Silia Vitoratou, Ioannis Ntzoufras and Irini Moustaki

Explaining the behavior of joint and marginal Monte Carlo estimators in latent variable models with independence assumptions

Article (Accepted version)
(Refereed)

Original citation: Vitoratou, Silia, Ntzoufras, Ioannis and Moustaki, Irini (2016) *Explaining the behavior of joint and marginal Monte Carlo estimators in latent variable models with independence assumptions.* Statistics and Computing, 26 (1). pp. 333-348. ISSN 0960-3174

DOI: 10.1007/s11222-014-9495-8

Explaining the behavior of joint and marginal Monte Carlo estimators in latent variable models with independence assumptions.

Silia Vitoratou, Ioannis Ntzoufras and Irini Moustaki.

Abstract

In latent variable models parameter estimation can be implemented by using the joint or the marginal likelihood, based on independence or conditional independence assumptions. The same dilemma occurs within the Bayesian framework with respect to the estimation of the Bayesian marginal (or integrated) likelihood, which is the main tool for model comparison and averaging. In most cases, the Bayesian marginal likelihood is a high dimensional integral that cannot be computed analytically and a plethora of methods based on Monte Carlo integration (MCI) are used for its estimation. In this work, it is shown that the joint MCI approach makes subtle use of the properties of the adopted model, leading to increased error and bias in finite settings. The sources and the components of the error associated with estimators under the two approaches are identified here and provided in exact forms. Additionally, the effect of the sample covariation on the Monte Carlo estimators is examined. In particular, even under independence assumptions the sample covariance will be close to (but not exactly) zero which surprisingly has a severe effect on the estimated values and their variability. To address this problem, an index of the sample's divergence from independence is introduced as a multivariate extension of covariance. The implications addressed here are important in the majority of practical problems appearing in Bayesian inference of multi-parameter models with analogous structures.

1 Introduction

Latent variable models are widely used to capture latent constructs by means of multiple observed indicators (items). From the early readings, the methods applied for the parameter estimation of model settings with latent variables relied either on the joint (Lord and Novick, 1968; Lord, 1980) or the marginal likelihood (Bock and Lieberman, 1970; Bock and Aitkin, 1981). The former suggests to estimate the observed and latent variable scores simultaneously while the latter to marginalize out the latent variables prior to the model parameter estimation. Similarly, counterpart approaches have been developed within the Bayesian context (for instance Mislevy, 1986; Gifford and Swaminathan, 1990; Kim et al., 1994; Baker, 1998; Patz and Junker, 1999).

The Bayes factors, posterior model probabilities and the corresponding odds (Kass and Raftery, 1995) require the computation of the Bayesian marginal (or integrated) likelihood which is defined as the expectation of the likelihood over the prior distribution. To separate from the marginal likelihood term used in the context of latent variables, we will refer to this as the Bayesian marginal likelihood (BML). In most cases the BML is a high dimensional integral which is not analytically tractable. Sophisticated Monte Carlo techniques have been developed throughout the years, such as the bridge sampling (Meng and Wong, 1996) and the Laplace-Metropolis estimator (Lewis and Raftery, 1997), among others. Despite the method implemented however, the BML can be estimated by considering either the joint or the marginal likelihood expressions.

Intuitively one expects the joint approach to be less efficient, especially as the number of dimensions increases. Empirical findings are available in the literature to verify this fact (see for instance Kim et al., 1994 and references within). However, to our knowledge, there are no theoretical justifications to illustrate the differences between the two approaches in terms of error. On the other hand, the use of the joint expression can be directly implemented through ready-to-use Markov chains Monte Carlo (MCMC) software such as WinBUGS. This favorites the full augmented scheme and the corresponding joint estimator, either for actual data analysis or for illustration purposes, as for instance in Chib and Jeliazkov (2001, Sec. 3.2 and 3.3), Friel and Pettitt (2008, Sec. 4.2) and Congdon (2010, Sec. 2.2.3). In this work, we obtain analytical expressions for the variances associated with the estimator of each approach and we consider the factors that influence their associated Monte Carlo error (MCE). In particular, we illustrate graphically and mathematically that even though the MCE is not by definition associated directly with the dimensionality of a model, the latter plays a key role through the variance components. In turn, the variance components are directly influenced by the number of the variables involved and their variability. Additionally, we demonstrate the effect of the sample covariation on the Monte Carlo estimates, which is considerably understated in the literature. In particular, for independent random variables the sample covariance is typically close but not exactly equal to zero. Here, we illustrate that, in high dimensions, even small sample covariances influence the estimators producing biased Monte Carlo estimates. This bias usually remains undetected, due to the fact that the effect of sample covariation also causes underestimation of the corresponding MCEs.

Concerns arise with respect to convergence, since the extensive use of simulation methods nowadays is not always followed by the necessary precautions to ensure accurate estimation of the quantity of interest. For instance, Koehler et al. (2009) reported that in a large number of articles with simulation studies, only a tiny proportion provided either a formal justification of the number of replications implemented or the actual estimate of the MCE. That is, integral approximations are based on an arbitrary number of replications, that are considered to be "large enough" to accurately estimate the quantity of interest. Nevertheless, in complex high dimensional problems, where the rate of convergence can be extremely low, millions of iterations may be required to achieve a desirable level of precision for the MC estimate of interest. Hence, in many cases the simulations are practically stopped "when patience runs out", as Jones et al. (2006) fluently describe. The remarks that are made in this paper facilitate the understanding of the error and bias mechanism of Monte Carlo methods under independence and conditional independence and hopefully will assist the researchers to accurately estimate the quantity of interest in high dimensions.

The findings of this article have applicability in a wide range of modern statistical models including factor analysis models, latent class models, random effects models, longitudinal models, cluster analysis models, data-augmented specific models, mixture models, negative binomial response models, zero-inflated models, bivariate Poisson models and Poisson difference (or Skellam distributed) models.

The structure of the paper is as follows. Section 2 presents a motivating example with regard to the estimation of the BML in a model with latent variables. Three popular MCMC methods are implemented, under both joint and marginal approaches. Key observations are made based on the comparison of the derived estimated values which motivate further research. Section 3 presents the Monte Carlo integration under the joint and marginal settings, with emphasis on high dimensional integrals where independence can be assumed for the integrand. The MCEs under both approaches are derived in Section 3.1 while the factors that affect the error are considered in Section 3.2. For illustration purposes a simple example is provided, that is, estimating the mean of the product of independent and identically distributed (i.i.d) *Beta* random variables. In Section 3.3, the variance reduction in the case of conditional independence is discussed. In Section 3.4 the total covariation of N variables is defined as a multivariate counterpart of covariance. A corresponding index that measures the sample's divergence from independence is developed and employed to amplify the factors that influence the total sample covariation. Finally, it is shown that in finite settings where the sample covariation is non zero, the MCE associated with the joint approach is underestimated.

2 A motivating example: BML estimation in generalised linear latent trait models

A broad and popular family of models that can handle continuous, discrete and categorical observed variables are the generalised linear latent variable models (GLLVM, Bartholomew et al., 2011). Due to GLLVM's versatile applicability, they are utilized in this section to amplify the difference between the joint and marginal likelihood approaches. In particular, we focus on a latent trait model (Moustaki and Knott, 2000) with binary observed items, under the Bayesian paradigm. The BML is computed in a simulated data set under the joint and marginal approaches. The derived estimations raise specific concerns which are discussed at the end of the section.

2.1 Model setting and estimation techniques

The GLLVM consist of four main components: (a) the multivariate random component $Y_i = (Y_{i1}, Y_{i2}, \ldots, Y_{ip})^T$ of the p response variables of subject i, $(i = 1, \ldots, N)$, (b) a set of k-dimensional latent vectors $\boldsymbol{Z}_i = (Z_{i1}, \ldots, Z_{ik})$ characterizing subject i, (c) the linear predictor $\boldsymbol{\eta}_i = (\eta_{i1}, \ldots, \eta_{ip})^T$ of the latent variables $\boldsymbol{Z}_i$ for subject i and (d) the link function $\upsilon(\cdot)$, that connects the previous three components. Hence, a GLLVM can be summarized as

$$Y_{ij}\,|\mathbf{Z}_i \sim ExpF, \quad \boldsymbol{\eta}_i = \boldsymbol{\alpha} + \boldsymbol{\beta}\boldsymbol{Z}_i, \quad \boldsymbol{Z}_i \sim \pi(\boldsymbol{Z}_i) \;\text{ and }\; \upsilon\big[E(Y_i|\mathbf{Z}_i\,)\big] = \boldsymbol{\eta}_i, \quad \text{for } \; i = 1, \ldots N,$$

where $ExpF$ is a member of the exponential family, $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_p)^T$, $\boldsymbol{\beta} = (\beta_{j\ell}\,; j = 1, \ldots, p, \ell = 1, \ldots, k)$ and $\boldsymbol{Z}_i \sim \pi(\boldsymbol{Z}_i)$ denotes that $\boldsymbol{Z}_i$ is random variable with density $\pi(\boldsymbol{Z}_i)$. In the above formulation, $\pi(\mathbf{Z}_i)$ needs to be specified for the latent variables. Typically, the latent variables are assumed to be a-priori distributed as independent standard normal distributions, that is, $\boldsymbol{Z}_i \sim N(\mathbf{0}, \boldsymbol{I}_k)$ for all individuals (Bartholomew et al., 2011), where $\boldsymbol{I}_k$ is the identity matrix of dimension $k \times k$.

In the following, we focus on models with binary responses and k latent vectors, which belong to the family of generalized latent trait models discussed in Moustaki and Knott (2000). The logistic model is used for the response probabilities:

$$\text{logit}\big[P_{ij}\,(\boldsymbol{Z}_i)\big] = \alpha_j + \sum_{\ell\,=1}^{k} \beta_{j\ell}\, Z_{i\ell}, \;\; i = 1, \ldots, N, \;\; j = 1, \ldots p,$$

where $P_{ij}(\boldsymbol{Z}_i)$ is the conditional probability of a positive response by the individual i to item j. The model assumes that for each subject i the responses are independent given the latent vector $\boldsymbol{Z}_i$ (*local independence* assumption) leading to either the joint (Lord and Novick, 1968; Lord, 1980)

$$f(\mathbf{Y}|\,\boldsymbol{\theta}, \mathbf{Z}) = \prod_{i=1}^{N} f(Y_i|\,\boldsymbol{\theta}, \boldsymbol{Z}_i) = \prod_{i=1}^{N}\prod_{j=1}^{p} P_{ij}\,(\boldsymbol{Z}_i)^{y_{ij}}\,\big[1 - P_{ij}\,(\boldsymbol{Z}_i)\big]^{(1-y_{ij})} \tag{1}$$

or the marginal likelihood (Bock and Lieberman, 1970; Bock and Aitkin, 1981; Moustaki and Knott, 2000)

$$f(\mathbf{Y}|\,\boldsymbol{\theta}) = \prod_{i=1}^{N} \int \prod_{j=1}^{p} P_{ij}\left(\boldsymbol{Z}_i\right)^{y_{ij}} \left[1 - P_{ij}\left(\boldsymbol{Z}_i\right)\right]^{(1-y_{ij})} d\,\boldsymbol{Z}_i, \tag{2}$$

where $\boldsymbol{\theta} = (\boldsymbol{\alpha}, \boldsymbol{\beta})$.

