The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*, 82:528–540, 1987.

[18] A. Skrondal and S. Rabe-Hesketh. *Generalized Latent Variable Modeling: Multilevel, Longitudinal and Structural Equation Models*. Chapman & Hall/CRC, Boca Raton, FL, 2004.

[19] J. Geweke and G. Zhou. Measuring the pricing error of the arbitrage pricing theory. *Review of Financial Studies*, 9:557–87, 1996.

[20] O. Aguilar and M. West. Bayesian dynamic factor models and portfolio allocation. *Journal of Business and Economic Statistics*, 18:338–357, 2000.

[21] H. F. Lopes and M. West. Bayesian model assessment in factor analysis. *Statistica Sinica*, 14:4167, 2004.

[22] J. Besag. A candidate's formula: A curious result in Bayesian prediction. *Biometrika*, 76:183, 1989.

[23] S. Chib and I. Jeliazkov. Inference in semiparametric dynamic models for binary longitudinal data. *Journal of the American Statistical Association*, 101:685–700, 2006.

[24] W.R. Gilks, S. Richardson, and D.J. Spiegelhalter. Introducing markov chan monte carlo. In S. Richardson W.R.Gilks and D.J. Spiegelhalter, editors, *Markov Chain Monte Carlo in Practice*, pages 165–187. London: Chapman and Hall, 1996.

[25] I. Ntzoufras, P. Dellaportas, and J. Forster. Bayesian variable and link determination for generalised linear models. *Journal of Statistical Planning and Inference*, 111:165–180, 2000.

[26] D. Fouskakis, I. Ntzoufras, and D. Draper. Bayesian variable selection using cost-adjusted BIC, with application to cost-effective measurement of quality of health care. *Annals of Applied Statistics*, 3:663–690, 2009.

[27] S. Rabe-Hesketh, A. Skrondal, and A. Pickles. Maximum likelihood estimation of limited and discrete dependent variable models with nested random effects. *Journal of Econometrics*, 128:301–323, 2005.

[28] S. Schilling and R.D Bock. High-dimensional maximum marginal likelihood item factor analysis by adaptive quadrature. *Psychometrika*, 70:533–555, 2005.

[29] P. Huber, E. Ronchetti, and M. P. Victoria-Feser. Estimation of generalized linear latent variable models. *Journal of the Royal Statistical Society, Series B*, 66:893–908, 2004.

[30] B. W. Schmeiser. Batch size effects in the analysis of simulation output. *Operations Research*, 30:556–568, 1982.

[31] P. Bratley, B. L. Fox, and L. E. Schrage. *A Guide to Simulation.* New York: Springer-Verlag, 1987.

[32] R. D. Bock and M. Lieberman. Fitting a response model for n dichotomously scored items. *Psychometrika*, 35:179–197, 1970.

[33] C. Airey, N. Tremlett, and R. Hamilton. The workplace industrial relations survey 1990. *Technical Report (Main and Panel Surveys)*, Social and Community Planning Research, 1992.

[34] D. A. van Dyk and X. L. Meng. The art of data augmentation. *Journal of Computational and Graphical Statistics*, 10:1–50, 2001.

Tables

Table 1: Simulated results: marginal likelihood estimates in Section 5.3.

Dataset	p	N	k_{true}	k_{model}	$\log LM$	$\log \widehat{\mathcal{L}}$	$\overline{\log \mathcal{L}}$	$MCE(\log \widehat{\mathcal{L}})$
(a)	6	600	1	1	-2175.3	-2175.2	-2175.1	0.016
				2	-2178.2	-2178.2	-2178.2	0.253
(b)	6	600	2	1	-2187.2	-2187.6	-2187.5	0.033
				2	-2070.8	-2071.3	-2071.2	0.066
(c)	7	800	3	1	-3422.4	-3422.3	-3422.5	0.029
				2	-3374.4	-3374.1	-3375.2	0.133
				3	-3341.3	-3339.1	-3339.3	0.332

p: number of items; N: number of individuals; k_{true} and k_{model}: number of factors in the true and evaluated model, respectively; LM and $\widehat{\mathcal{L}}$: Laplace-Metropolis and Chib and Jeliazkov estimates of the marginal likelihood; $\overline{\log \mathcal{L}}$: Batch mean estimator of the log-marginal likelihood; $MCE(\log \widehat{\mathcal{L}})$: Monte Carlo error of the $\log \widehat{\mathcal{L}}$ obtained as the standard deviation of 30 batches of equal size as the estimate reported in 7th column of the table.