For the Bayesian counterpart of the model, priors distributions are additionally assigned on model parameters $\boldsymbol{\theta}$. The prior specification of the model used here is based on the ideas presented by Ntzoufras et al. (2003) and further explored in the context of generalized linear models by Fouskakis et al. (2009, equation 6). For a GLLVM with binary responses, this prior corresponds to a $N(0, 4)$. In the case of $k > 1$ latent variables, constraints need to be imposed on the loadings $\boldsymbol{\beta}$ to ensure identification of the model. To achieve a unique solution, the loadings matrix is constrained to be a full rank lower triangular matrix (see also Geweke and Zhou, 1996, Aguilar and West, 2000 and Lopes and West, 2004), by setting $\beta_{j\ell} = 0$ for all $j < \ell$ and $\beta_{\ell\ell} > 0$. The prior is summarized as follows:

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

where $X \sim LN(0, 1)$ is the log-normal distribution with zero mean and the variance equal to one for $\log X$. For diagonal elements $\beta_{\ell\ell}$, the $LN(0, 1)$ was selected as a prior in order to approximately match the prior standard deviation used for the rest of the parameters. Moreover, this is one of the default prior choices for such parameters in the relevant literature; see for example in Kang and Cohen (2007) and references therein.

In analogy with (1) and (2), under the local independence assumption there are two equivalent formulations of the BML, namely

$$f(\mathbf{Y}) = \int \prod_{i=1}^{N} f(Y_i|\,\boldsymbol{\theta}, \boldsymbol{Z}_i)\, \pi(\boldsymbol{\theta}, \mathbf{Z})\, d(\boldsymbol{\theta}, \mathbf{Z}) \tag{3}$$

and

$$f(\mathbf{Y}) = \int f(\boldsymbol{Y}\,|\,\boldsymbol{\theta})\, \pi(\boldsymbol{\theta})\, d\boldsymbol{\theta} = \int \left[\prod_{i=1}^{N} \int f(Y_i|\,\boldsymbol{\theta}, \boldsymbol{Z}_i)\, \pi(\boldsymbol{Z}_i)\, d\boldsymbol{Z}_i\right] \pi(\boldsymbol{\theta})\, d\boldsymbol{\theta}\ . \tag{4}$$

Hereafter we refer to (3) with the term joint approach and to (4) with the term marginal approach for the BML and we compare them within the Bayesian framework.

For both approaches, we employ three popular BML estimators namely: the reciprocal mean estimator (*RM*; Gelfand and Dey 1994), the bridge harmonic estimator (*BH*; Meng and Wong 1996, often refer to as the generalised harmonic mean) and the bridge geometric estimator (BG; Meng and Wong 1996). The identities that correspond to these estimators are provided in the Appendix. In order to construct the estimators using the joint approach, the parameter vector is augmented to include the latent variables, that is $\boldsymbol{\vartheta} = \{\boldsymbol{\theta}, \mathbf{Z}\} = \{\boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{Z_1}, \ldots, \mathbf{Z_N}\}$, while for the marginal approach it holds $\boldsymbol{\vartheta} = \boldsymbol{\theta} = \{\boldsymbol{\alpha}, \boldsymbol{\beta}\}$.

The estimators require also an *importance* function $g(\boldsymbol{\vartheta})$. The objective and recommendation of many authors (DiCiccio et al., 1997; Gelman and Meng, 1998; Meng and Schilling, 2002; Meng and Wong, 1996), is to choose a density similar to the target distribution (here the posterior). In the current example, we use an approximation based on the posterior moments for each parameter, with structure $g(\boldsymbol{\theta}) = g(\boldsymbol{\alpha})g(\boldsymbol{\beta}_e)$ where

$$g(\boldsymbol{\alpha}) \sim MN(\widetilde{\mathbf{m}}_{\boldsymbol{\alpha}}, \widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{\alpha}}) \;\text{ and }\; g(\boldsymbol{\beta}_e) \sim MN(\widetilde{\mathbf{m}}_{\boldsymbol{\beta}_e}, \widetilde{\boldsymbol{\Sigma}}_{\beta_e}),\, \boldsymbol{\beta}_e = \beta_{j\ell},\; j \geq \ell$$

and $\boldsymbol{\beta}_e$ refers to the non-zero components of $\boldsymbol{\beta}$ with elements $\log \beta_{\ell\ell}$ for $j = 1, \ldots, p$ and $\beta_{j\ell}$ for $j > \ell$. The $MN(\widetilde{\mathbf{m}}, \widetilde{\boldsymbol{\Sigma}})$ denotes a multivariate normal distribution whose parameters $(\widetilde{\mathbf{m}}, \widetilde{\boldsymbol{\Sigma}})$ are the posterior mean and variance-covariance matrix estimated from the MCMC output. For the joint approach, the $g(\boldsymbol{\vartheta})$ is simply augmented for the latent vector

$$g(\boldsymbol{\vartheta}) = g(\boldsymbol{\alpha})g(\boldsymbol{\beta}_e) \prod_{i=1}^{N} \prod_{\ell=1}^{k} g(Z_{i\ell}),$$

where $g(Z_{i\ell}) \sim N(\tilde{m}_{Z_{i\ell}}, \tilde{s}^2_{Z_{i\ell}})$, with parameters estimated from the MCMC output used to approximate the posterior $\pi(Z_{i\ell}|\boldsymbol{Y})$.

2.2 Simulation study

A simulated data set with $p = 6$ items, $N = 600$ cases and $k = 2$ factors is firstly considered. The model parameters were selected randomly from a uniform distribution U(-2,2). Using a Metropolis within Gibbs algorithm, 50,000 posterior observations were obtained after discarding a period of 10,000 iterations and considering a thinning interval of 10 iterations to diminish autocorrelations. The posterior moments involved in the construction of the importance function were estimated from the final output and an additional sample of equal size was generated from $g(\boldsymbol{\theta})$. The MCMC estimators were computed in two versions, joint and marginal, using the entire MCMC output of 50,000 iterations. In a second step, the simulated sample was divided into 50 batches (of 1,000 iterations) and the integrated log-likelihood was estimated at each batch. The standard deviation of the log-BML estimators over the different batches is considered here as its MCE estimate (Schmeiser 1982, Bratley et al. 1987, Carlin and Louis 2000).

The batch mean method is the most frequently met approach for the estimation of the Monte Carlo error of the marginal likelihood estimates. The method is considered reliable and it can be implemented in a straight-forward manner. However, motivated by a referee's comment, we have also implemented the overlapping batch means method (Meketon and

Schmeiser, 1984) which is proven to be strongly consistency according to Flegal and Jones (2010). The two methods gave identical results both in terms of the estimated marginal likelihood and in terms of MCEs. Therefore, only the results based on the batch mean approach are presented in the following.

The BML (4) was calculated by approximating the inner integrals with fixed Gauss-Hermite quadrature points. This way, the computational burden is considerably reduced without compromising the accuracy, since such approximations are fairly precise in low dimensions. For instance, in the current example where the inner integrals are two-dimensional ($k = 2$), Monte Carlo integration leads to identical results with the Gauss-Hermite method, with the latter being more than 80% faster to implement. Other approximations can be alternatively used, such as the adaptive quadrature points (Rabe-Hesketh et al. 2005, Schilling and Bock 2005) or Laplace approximations (Huber et al., 2004). 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. The estimated values for each case are presented in Table 1.

Table 1: **BML estimates (log scale) for the GLLVM example**

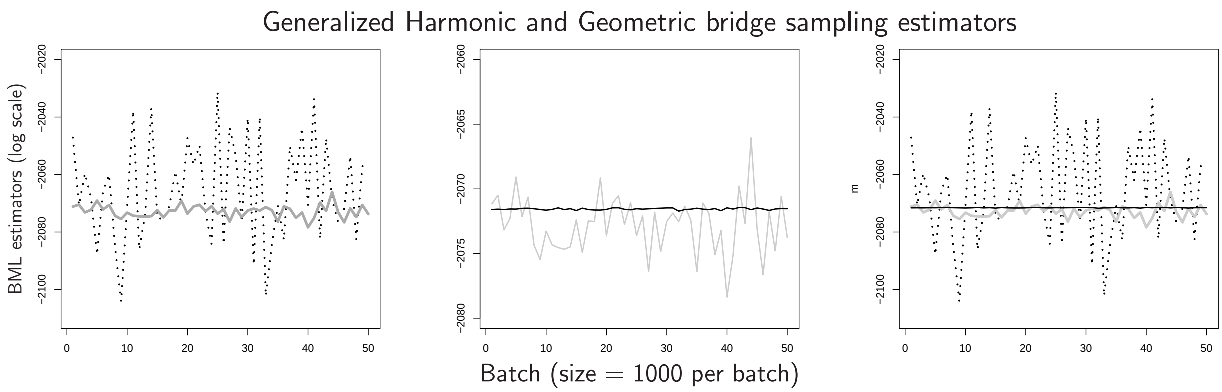
Approach	Estimator	Estimation	Batch mean	$\widehat{MCE}$
Joint	RM	-2062.3	-2053.9	3.46
	BH	-2068.8	-2065.5	17.92
	BG	-2073.3	-2072.8	2.21
Marginal	RM	-2071.3	-2071.2	0.28
	BH	-2069.6	-2069.3	2.11
	BG	-2071.6	-2071.6	0.07

The estimated BML of a GLLVM model with $p = 6$ items, $N = 600$ cases and $k = 2$ factors. Each estimation was computed over a sample of 50,000 simulated points while the batch mean and the associated error were computed over 50 batches of 1,000 points each. RM: Reciprocal mean estimator, BH: Bridge harmonic estimator and BG: Bridge geometric estimator.

2.3 Estimations and key observations

The first observation derived from the current example refers to the variability differences between the estimators and between their joint and marginal counterparts. For illustration purposes we focus on the two bridge sampling estimators. The joint bridge harmonic (BH_J) and bridge geometric (BG_J) estimators are depicted in Figure 1(a) over the 50 batches. The variability differences between them is striking, implying that the geometric estimator is a variance reduction technique as opposed to the harmonic. The next step in our investigation was to compare the less variant estimator with its marginal counterpart. Figure 1(b) illustrates that further variance reduction can be achieved by implementing the marginal rather than the joint geometric estimator. It becomes apparent that even the efficient bridge geometric estimator was considerably improved by employing the marginal approach. This fact is typical in high dimensional models and often expected intuitively.

The second observation was less imaginable and it refers to the estimated values per se. In particular, Figure 1(c) illustrates that the BH_J, BG_J and BG_M estimators vary around a common estimated value for the BML and the divergencies present in Table 1 are within the margins of their corresponding errors. However this is not true in the case of the reciprocal estimator. As opposed to the bridge estimators, Figure 2(a) illustrates that substantially distant estimations were derived by the joint (RM_J) and marginal (RM_M) reciprocal estimators. The difference in the estimated values is about 10 units in log-scale, meaning that it far exceeds the corresponding MCEs and hence cannot be explained solely by variability. In addition, it is interesting to notice that the RM_J occurs to be much more divergent than the BH_J, even though the latter is associated with 5 times higher error (Table 1). The three joint estimators are depicted in Figure 2(b) and their marginal counterparts are illustrated in Figure 2(c).



(a) BH_J and BG_J (b) BG_J and BG_M (c) BH_J, BG_J and BG_M)

Figure 1: The *joint* bridge harmonic estimator BH_J (dotted line), the *joint* bridge geometric estimator BG_J (gray solid line) and the *marginal* bridge geometric estimator BG_M (black solid line), for the BML (log scale), implementing a simulated data set with $p = 6$ binary items, $N = 600$ cases and $k = 2$ factors, over 50 batches.

Several concerns arise therefore with regard to the convergence of the estimators in finite settings, listed below:

a) What is the mechanism which produces these differences?

b) Can the differences in the error be ameliorated to some extend by increasing the simulated sample size in finite settings?

c) By increasing the number of the simulated points, do the discrepancies in the estimated values reduce? Where is this type of bias coming from?

Regarding the mechanism, we state here that is related to the model assumptions. Specifically, consider the model parameters $\boldsymbol{\theta}$ fixed in the BML expressions (3) and (4). It occurs that the joint expression implements the mean of the product of the independent variables $f_{\boldsymbol{\theta}}(\mathbf{Y}_i|\boldsymbol{Z}_i)$ while the marginal expression employs the product of their means. The former is a generally applicable approach while the latter occurs explicitly under independence. We conclude that the joint approach makes subtle use of the local independence assumption. This fact has direct implications on the estimated value and the associated error which are thoroughly examined in the following section.

3 Joint and marginal Monte Carlo estimators under independence assumptions

The Monte Carlo integration techniques are reviewed here in a general framework, since the subsequent theoretical findings extend beyond models with latent variables. In particular, we consider any multi-dimensional integral of the form

$$I = \int \phi(\mathbf{Y})h(\mathbf{Y})\,d\mathbf{Y}, \text{ where } \mathbf{Y} = (Y_1, Y_2, ..., Y_N). \tag{5}$$

The MC approximation of the integral (5) corresponds to the expected value of $\phi(\boldsymbol{Y})$ over $h(\boldsymbol{Y})$. Specifically, if $\boldsymbol{y}^{(r)} = \big(y_1^{(r)}, y_2^{(r)}, ..., y_N^{(r)}\big)$ and $\mathbf{y}^R = \big\{\boldsymbol{y}^{(r)}\big\}_{r=1}^R$ is a random sample of points generated from the distribution $h(\boldsymbol{y})$, then the estimator $\widehat{I} = \overline{\phi} = \frac{1}{R}\sum_{r=1}^R \phi\big(\boldsymbol{y}^{(r)}\big)$ will approach (5) for sufficiently large sample size R. The degree of accuracy associated with the Monte Carlo estimator is directly related to the size of the simulated sample R. The standard deviation of $\overline{\phi}$ is the MCE of the estimator. The MCE is therefore defined as the standard deviation of the estimator across simulations of the same number of replications R and is given by:

$$MCE = \sqrt{\mathrm{Var}(\,\overline{\phi}\,)} = \frac{\sigma_\phi}{\sqrt{R}}, \tag{6}$$

while an obvious estimator of MCE is given by $\widehat{MCE} = S_\phi/\sqrt{R}$, provided that an estimator S_ϕ^2 of the integrand's variance σ_ϕ^2 is available. From (6), it occurs that the MCE directly depends on σ_ϕ and R.

Here we focus on the estimation of the expected value of $\phi(\boldsymbol{Y}) = \prod_{i=1}^{N} \phi_i(Y_i)$ given by

$$I = E[\phi(\mathbf{Y})] = E\left[\prod_{i=1}^{N} \phi_i(Y_i)\right] = \int \prod_{i=1}^{N} \phi_i(Y_i) h(Y_1, Y_2, ..., Y_N)\, d(Y_1, Y_2, ..., Y_N)\ . \tag{7}$$

Under the assumption of independence for Y_i, we can rewrite (7) as

$$I = \prod_{i=1}^{N} E\left[\phi_i(Y_i)\right] = \prod_{i=1}^{N} \int \phi_i(Y_i) h_i(Y_i) dY_i\ . \tag{8}$$

The expressions (7) and (8) can be used to construct two unbiased Monte Carlo estimators of I, described in Definitions 3.1 and 3.2 that follow.

Definition 3.1 Joint estimator of I. *For any random sample* $\left\{y_1^{(r)}, y_2^{(r)}, ..., y_N^{(r)}\right\}_{r=1}^{R}$ *from* h*, the joint estimator of* I *is defined as*

$$\widehat{I}_J = \overline{\phi} = \frac{1}{R}\sum_{r=1}^{R} \phi\left(y_1^{(r)}, y_2^{(r)}, ..., y_N^{(r)}\right) = \frac{1}{R}\sum_{r=1}^{R}\left[\prod_{i=1}^{N} \phi_i\left(y_i^{(r)}\right)\right]\ . \tag{9}$$

Definition 3.2 Marginal estimator of I. *For any random sample* $\left\{y_1^{(r)}, y_2^{(r)}, ..., y_N^{(r)}\right\}_{r=1}^{R}$ *from* h*, the marginal estimator of* I *is defined as*

$$\widehat{I}_M = \prod_{i=1}^{N}\left[\frac{1}{R}\sum_{r=1}^{R} \phi_i\left(y_i^{(r)}\right)\right] = \prod_{i=1}^{N} \overline{\phi}_i. \tag{10}$$

In the remaining of the paper we examine the divergencies between the two estimators in finite settings, as a result of disregarding the assumption of independence.

3.1 Monte Carlo errors

The exact MCEs for the joint and marginal estimators are expressed in terms of their variances. In particular, the variance of the joint estimator (9) is directly linked to the variance of the product of N independent variables since

$$Var(\widehat{I}_J) = Var\left[\frac{1}{R}\sum_{r=1}^{R}\left\{\prod_{i=1}^{N} \phi_i\big(y_i^{(r)}\big)\right\}\right] = \frac{Var\left[\prod_{i=1}^{N} \phi_i(Y_i)\right]}{R}. \tag{11}$$

On the other hand, the variance of the marginal estimator (10) is given by the variance of the product of N univariate MC estimators, that is

$$Var(\widehat{I}_M) = Var\left[\prod_{i=1}^{N} \overline{\phi}_i\right]. \tag{12}$$

The difference between (11) and (12) becomes apparent if the early findings of Goodman (1962) are reviewed within the framework of Monte Carlo integration. Goodman (1962, eq. 1 and 2) provides the variance σ^2 of the product of N independent variables Y_i, $(i = 1, ..., N)$ with probability or density functions $h_i(Y_i)$. For our purposes, we expand it to the case of functions $\phi_i(Y_i)$ of the original independent random variables, leading to

$$Var\left(\prod_{i=1}^{N}\phi_i(Y_i)\right) = \sum_{i=1}^{N} V_i \prod_{i'\neq i}^{N} E_{i'}^2 + \sum_{i_1<i_2}^{N} V_{i_1}V_{i_2} \prod_{i'\neq i_1,i_2}^{N} E_{i'}^2 + ... + V_1V_2\cdots V_N, \quad (13)$$

where $E_{i'} = E[\phi_{i'}(Y_{i'})]$ and $V_i = Var[\phi_i(Y_i)]$, $(i, i' \in \{1, ..., N\})$, with all moments being calculated over the corresponding densities $h_i(Y_i)$.

Equation (13) can be written as

$$Var\left(\prod_{i=1}^{N}\phi_i(Y_i)\right) = \sum_{k=1}^{N}\sum_{\mathcal{C}\in\binom{\mathcal{N}}{k}}\left[\prod_{i\in\mathcal{C}}V_i\prod_{j\in\mathcal{N}\setminus\mathcal{C}}E_j^2\right], \quad (14)$$

where $\binom{\mathcal{N}}{k}$ is the set of all possible combinations of k elements of $\mathcal{N} = \{1, 2, \ldots, N\}$ and any product over the empty set is specified to be equal to one.

The variances of the two Monte Carlo estimators in (11) and (12) may now be expressed in terms of (13). Specifically, the variance of the joint estimator is directly obtained by dividing the integrand's variance in (13) with the simulated sample size R. For the marginal estimator, the variance (12) can be obtained by substituting V_i by V_i/R in (14). The variance components that correspond to the MCEs in each case are presented in the following lemma.

Lemma 3.1 *The variances of the joint (9) and marginal estimators (10) are given by*

$$Var(\widehat{I}_J) = \frac{1}{R}\sum_{i\in\mathcal{N}}V_i\prod_{j\in\mathcal{N}\setminus\{i\}}E_j^2 + \sum_{k=2}^{N}\left[\frac{1}{R}\sum_{\mathcal{C}\in\binom{\mathcal{N}}{k}}\prod_{i\in\mathcal{C}}V_i\prod_{j\in\mathcal{N}\setminus\mathcal{C}}E_j^2\right],$$

and

$$Var(\widehat{I}_M) = \frac{1}{R}\sum_{i\in\mathcal{N}}V_i\prod_{j\in\mathcal{N}\setminus\{i\}}^{N}E_j^2 + \sum_{k=2}^{N}\left[\frac{1}{R^k}\sum_{\mathcal{C}\in\binom{\mathcal{N}}{k}}\prod_{i\in\mathcal{C}}V_i\prod_{j\in\mathcal{N}\setminus\mathcal{C}}E_j^2\right],$$

In each case, the associated MCE equals the square root of the corresponding variance in Lemma 3.1. The variances (and therefore the MCEs) are asymptotically equivalent, since both converge to zero with rate of order $\mathcal{O}(R^{-1})$. However, with the exception of the first term in $Var(\widehat{I}_M)$, the rest of the components in the summation converge faster to zero with rates $\mathcal{O}(R^{-k})$ for any $k \geq 2$. Hence, in finite settings the joint estimator will always have larger error. The factors that influence the magnitude of this difference are discussed in the next section.

3.2 Determinants of Monte Carlo error difference

In this section, we study the difference in the errors associated with the joint and marginal estimators. We illustrate how it depends on the dimensionality of the problem at hand (N), the variation of the variables involved and the simulated sample's size (R).