Table 2: Simulated results: Bayes Factor estimates in Section 5.3.

Dataset details				Comparison	$\log \widehat{BF}$		Batch summaries of $\log \widehat{\mathcal{L}}$			
#	p	N	k_{true}	k_1 vs. k_2	LM	CJ	Mean	S.D.	1^{st}Q	3^{rd}Q
(a)	6	600	1	$1-2$	3.1	3.0	3.1	0.25	2.5	3.3
(b)	6	600	2	$2-1$	116.3	116.3	116.3	0.08	116.0	116.5
(c)	7	800	3	$3-1$	81.1	83.3	83.2	0.33	81.5	84.5
				$3-2$	33.3	35.0	35.9	0.35	34.3	37.7

p: number of items; N: number of individuals; k_{true}: number of factors in the true model; k_1 vs. k_2: the Bayes factor comparing the k_1 versus the k_2-factor model is estimated; $\widehat{BF}$: Estimated Bayes factors based on Laplace-Metropolis (LM) and Chib and Jeliazkov (CJ) estimates of the marginal likelihood; Batch summaries of $\log \widehat{\mathcal{L}}$: Summaries based on 30 batches of $\log \widehat{\mathcal{L}}$ (mean=Batch mean estimate, S.D.= standard deviation - provides an estimate for the Monte Carlo Error, 1^{st}Q and 3^{rd}Q: first and third quartiles).

Table 3: Marginal Likelihood and Bayes Factor for the real data: LSAT and WIRS

	$\log LM$			$\log \widehat{\mathcal{L}}$		
Dataset	1-factor	2-factor	$\log \widehat{BF}_{21}^{(LM)}$	1-factor	2-factor	$\log \widehat{BF}_{21}^{(CJ)}$
1. LSAT	-2494.8	-2496.2	-1.4	-2495.1	-2496.6	-1.5
2. WIRS-6 items	-3456.1	-3387.1	69.0	3456.2	-3387.3	68.9
3. WIRS-5 items	-2786.6	-2782.8	3.8	-2786.8	-2783.1	3.7

LM and $\widehat{\mathcal{L}}$: Laplace-Metropolis and Chib and Jeliazkov estimates of the marginal likelihood; 1-factor and 2-factor columns: estimates of the log-marginal likelihood for the 1-factor and 2-factor models, respectively; $\widehat{BF}_{21}^{(LM)}$ and $\widehat{BF}_{21}^{(CJ)}$: Estimated Bayes factors of 2-factor versus 1-factor model based on LM and $\widehat{\mathcal{L}}$, respectively.

Figure labels

Figure 1

Main caption

Ergodic $\log \widehat{\mathcal{L}}$ using three posterior measures of central location (mean, median and mode) for different M (number of values generated from the proposal) and for different R (number of MCMC iterations); p=4 items, $N = 400$ individuals and k=1 latent factor.

Sub captions

Figure 1(a): Sensitivity of $\log \widehat{\mathcal{L}}$ (based on CJ^I) on different M with $R = 1000$.

Figure 1(b): Sensitivity of $\log \widehat{\mathcal{L}}$ (based on CJ^I) on different R with $M = 50$.

Figure 2

Main caption

Log-likelihood estimated via CJ^I (dotted line) over 30 batches of size R=1000 compared with the corresponding Laplace-Metropolis estimate (solid line) using MCMC output of 30,000 iterations and the posterior median, mean or mode as measures of central location; p=4 items, $N = 400$ individuals and $k = 1$ factor.