To begin with, if both estimators $\widehat{I}_J$ and $\widehat{I}_M$ are applied with the same finite R, then according to Lemma 3.1, the difference in their variances is given by

$$Var(\widehat{I}_J) - Var(\widehat{I}_M) \quad = \quad \frac{1}{R}\sum_{k=2}^{N}\left[\left(1-\frac{1}{R^{k-1}}\right)\sum_{\mathcal{C}\in\binom{\mathcal{N}}{k}}\prod_{i\in\mathcal{C}}V_i\prod_{j\in\mathcal{N}\setminus\mathcal{C}}E_j^2\right],$$

As the number of the variables increases, more positive terms are added to (15) and this explains the indirect effect of the dimensionality. The effect of the moments E_i and $V_i,\ i = 1\ldots N$, can be expressed in terms of the corresponding coefficients of variation (CV_i^2), according to the following lemma.

Lemma 3.2 *Without loss of generality, let $\{Y_i, i\in\mathcal{N}_0\}$ be the sub-set of $\{Y_1, Y_2, \ldots, Y_N\}$ random variable with zero expectations. The variances of the joint (9) and marginal (10) estimators are given by:*

$$Var(\widehat{I}_J) = \frac{1}{R}\times\prod_{i\in\mathcal{N}_0}V_i \ \times\prod_{i\in\overline{\mathcal{N}}_0}E_i^2\times\left(\prod_{i\in\overline{\mathcal{N}}_0}(CV_i^2+1) - I(\mathcal{N}_0=\emptyset)\right)$$

and

$$Var(\widehat{I}_M) = \frac{1}{R^{\mathcal{N}_0}}\times\prod_{i\in\mathcal{N}_0}V_i \ \times\prod_{i\in\overline{\mathcal{N}}_0}E_i^2\times\left(\prod_{i\in\overline{\mathcal{N}}_0}\left(\frac{CV_i^2}{R}+1\right) - I(\mathcal{N}_0=\emptyset)\right)$$

where $\mathcal{N}_0 \subseteq \mathcal{N} = \{0,1,...,N\}$, $\overline{\mathcal{N}}_0 = \mathcal{N}\setminus\mathcal{N}_0$ is the index of variables Y_i with non-zero expectations, $\prod_{i\in\emptyset}Q_i = 1$ for any Q_i and $I(\mathcal{N}_0=\emptyset)$ is equal to one if $E_i\neq 0$ for all $i\in\mathcal{N}$ and zero otherwise.

▷ The proof of Lemma 3.2 is given at the Appendix. □

Based on Lemma 3.2, the difference in the variances of the estimators becomes larger as the variability of the Y_is increases. The maximum difference occurs when all variables involved have zero means, in which case $Var(\widehat{I}_J) = R^N\,Var(\widehat{I}_M)$. On the contrary, when all means are non zero, the difference mainly depends on the coefficients of variation. Based on Lemma 3.2, we may also consider the case where the two estimators have the same variance, that is $Var(\widehat{I}_J) = Var(\widehat{I}_M)$, which can be achieved under different number of replications, R_J and

R_M. The number of replications that the joint estimator requires, in order to archive the same error with the marginal estimator, is defined at the following corollary.

Corollary 3.1 *The joint (9) and marginal (10) estimators achieve the same accuracy when*

$$R_J = R_M^{N_0} \times \omega(N, N_0, \mathcal{CV})$$

with

$$\omega(N, N_0, \mathcal{CV}) = \begin{cases} R_M^{N-N_0} & \text{if } \mathcal{N}_0 = \mathcal{N} \\ \frac{\prod_{i=1}^{N}(CV_i^2+1)-1}{\prod_{i=1}^{N}(CV_i^2/R_M+1)-1} & \text{if } \mathcal{N}_0 = \emptyset \\ \prod_{i \in \overline{\mathcal{N}}_0} \frac{CV_i^2+1}{CV_i^2/R_M+1} & \text{otherwise} \end{cases}$$

where $N_0 = |\mathcal{N}_0|$ denotes the number of the zero mean variables, $\mathcal{CV} = \{CV_i : i \in \overline{\mathcal{N}}_0\}$ and R_J, R_M are the number of iterations for the joint and marginal estimators, respectively.

Corollary 3.1 states that the joint estimator achieves the same MCE when its number of iterations R_J is equal to the number of iterations of the marginal estimator R_M raised to the number of variables with zero expectations and multiplied by a factor $\omega(N, N_0, \mathcal{CV}) > 1$ for $R_M > 1$. Hence, in order to achieve the same precision for the two estimations, the joint estimator will always require more iterations R_J than the marginal one R_M. The multiplicative factor ω heavily depends on the number of variable with zero expectations and on the variability of the Y_is (through CVs) for the non-zero variables. In the special case where all expectations E_i are zero, the required number of iterations is R_J=R_M^N. Lemma 3.2 and Corollary 3.1 indicate that the error of the joint estimator may not be always manageable. That is, if the number of variables is large or if their variability is high, then the joint estimator requires simulated samples that can be unreasonably large.

For illustration purposes, we implement a toy example of N independent and identically distributed (i.i.d) Beta random variables $Y_i \sim Beta(\lambda_1, \lambda_2)$ $(i = 1, ..., N)$. The mean of their product is given by:

$$E\left(\prod_{i=1}^{N} Y_i\right) = \left(\frac{\lambda_1}{\lambda_1 + \lambda_2}\right)^N .$$

Fifty samples with size ranging from 5 to 250 thousands simulated points, were generated from $N = 10$ $Beta(1, 2)$ distributions. The two estimators were computed and depicted in Figure 3(a). The same procedure was repeated for $N = 50$ and $N = 150$ and is graphically represented in Figures 3(b) and 3(c).

In the low dimensional case ($N = 10$), the error of the joint estimator ($\widehat{I}_J$: light grey line) is rather comparable with the error of marginal ($\widehat{I}_M$: dark grey line). When R reaches 250 thousands, both estimators reach the true mean (I_T: dashed line). However, if the number of variables is increased to $N = 50$ and $N = 150$, the variability differences between the two approaches remain large even for $R = 250,000$; see Table 2.

The exercise was also replicated for $N = 10, 50$ and 150 i.i.d. $Beta(0.1, 0.2)$ variables. The true mean is the same with the previous setting (equal to 1/3), but the coefficient of variation (CV) is now approximately 77% higher. For the same R and N, the difference in the errors of the two estimators is even larger (Figures 3(d) to 3(e)), indicating the role of the variability of the variables involved. The estimated values and the corresponding errors are summarized in Table 2. Although, this example is simple assuming i.i.d random variables, the same picture can be reproduced for non identically distributed random variables.

Table 2: **Estimated mean of the product of i.i.d Beta variables (log scale)**

Distribution	N	I_T	$\widehat{I}_M$	$M\widehat{C}E_M$	$\widehat{I}_J$	$M\widehat{C}E_J$
	10	-10.99	-10.98	0.02	-10.97	0.07
$Beta(1, 2)$	50	-54.93	-54.93	0.06	-52.01	2.03
	150	-164.79	-164.79	0.09	-176.94	3.37
	10	-10.99	-10.98	0.04	-11.05	1.07
$Beta(0.1, 0.2)$	50	-54.93	-54.90	0.10	-113.81	13.77
	150	-164.79	-164.80	0.17	-595.13	28.50

N: Number of i.i.d variables; I_T: true mean; $\widehat{I}_{(J\ or\ M)}$: the estimated value via the joint or the marginal approach respectively, over R =250,000 iterations; $\widehat{I}_{(M\ or\ J)}$ and $M\widehat{C}E_{(M\ or\ J)}$: batch mean error over 25 batches of 10,000 points each (obtained as the standard deviation of the log estimates).

3.3 Variance reduction under conditional independence

In this section, we demonstrate how we can extend the previous results in the case of conditional independence which is more realistic in practice and it frequently met in hierarchical models with latent variables.

Specifically, let us substitute $\boldsymbol{Y}$ by $(\boldsymbol{U}, \boldsymbol{V})$. In analogy with the previous setting, let $\boldsymbol{U}_i$ (with $i = 1, 2, \ldots, N$) be conditionally independent random variables when $\boldsymbol{V}$ are given with densities denoted by $h(\boldsymbol{u}_i|\boldsymbol{v})$. We are interested in estimating the integral

$$\mathcal{I} = \int \Big[\prod_{i=1}^{N} \varphi_i(\boldsymbol{u}_i, \boldsymbol{v})\Big] h(\boldsymbol{u}, \boldsymbol{v})\, d(\boldsymbol{u}, \boldsymbol{v}), \tag{15}$$

that now corresponds to the expected value of $\varphi(\boldsymbol{u}, \boldsymbol{v}) = \prod_{i=1}^{N} \varphi_i(\boldsymbol{u}_i, \boldsymbol{v})$ over $h(\boldsymbol{u}, \boldsymbol{v})$. This can be directly estimated by the joint estimator

$$\widehat{\mathcal{I}}_J = \frac{1}{R}\sum_{r=1}^{R}\left[\prod_{i=1}^{N}\varphi_i\big(\boldsymbol{u}_i^{(r)}, \boldsymbol{v}^{(r)}\big)\right] \tag{16}$$

assuming that we can generate a random sample $\big\{\boldsymbol{u}^{(r)}, \boldsymbol{v}^{(r)}\big\}_{r=1}^{R}$ from $h(\boldsymbol{u}, \boldsymbol{v})$.

If we use the conditional independence assumption, (15) can be written as

$$\mathcal{I} = \int \left\{\prod_{i=1}^{N}\Big[\int \varphi_i(\boldsymbol{u}_i, \boldsymbol{v})h(\boldsymbol{u}_i|\boldsymbol{v})\,d\boldsymbol{u}_j\Big]\right\} h(\boldsymbol{v})\,d\boldsymbol{v} = \int \prod_{i=1}^{N} E\big(\varphi_i\big|\boldsymbol{v}\big)h(\boldsymbol{v})\,d\boldsymbol{v}, \tag{17}$$

where $E\big(\varphi_i\big|\boldsymbol{v}\big)$ is the conditional expectation of $\varphi_i(\boldsymbol{u}_i, \boldsymbol{v})$ with respect to $h(\boldsymbol{u}_i|\boldsymbol{v})$. From (17) we can directly obtain the corresponding marginal estimator by

$$\widehat{\mathcal{I}}_M = \frac{1}{R_1}\sum_{r_1=1}^{R_1}\Big[\prod_{i=1}^{N}\overline{\varphi}_i^{(r_1)}\Big] \text{ with } \overline{\varphi}_i^{(r_1)} = \frac{1}{R_2}\sum_{r_2=1}^{R_2}\varphi_i\big(u_i^{(r_2)}, \boldsymbol{v}^{(r_1)}\big), \tag{18}$$

calculated by a nested Monte Carlo experiment; where $\big\{\boldsymbol{v}^{(r_1)}\big\}_{r_1=1}^{R_1}$ is a sample from $h(\boldsymbol{v})$ and $\big\{u_i^{(r_2)}\big\}_{r_2=1}^{R_2}$ is a sample obtained by the conditional distribution $h\big(u_i|\boldsymbol{v} = \boldsymbol{v}^{(r_1)}\big)$.

Lemma 3.3 *The variances of the joint (16) and marginal estimators (18) under the assumption of conditional independence are given by*

$$Var(\widehat{I}_J) = \frac{1}{R}Var_{\boldsymbol{v}}\Big[\prod_{i=1}^{N}E\big(\varphi_i\big|\boldsymbol{v}\big)\Big] + \frac{1}{R}\sum_{k=1}^{N}\sum_{\mathcal{C}\in\binom{\mathcal{N}}{k}} E_{\boldsymbol{v}}\Big[\prod_{i\in\mathcal{C}}V\big(\varphi_i\big|\boldsymbol{v}\big)\prod_{j\in\mathcal{N}\setminus\mathcal{C}}E\big(\varphi_j\big|\boldsymbol{v}\big)^2\Big]$$

and

$$Var(\widehat{I}_M) = \frac{1}{R_1}Var_{\boldsymbol{v}}\Big[\prod_{i=1}^{N}E\big(\varphi_i\big|\boldsymbol{v}\big)\Big] + \frac{1}{R_1}\sum_{k=1}^{N}\frac{1}{R_2^k}\sum_{\mathcal{C}\in\binom{\mathcal{N}}{k}} E_{\boldsymbol{v}}\Big[\prod_{i\in\mathcal{C}}V\big(\varphi_i\big|\boldsymbol{v}\big)\prod_{j\in\mathcal{N}\setminus\mathcal{C}}E\big(\varphi_j\big|\boldsymbol{v}\big)^2\Big]$$

where $E_{\boldsymbol{v}}\big[g(\boldsymbol{v})\big]$ and $Var_{\boldsymbol{v}}\big[g(\boldsymbol{v})\big]$ denote the expectation and the variance of $g(\boldsymbol{v})$ with respect to $h(\boldsymbol{v})$ and $V\big(\varphi_i\big|\boldsymbol{v}\big)$ is, in analogy to $E\big(\varphi_i\big|\boldsymbol{v}\big)$, the conditional variance of $\varphi_i(\boldsymbol{u}_i, \boldsymbol{v})$ with respect to $h(\boldsymbol{u}_i|\boldsymbol{v})$.

$\triangleright$ The proof of Lemma 3.3 is given at the Appendix. $\square$

Lemma 3.3 is an extension of Lemma 3.1 for the case of conditional independence. For this reason, similar statements about the behaviour and the error of the joint and the marginal estimators also hold for the case of conditional independence. The main difference is the first term of variances of the estimators which is common and it is due to the additional variability of $\boldsymbol{v}$ which is of order $\mathcal{O}(R^{-1})$. Moreover, for $R_1 = R$ and any $R_2 > 1$ the marginal estimator is better since $Var(\widehat{I}_M) < Var(\widehat{I}_J)$. It would be interesting to examine the case of using the exactly the same computation effort in terms of Monte Carlo iterations. Nevertheless, setting $R = R_1 R_2$, then no clear conclusion can be drawn since the first common term will be of different order. For example, if we consider $R_1 = R_2 = r$ and $R = r^2$ then the two variances are given by

$$Var(\widehat{I}_J) \;=\; \frac{1}{r^2} Var_{\boldsymbol{v}}\Big[\prod_{i=1}^{N} E\big(\varphi_i|\boldsymbol{v}\big)\Big] + \frac{1}{r^2}\sum_{i=1}^{N} E_{\boldsymbol{v}}\Big[V\big(\varphi_i|\boldsymbol{v}\big)\prod_{j\in\mathcal{N}\setminus\{i\}} E\big(\varphi_j|\boldsymbol{v}\big)^2\Big] + \mathcal{O}(r^{-2})$$

and

$$Var(\widehat{I}_M) \;=\; \frac{1}{r} Var_{\boldsymbol{v}}\Big[\prod_{i=1}^{N} E\big(\varphi_i|\boldsymbol{v}\big)\Big] + \frac{1}{r^2}\sum_{i=1}^{N} E_{\boldsymbol{v}}\Big[V\big(\varphi_i|\boldsymbol{v}\big)\prod_{j\in\mathcal{N}\setminus\{i\}} E\big(\varphi_j|\boldsymbol{v}\big)^2\Big] + \mathcal{O}(r^{-3})$$

Finally, in the case that instead of nested Monte Carlo, we use a numerical method which approximates very well the expectations $E(\varphi_i|\boldsymbol{v})$ then the second term of the the variance of the corresponding marginal estimator will be zero making the method considerably more accurate and faster to converge than the joint estimator.

Due to the fact that Lemma 3.3 also incorporates similar expressions as in Lemma 3.2, the remarks made on the error differences with regard to the sample size, the number of variables and their variability apply also in the case of conditional independence assumption. We may now explain the different behaviour of the three BML estimators at the GLLVM example (Section 2), where $\boldsymbol{u}_i = \boldsymbol{Z}_i$ are the latent variables and $\boldsymbol{v} = (\boldsymbol{\alpha}, \boldsymbol{\beta})$ are the model parameters. The error differences observed in Figure 1(a) between the BH_J and BG_J estimators (for the same N and R) can be now attributed to the different coefficients of variation of the averaged quantities involved. For both estimators, the expectation in the numerator is taken over $g(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{Z}) = g(\boldsymbol{\alpha})g(\boldsymbol{\beta})\prod_{i=1}^{N}(\boldsymbol{Z}_i)$. However, the N averaged variables differ according to (24) and (25). Specifically for $i = 1, \ldots, N$ the averaged variables were:

(a) $\varphi_i(\cdot) = \Big[g(\boldsymbol{\alpha})^{1/N} g(\boldsymbol{\beta})^{1/N} g(\mathbf{Z}_i)\Big]^{-1}$, in the case of BH$_J$ and

(b) $\varphi_i'(\cdot) = \left\{\frac{f(Y_i|\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{Z}_i)\pi(\boldsymbol{Z}_i)}{g(\boldsymbol{Z}_i)}\left[\frac{\pi(\boldsymbol{\alpha})\pi(\boldsymbol{\beta})}{g(\boldsymbol{\alpha})g(\boldsymbol{\beta})}\right]^{1/N}\right\}^{1/2}$, in the case of BG$_J$.

Moreover, none of the conditional expectations will be equal to zero since ϕ_i and ϕ_i' are both positive. Therefore, following Lemma 3.2 we may rewrite the variances of the estimators as functions of the corresponding coefficients of variation

$$Var(\widehat{I}_J) = \frac{1}{R} Var_{\boldsymbol{v}}\Big[\prod_{i=1}^{N} E\big(\varphi_i|\boldsymbol{v}\big)\Big] + \frac{1}{R} E_{\boldsymbol{v}}\left[\prod_{i=1}^{N} E\big(\varphi_i|\boldsymbol{v}\big)^2 \left\{\prod_{i=1}^{N}\Big[CV\big(\varphi_i|\boldsymbol{v}\big)^2 + 1\Big] - 1\right\}\right]$$

and

$$Var(\widehat{I}_M) = \frac{1}{R_1} Var_{\boldsymbol{v}}\Big[\prod_{i=1}^{N} E\big(\varphi_i|\boldsymbol{v}\big)\Big] + \frac{1}{R_1} E_{\boldsymbol{v}}\left[\prod_{i=1}^{N} E\big(\varphi_i|\boldsymbol{v}\big)^2 \left\{ \prod_{i=1}^{N} \Big[\tfrac{CV(\varphi_i|\boldsymbol{v})^2}{R_2} + 1\Big] - 1 \right\}\right]$$

From the above equations, it is obvious that the variances of the estimators will explode for large N in the (a) case since we expect values of $\varphi_i > 1$ demanding a large number of iterations to reach a required precision level. The effect will be more evident in the joint estimator, since the marginal estimator some of these effects will be eliminated for large R_2 (or using well behaved numerical methods). For case (b), the situation seems much better, since (assuming that g is a good proxy for the posterior) the expectation in the first term (which is common in both approaches) will estimate the normalizing constant of $f(\boldsymbol{\alpha}, \boldsymbol{\beta}|\boldsymbol{y})$ for given values of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$. These values are usually small and therefore will not too greatly influenced by N. Therefore this term will be eliminated for reasonably small R and R_1. If this is the case, the second term will behave as in described in previous sections and therefore any action of marginalizing will greatly improve the Monte Carlo errors.

To verify this, we used the last 5000 iterations to calculate the corresponding CVs. For the bridge harmonic estimator, the CVs of the N quantities in (a) varied in log scale from 0.20 to 0.52 (median CV=0.27). In the case of the bridge geometric estimator, the CVs of the corresponding variables in (b) were substantially lower, varying from 0.01 to 0.10 (median CV=0.02). Similar results occurred for the denominators of the two bridge sampling estimators (harmonic: CV from 0.2 to 0.9 /geometric: CV less than 0.006).

The conditional independence setting considered here, applies to a plethora of high dimensional models involving latent vectors and it provides formally the rationale behind choosing to marginalize out the latent variables. In such settings, the rate of convergence is extremely slow and millions of iterations may be required to achieve a desirable level of precision for the joint estimator. However, convergence is not only a matter of the associated MCE, as will be explained in the next section.

3.4 The role of the sample covariation

Up to this point, we have studied the variability differences between the two approaches under consideration. In this section, we focus on the estimations themselves and show how they are influenced by sample covariation.

In the bivariate case, the difference between the mean of the product of two variables and the product of their means is by definition their covariance. Let us refer to a multivariate analogue of covariance with the general term *total covariation* defined as:

$$TCI(\boldsymbol{Y}) = E\Big(\prod_{i=1}^{N} Y_i\Big) - \prod_{i=1}^{N} E(Y_i), \tag{19}$$

which is actually the difference between the expectations under the joint and marginal approaches in their simplest forms. That is, it coincides with the difference between the expressions in (7) and (8) if in (19) we use the random variables $\phi_i(Y_i),\ i = 1, ..., N$ (for simplicity in the notation hereafter we proceed with the original variables without loss of generality). Therefore, by definition, the difference between the estimated means provided by $\widehat{I}_J$ and $\widehat{I}_M$ reflects the total sample covariation between the N variables. In fact, the total sample covariation is accountable for and completely explains the estimation differences which occur in the simulated example of Section 2.2 and cannot be attributed to the associated Monte Carlo errors of the two estimators.

Therefore, while independence demands zero total sample covariation, that is not the case in finite samples. When N random variables are simulated independently, even the smallest dependencies between the variables will result in non zero total sample covariation. The covariance induced by the simulation procedure cannot be ignored even for samples of several hundreds of thousands points (see Tables 1 and 2). We proceed with exploring the factors that affect the total sample covariation.

The identity (19) is not useful into gaining insight on the factors that affect the total sample covariation. Here, we provide an alternative expression which assesses the total covariation among N random variables, in terms of their expected means $E(Y_i)$, $i = 1, ..., N$ and covariances of the form:

$$Cov_{(k)}(\boldsymbol{Y}) = Cov\Big(\prod_{i=1}^{k-1} Y_i, Y_k\Big). \tag{20}$$

Lemma 3.4 *The total covariation among N variables, is given by:*

$$TCI(\boldsymbol{Y}) = Cov_{(N)}(\boldsymbol{Y}) + \sum_{k=1}^{N-2}\left[\left(\prod_{i=N-k+1}^{N} E(Y_i)\right) Cov_{(N-k)}(\boldsymbol{Y})\right], \tag{21}$$

where $N \geq 3$ and $E(Y_{N+1}) = 1$.

$\triangleright$ The proof of Lemma 3.4 is given at the Appendix. □

The total covariation among N random variables is therefore assessed through a weighted sum of N-1 covariance terms. The means of the variables serve as weights that adjust the contribution to the total covariation for each additional variable.

The total covariance index (TCI) defined in (19) will be zero if $\boldsymbol{Y}$ are independent. Nevertheless, the reverse statement does not hold, that is a zero value of TCI does not ensure independence. In the case where N random variables are generated independently, the sample covariances participating in (21) will be non-zero resulting in a sample total covariation far away from zero. Therefore, it can be viewed as an index of the sample's divergence from independence. By definition, the total sample covariation is accountable for and completely explains the estimation differences that were illustrated in the our examples.

Equation (21) implies that any divergence from the independence assumption is also affected by the number of variables N, their expectations, their covariation and the simulated sample size R, as already illustrated graphically in Figures 3(a) to 3(f). In the case of independent variables, the sample covariation converges to zero as R goes to infinity. The Cauchy-Schwartz inequality provides an upper bound, according to the following corollary.

Corollary 3.2 *An upper bound for the absolute value of $TCI(\boldsymbol{Y})$ is given by:*

$$|TCI(\boldsymbol{Y})| \ \le \ \sum_{k=0}^{N-2} \left[\left(\prod_{i=N+1-k}^{N+1} | \, E(Y_i)| \right) \sqrt{Var\left(\prod_{j=1}^{N-k-1} Y_i \right) Var(Y_{N-k})} \ \right] .$$

$\triangleright$ Corollary 3.2 immediately follows from Lemma 3.4 by further implementing the Cauchy-Schwartz inequality. $\square$

Corollary 3.2 provides only an upper end to the total covariation. Therefore, we cannot draw any firm conclusion regarding its magnitude especially in combination with several quantities of interest (e.g. means, variances or the number N of the variables under consideration).

However, in a vise versa point of view, it suggests that:

- The lower the expected means of the variables (in absolute value) are, the lower the index is expected to be (due to the lower bound).
- The lower the variances of the variables are, the lower the index is expected to be (due to the lower bound).
- Fewer variables (lower N) correspond to lower number of positive terms added to the right part of the inequality and therefore to lower total covariation.

The total sample covariation affects also the estimated variance of the joint estimator. Let us denote with R_0, the number of iterations required to overcome the sample covariation effect. For simulated samples less than R_0, the variance of the joint estimator is underestimated by a factor of $TCI(\boldsymbol{Y})^2$, according to the following lemma.

Lemma 3.5 *The variance of the product of N variables, equals their variance under assumed independence minus the square of their total covariation,*

$$Var\left(\prod_{i=1}^{N} Y_i\right) = Var\left(\prod_{i=1}^{N} Y_i \middle| Independence\right) - TCI(\boldsymbol{Y})^2 \ , \tag{22}$$

where $Var\left(\prod_{i=1}^{N} Y_i \middle| Independence\right)$ is the variance of the product under the assumption of independence.

$\triangleright$ The proof of Lemma 3.5 is given at the Appendix. $\square$

According to Lemma 3.5, in the presence of sample total covariation, the joint approach leads in practice to a false sense of accuracy. Once the simulated sample is large enough (larger than R_0), the covariation effect vanishes ($TCI(\boldsymbol{Y})^2 \simeq 0$), yet the variance of the joint estimator is always larger than the one associated with the marginal estimator, according to (15).

Based on the sample total covariation of $\boldsymbol{\Phi} = \big(\phi_1(Y_1), \ldots, \phi_N(Y_N)\big)$, it is now possible to explain why at the GLLVM example (Section 2) MCMC estimators lead to biased estimations, even though they were associated with low MCE. In particular, the sample covariation does not seem to affect the bridge harmonic (BH_J) estimator while it is clearly present in the case of the reciprocal (RM_J) estimator (see Table 1). To explain this phenomenon, we need first to underline that the bridge harmonic estimator is a ratio. Based on the last 5,000 draws, the sample total covariation between the averaged variables at the numerator of BH_J was -723.8 and -730.5 at the denominator. These values are substantially larger than the sample covariation among the averaged variables in the case of the reciprocal estimator (equal to -23.0). However, since BH_J is a ratio the sample covariations estimated at the numerator and the denominator cancel out, which is not the case for the reciprocal estimator. Similarly, the sample covariation effect also cancels out in the case of the bridge geometric estimator. On the other hand, the sample covariation leads to underestimation of the marginal likelihood which, in turn, leads to underestimation of the MCE (Lemma 3.5). Therefore, estimators that are affected by sample covariation appear to have smaller MCE than they actually do.

4 Discussion

In the presence of independence assumptions, the mean of the product of N variables can be either estimated by implementing the joint or the marginal approaches. Early in the 90's, Kim et al. (1994) had identified the significance and the distinction of the two approaches comparing them empirically (for applications, see also references within). As pointed out by referees, modern-day practitioners less frequently select joint estimators for the evaluation of the marginal likelihood, especially in settings with high-dimensional latent vectors. Nevertheless, the easy-to-use specification of the MCMC sampler and their direct implementation through ready-to-use MCMC software such as WinBUGS, may lead researchers to marginal likelihood estimators based on the full augmented scheme and the corresponding joint estimators either for illustration purposes or for actual data analysis. For example, the efficiency of the marginal likelihood estimation could have been increased if the "marginal" approach have been adopted in the longitudinal examples of Chib and Jeliazkov (2001, Sec. 3.2 and 3.3) and Friel and Pettitt (2008, Sec. 4.2). Another example of the implementation of the joint estimator can be found in Congdon (2010, Sec. 2.2.3), used for illustration purposes in hierarchical models. In all the above cases, there is no indication that the estimators have not converged to their target value but, according to our ndings, similar levels of precision could have been achieved in less iterations if the marginal approach have been used instead.

On the other hand, the intuitive tendency to use the marginal estimators is mainly based on empirical findings rather than theoretical justifications, as the ones provided here. The absence of explicit evidence against the use of the joint marginal likelihood estimators in the relevant literature, makes its use even more attractive since it is directly available from easily applicable Gibbs sampling schemes.

The findings of this article have applicability in a wide range of modern statistical models that can be expressed with the incorporation of latent variables. A (non-exhaustive) list of such models include the Gaussian factor models, factor models with non-normal responses, latent class models, random effects models, longitudinal models, cluster analysis models, data-augmented specific models, mixture models, negative binomial response models, zero-inflated models, bivariate Poisson models and Poisson difference (or Skellam distributed) models.

APPENDIX

The identities of the MCMC estimators used in the Section 2.1 are

- *Reciprocal importance (RM) sampling* estimator (Gelfand and Dey, 1994)

$$f(\boldsymbol{Y}) = \left[\int \frac{g(\boldsymbol{\vartheta})}{f(\mathbf{Y}|\,\boldsymbol{\vartheta})\,\pi(\boldsymbol{\vartheta})}\,\pi(\boldsymbol{\vartheta}|\,\mathbf{Y})\,d\boldsymbol{\vartheta}\right]^{-1}, \tag{23}$$

- *Generalized harmonic bridge (BH) sampling* estimator (Meng and Wong, 1996)

$$f(\boldsymbol{Y}) = \frac{\displaystyle\int [\,g(\boldsymbol{\vartheta})]^{-1}\,g(\boldsymbol{\vartheta})\,d\boldsymbol{\vartheta}}{\displaystyle\int [f(\mathbf{Y}|\,\boldsymbol{\vartheta})\pi(\boldsymbol{\vartheta})]^{-1}\,\pi(\boldsymbol{\vartheta}|\,\mathbf{Y})\,d\boldsymbol{\vartheta}}, \tag{24}$$

- *Geometric bridge (BG) sampling* estimator (Meng and Wong, 1996)

$$f(\mathbf{Y}) = \frac{\displaystyle\int \left[\frac{f(\mathbf{Y}|\,\boldsymbol{\vartheta})\pi(\boldsymbol{\vartheta})}{g(\boldsymbol{\vartheta})}\right]^{1/2} g(\boldsymbol{\vartheta})\,d\boldsymbol{\vartheta}}{\displaystyle\int \left[\frac{f(\mathbf{Y}|\,\boldsymbol{\vartheta})\pi(\boldsymbol{\vartheta})}{g(\boldsymbol{\vartheta})}\right]^{-1/2} \pi(\boldsymbol{\vartheta}|\,\mathbf{Y})\,d\boldsymbol{\vartheta}}. \tag{25}$$

Proof of Lemma 3.2

According to Goodman (1962), the variance of the product of N variables is given by

$$Var\left(\prod_{i=1}^{N}\phi_i(Y_i)\right) = \prod_{i=1}^{N}\left(V_i + E_i^2\right) - \prod_{i=1}^{N}E_i^2. \tag{26}$$

Hence we can write

$$\begin{aligned}
Var\left(\prod_{i=1}^{N}\phi_i(Y_i)\right) &= \prod_{i\in\mathcal{N}_0}\left(V_i + E_i^2\right)\prod_{i\in\overline{\mathcal{N}}_0}\left(V_i + E_i^2\right) - \prod_{i\in\mathcal{N}_0}E_i^2\prod_{i\in\overline{\mathcal{N}}_0}E_i^2. \\
&= \prod_{i\in\mathcal{N}_0}V_i\prod_{i\in\overline{\mathcal{N}}_0}\left[E_i^2\left(CV_i^2+1\right)\right] - \prod_{i\in\mathcal{N}_0}E_i^2\prod_{i\in\overline{\mathcal{N}}_0}E_i^2. \\
&= \prod_{i\in\overline{\mathcal{N}}_0}E_i^2 \times \left[\prod_{i\in\mathcal{N}_0}V_i\prod_{i\in\overline{\mathcal{N}}_0}\left(CV_i^2+1\right) - \prod_{i\in\mathcal{N}_0}E_i^2\right].
\end{aligned}$$

Note that $\prod_{i\in\mathcal{N}_0} E_i^2$ will be the value of one if $\mathcal{N}_0 = \emptyset$ and zero otherwise. Therefore we can write $\prod_{i\in\mathcal{N}_0} E_i^2 = \prod_{i\in\mathcal{N}_0} E_i^2 \times \prod_{i\in\mathcal{N}_0} V_i^2$ resulting in

$$\begin{aligned} Var\left(\prod_{i=1}^{N} \phi_i(Y_i)\right) &= \prod_{i\in\mathcal{N}_0} V_i \times \prod_{i\in\overline{\mathcal{N}}_0} E_i^2 \times \left[\prod_{i\in\overline{\mathcal{N}}_0} \big(CV_i^2+1\big) - \prod_{i\in\mathcal{N}_0} E_i^2\right]. \\ &= \prod_{i\in\mathcal{N}_0} V_i \times \prod_{i\in\overline{\mathcal{N}}_0} E_i^2 \times \left[\prod_{i\in\overline{\mathcal{N}}_0} \big(CV_i^2+1\big) - I(\mathcal{N}_0=\emptyset)\right], \end{aligned}$$

which gives

$$Var\left(\prod_{i=1}^{N} \phi_i(Y_i)\right) =$$

$$= \begin{cases} \prod_{i=1}^{N} V_i & \text{if } \mathcal{N}_0 = \mathcal{N} \text{ (all expectations are zero)} \\ \prod_{i=1}^{N} E_i^2 \times \left[\prod_{i=1}^{N} \big(CV_i^2+1\big) - 1\right] & \text{if } \mathcal{N}_0 = \emptyset \text{ (all expectations are non-zero)} \\ \prod_{i\in\mathcal{N}_0} V_i \times \prod_{i\in\overline{\mathcal{N}}_0} E_i^2 \times \prod_{i\in\overline{\mathcal{N}}_0} \big(CV_i^2+1\big) & \text{otherwise} \end{cases}$$

The proof is completed by placing the general expression for the integrand's variance in (11) and (12) respectively. □

Proof of Lemma 3.3

$$\begin{aligned} Var(\widehat{I}_J) &= Var_{(\boldsymbol{u},\boldsymbol{v})}\left\{\frac{1}{R}\sum_{r=1}^{R}\left[\prod_{i=1}^{N}\varphi_i\big(\boldsymbol{u}_i^{(r)},\boldsymbol{v}^{(r)}\big)\right]\right\} \\ &= \frac{1}{R} Var_{(\boldsymbol{u},\boldsymbol{v})}\left[\prod_{i=1}^{N}\varphi_i\big(\boldsymbol{u}_i,\boldsymbol{v}\big)\right] \\ &= \frac{1}{R} Var_{\boldsymbol{v}}\left\{E_{\boldsymbol{u}|\boldsymbol{v}}\left[\prod_{i=1}^{N}\varphi_i\big(\boldsymbol{u}_i,\boldsymbol{v}\big)\,\Big|\boldsymbol{v}\right]\right\} + \frac{1}{R} E_{\boldsymbol{v}}\left\{Var_{\boldsymbol{u}|\boldsymbol{v}}\left[\prod_{i=1}^{N}\varphi_i\big(\boldsymbol{u}_i,\boldsymbol{v}\big)\,\Big|\boldsymbol{v}\right]\right\} \end{aligned} \quad (27)$$

Due to conditional independence we have that

$$E_{\boldsymbol{u}|\boldsymbol{v}}\left[\prod_{i=1}^{N}\varphi_i\big(\boldsymbol{u}_i,\boldsymbol{v}\big)\,\Big|\boldsymbol{v}\right] = \prod_{i=1}^{N}E_{\boldsymbol{u}|\boldsymbol{v}}\left[\varphi_i\big(\boldsymbol{u}_i,\boldsymbol{v}\big)\,\Big|\boldsymbol{v}\right] = \prod_{i=1}^{N}E\left(\varphi_i|\boldsymbol{v}\right). \tag{28}$$

Moreover, from (14) we have that

$$Var_{\boldsymbol{u}|\boldsymbol{v}}\left[\prod_{i=1}^{N}\varphi_i\big(\boldsymbol{u}_i,\boldsymbol{v}\big)\,\Big|\boldsymbol{v}\right] = \sum_{k=1}^{N}\sum_{\mathcal{C}\in\binom{\mathcal{N}}{k}}\Big[\prod_{i\in\mathcal{C}}V\big(\varphi_i|\boldsymbol{v}\big)\prod_{j\in\mathcal{N}\setminus\mathcal{C}}E\big(\varphi_j|\boldsymbol{v}\big)^2\Big] \tag{29}$$

By substituting (28) and (29) in (27), we obtain the variance of the joint estimator of Lemma 3.3.

Similarly, for the marginal estimator we have

$$\begin{aligned}
Var\left(\widehat{\mathcal{I}}_M\right) &= Var_{(\boldsymbol{u},\boldsymbol{v})}\left[\frac{1}{R_1}\sum_{r_1=1}^{R_1}\prod_{i=1}^{N}\overline{\varphi}_i^{(r_1)}\right] = \frac{1}{R_1}Var_{(\boldsymbol{u},\boldsymbol{v})}\left[\prod_{i=1}^{N}\overline{\varphi}_i\right] \\
&= \frac{1}{R_1}Var_{\boldsymbol{v}}\left\{E_{\boldsymbol{u}|\boldsymbol{v}}\Big[\prod_{i=1}^{N}\overline{\varphi}_i\,\Big|\boldsymbol{v}\Big]\right\} + \frac{1}{R_1}E_{\boldsymbol{v}}\left\{Var_{\boldsymbol{u}|\boldsymbol{v}}\Big[\prod_{i=1}^{N}\overline{\varphi}_i\,\Big|\boldsymbol{v}\Big]\right\}
\end{aligned} \tag{30}$$

Due to conditional independence we have that

$$E_{\boldsymbol{u}|\boldsymbol{v}}\Big[\prod_{i=1}^{N}\overline{\varphi}_i\,\Big|\boldsymbol{v}\Big] = \prod_{i=1}^{N}E_{\boldsymbol{u}|\boldsymbol{v}}\left[\overline{\varphi}_i\,\Big|\boldsymbol{v}\right] = \prod_{i=1}^{N}E\left(\varphi_i|\boldsymbol{v}\right). \tag{31}$$

Moreover, from Lemma 3.1 we have that

$$Var_{\boldsymbol{u}|\boldsymbol{v}}\Big[\prod_{i=1}^{N}\overline{\varphi}_i\,\Big|\boldsymbol{v}\Big] = \sum_{k=1}^{N}\left[\frac{1}{R_2^k}\sum_{\mathcal{C}\in\binom{\mathcal{N}}{k}}\prod_{i\in\mathcal{C}}V_i\prod_{j\in\mathcal{N}\setminus\mathcal{C}}E_j^2\right], \tag{32}$$

Substituting (31) and (32) in (30) gives the expression of the variance of the marginal estimator of Lemma 3.3.

Proof of Lemma 3.4

The proof of Lemma 3.4 can be obtained by induction. The statement of the Lemma holds for $N = 3$ with $\boldsymbol{Y}_3 = (Y_1, Y_2, Y_3)$ since

$$\begin{aligned}
Cov_{(3)}(\boldsymbol{Y}) + \sum_{k=1}^{1}\left[\left(\prod_{i=4-k}^{3} E(Y_i)\right) Cov_{(3-k)}(\boldsymbol{Y})\right] &= Cov_{(3)}(\boldsymbol{Y}) + \left(\prod_{i=3}^{3} E(Y_i)\right) Cov_{(2)}(\boldsymbol{Y}) \\
&= Cov(Y_1Y_2, Y_3) + E(Y_3)Cov(Y_1, Y_2) \\
&= E(Y_1Y_2Y_3) - E(Y_1Y_2)E(Y_3) + E(Y_3)[E(Y_1Y_2) - E(Y_1)E(Y_2)] \\
&= TCI(\boldsymbol{Y}_3)\ .
\end{aligned}$$

which is true by the definition of TCI (see equation 19) for vectors $\boldsymbol{Y}$ of length equal to three.

Let us now assume that (21) it is true for any vector $\boldsymbol{Y}_N$ of length $N > 3$. Then, for $\boldsymbol{Y}_{N+1} = (\boldsymbol{Y}_N, Y_{N+1}) = (Y_1, \ldots, Y_N, Y_{N+1})$ the equation

$$\mathrm{TCI}(\boldsymbol{Y}_{N+1}) = Cov_{(N+1)}(\boldsymbol{Y}) + \sum_{k=1}^{N-1}\left[\left(\prod_{i=N-k+2}^{N+1} E(Y_i)\right) Cov_{(N+1-k)}(\boldsymbol{Y})\right], \tag{33}$$

is also true since

$$\begin{aligned}
TCI(\boldsymbol{Y}_{N+1}) = E\left(\left[\prod_{i=1}^{N} Y_i\right] \times Y_{N+1}\right) - \left[\prod_{i=1}^{N} E(Y_i)\right] E(Y_{N+1}) \\
&= Cov_{(N+1)}(\boldsymbol{Y}) + E\left(\prod_{i=1}^{N} Y_i\right) E(Y_{N+1}) - \left[\prod_{i=1}^{N} E(Y_i)\right] E(Y_{N+1}) \\
&= Cov_{(N+1)}(\boldsymbol{Y}) + TCI(\boldsymbol{Y}_N)E(Y_{N+1}) \\
&= Cov_{(N+1)}(\boldsymbol{Y}) + \left\{Cov_{(N)}(\boldsymbol{Y}) + \sum_{k=1}^{N-2}\left[\left(\prod_{i=N-k+1}^{N} E(Y_i)\right) Cov_{(N-k)}(\boldsymbol{Y})\right]\right\} E(Y_{N+1}) \\
&\qquad \textit{(from eq. 21)} \\
&= Cov_{(N+1)}(\boldsymbol{Y}) + Cov_{(N)}(\boldsymbol{Y})E(Y_{N+1}) + \sum_{k=1}^{N-2}\left[\left(\prod_{i=N-k+1}^{N+1} E(Y_i)\right) Cov_{(N-k)}(\boldsymbol{Y})\right] \\
&= Cov_{(N+1)}(\boldsymbol{Y}) + Cov_{(N)}(\boldsymbol{Y})E(Y_{N+1}) + \sum_{k'=2}^{N-1}\left[\left(\prod_{i=N-k'+2}^{N+1} E(Y_i)\right) Cov_{(N-k'+1)}(\boldsymbol{Y})\right] \\
&\qquad \textit{(we set } k' = k+1 \textit{)} \\
&= Cov_{(N+1)}(\boldsymbol{Y}) + \sum_{k'=1}^{N-1}\left[\left(\prod_{i=N-k'+2}^{N+1} E(Y_i)\right) Cov_{(N-k'+1)}(\boldsymbol{Y})\right]
\end{aligned}$$

[for $k = 1$, the term in the summation of (33) is equal to $Cov_{(N)}(\boldsymbol{Y})E(Y_{N+1})$].

Proof of Lemma 3.5

$$\begin{aligned}
Var\Big(\prod_{i=1}^{N} Y_i\Big) &= E\left[\prod_{i=1}^{N} Y_i - E\Big(\prod_{i=1}^{N} Y_i\Big)\right]^2 \\
&= E\left[\left(\prod_{i=1}^{N} Y_i - \prod_{i=1}^{N} E(Y_i)\right) - TCI(\boldsymbol{Y})\right]^2 \\
&= E\left[\prod_{i=1}^{N} Y_i - \prod_{i=1}^{N} E(Y_i)\right]^2 + TCI(\boldsymbol{Y})^2 - 2\,E\left\{TCI(\boldsymbol{Y})\Big[\prod_{i=1}^{N} Y_i - \prod_{i=1}^{N} E(Y_i)\Big]\right\} \\
&= E\left[\prod_{i=1}^{N} Y_i - \prod_{i=1}^{N} E(Y_i)\right]^2 = Var\left(\prod_{i=1}^{N} Y_i \Big| Independence\right) - TCI(\boldsymbol{Y})^2.
\end{aligned}$$

since $E\left\{TCI(\boldsymbol{Y})\Big[\prod_{i=1}^{N} Y_i - \prod_{i=1}^{N} E(Y_i)\Big]\right\} = TCI(\boldsymbol{Y})E\Big[\prod_{i=1}^{N} Y_i - \prod_{i=1}^{N} E(Y_i)\Big] = 0.$ $\square$

References

Aguilar, O. and West, M. (2000). Bayesian Dynamic Factor Models and portfolio allocation. *Journal of Business and Economic Statistics*, 18:338–357.

Baker, F. (1998). An investigation of the item parameter recovery characteristics of a Gibbs sampling procedure. *Applied Psychological Measurement*, 22:153–169.

Bartholomew, D., Knott, M., and Moustaki, I. (2011). *Latent variable models and factor analysis: a unified approach*. Wiley Series on Probability and Statistics. John Wiley and Sons, London, UK, 3rd edition.

Bock, R. and Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, 46:443–459.

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

Bratley, P., Fox, B. L., and Schrage, L. (1987). *A guide to simulation*. Springer, second edition.

Carlin, B. P. and Louis, T. A. (2000). *Bayes and Empirical Bayes methods for data analysis*. Chapman & Hall/CRC, second edition.

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

Congdon, P. (2010). *Applied Bayesian Hierarchical Methods*. Chapman and Hall/CRC.

DiCiccio, T. J., Kass, R. E., Raftery, A., and Wasserman, L. (1997). Computing Bayes Factors by combining simulation and asymptotic approximations. *Journal of the American Statistical Association*, 92(439):903–915.

Flegal, J. and Jones, G. (2010). Batch means and spectral variance estimators in markov chain monte carlo. *Annals of Statistics*, 38:1034–1070.

Fouskakis, D., Ntzoufras, I., and Draper, D. (2009). 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.

Friel, N. and Pettitt, N. (2008). Marginal likelihood estimation via power posteriors. *Journal of the Royal Statistical Society Series B (Statistical Methodology)*, 70(3):589–607.

Gelfand, A. E. and Dey, D. K. (1994). Bayesian Model Choice: Asymptotics and exact calculations. *Journal of the Royal Statistical Society. Series B (Methodological)*, 56(3):501–514.

Gelman, A. and Meng, X.-L. (1998). Simulating normalizing constants: From Importance sampling to Bridge sampling to Path sampling. *Statistical Science*, 13(2):163–185.

Geweke, J. and Zhou, G. (1996). Measuring the pricing error of the Arbitrage Pricing Theory. *Review of Financial Studies*, 9:557–587.

Gifford, J. A. and Swaminathan, H. (1990). Bias and the effect of priors in Bayesian estimation of parameters of Item Response Models. *Applied Psychological Measurement*, 14:33–43.

Goodman, L. A. (1962). The variance of the product of K random variables. *Journal of the American Statistical Association*, 57:54–60.

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

Jones, G., Haran, M., Caffo, B., and Neath, R. (2006). Fixed-width output analysis for Markov Chain Monte Carlo. *Journal of the American Statistical Association*, 101:1537–1547.

Kang, T. and Cohen, A. S. (2007). Irt model selection methods for dichotomous items. *Applied Psychological Measurement*, 31(4):331358.

Kass, R. and Raftery, A. (1995). Bayes factors. *Journal of the American Statistical Association*, 90:773–795.

Kim, S.-H., Cohen, A. S., Baker, F. B., Subkoviak, M. J., and Leonard, T. (1994). An investigation of hierarchical Bayes procedures in item response theory. *Psychometrika*, 59(3):405–421.

Koehler, E., Brown, E., and Haneuse, S. J.-P. A. (2009). On the assessment of Monte Carlo error in simulation-based statistical analyses. *The American Statistician*, 63(2):155–162.

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

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

Lord, F. M. (1980). *Applications of Item Response Theory to practical testing problems*. Erlbaum Associates, Hillsdale, NJ.

Lord, F. M. and Novick, M. R. (1968). *Statistical theories of mental test scores*. Addison-Wesley, Oxford, UK.

Meketon, M. S. and Schmeiser, B. W. (1984). "Overlapping batch means: Something for nothing?". in *Proceedings of the 1984 Winter Simulation Conference*, pages 227–230, Piscataway, NJ. Institute of Electrical and Electronics Engineers Inc.

Meng, X.-L. and Schilling, S. (2002). Warp Bridge Sampling. *Journal of Computational and Graphical Statistics*, 11(3):552–586.

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

Mislevy, R. (1986). Bayes modal estimation in Item Response Models. *Psychometrika*, 51:177–195.

Moustaki, I. and Knott, M. (2000). Generalized Latent Trait Models. *Psychometrika*, 65:391–411.

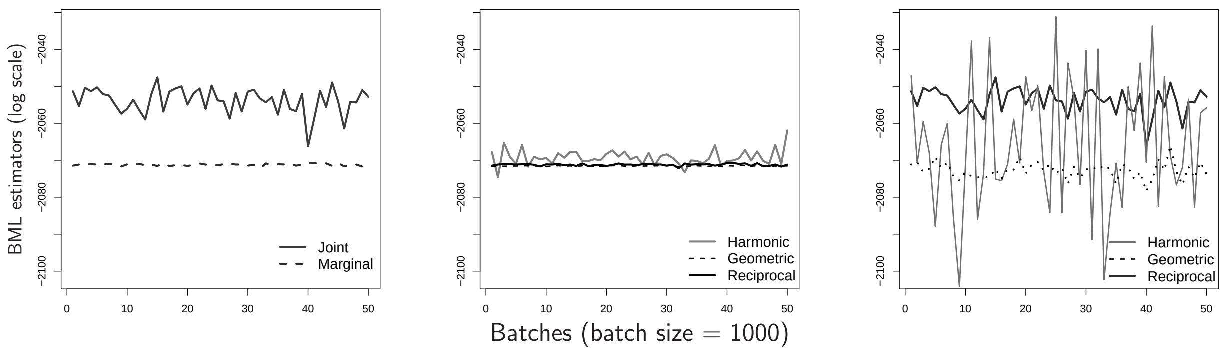
Ntzoufras, I., Dellaportas, P., and Forster, J. (2003). Bayesian variable and link determination for Generalised Linear Models. *Journal of Statistical Planning and Inference*, 111(1-2):165–180.

Patz, R. and Junker, B. (1999). A straightforward approach to Markov Chain Monte Carlo methods for Item Response Models. *Journal of Educational and Behavioral Statistics*, 24:146–178.

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

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

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

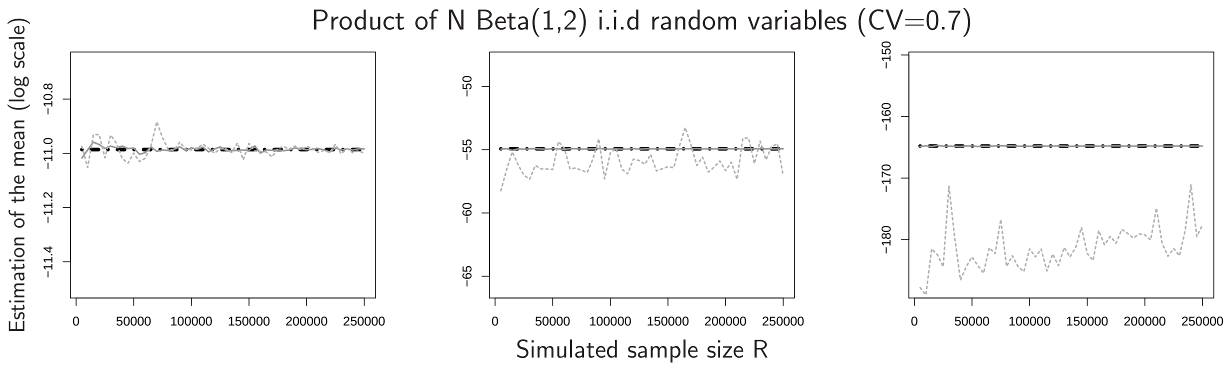


(a) Reciprocal estimator

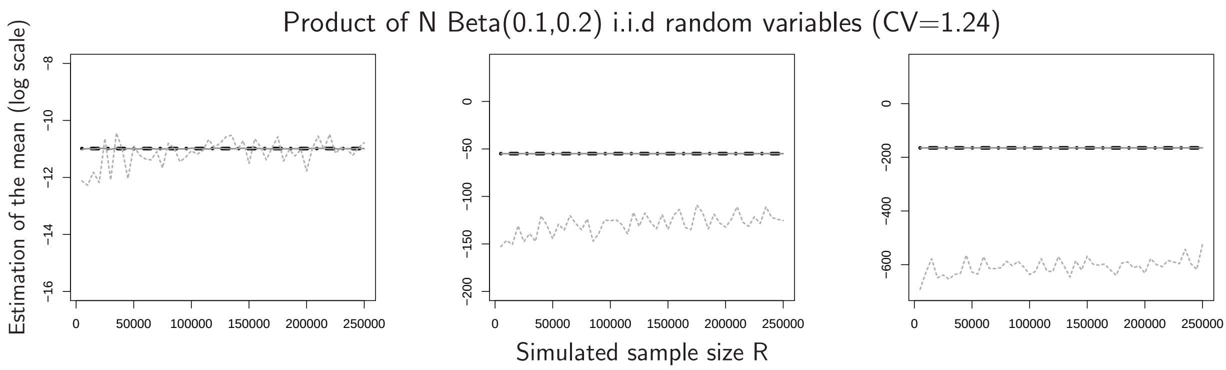
(b) Marginal estimators

(c) Joint estimators

Figure 2: Joint and marginal approaches for the reciprocal (RM), bridge harmonic (BH) and bridge geometric (BG) estimators of the BML (log scale), implementing a simulated data set with $p = 6$ binary items, $N = 600$ cases and $k = 2$ factors, over 50 batches.



(a) N=10 (b) N=50 (c) N=150



(d) N=10 (e) N=50 (f) N=150

Figure 3: The joint estimator $\widehat{I}_J$ (dashed grey line) and the marginal estimator $\widehat{I}_M$ (solid grey line) compared with the true mean (black dashed line) of the product of N i.i.d $Beta(\lambda_1, \lambda_2)$ variables, as the size of simulated the samples increases from 5000 to 250000 and for $N = 20, 50,$ and 150.