ATHENS UNIVERSITY
OF ECONOMICS AND BUSINESS

DEPARTMENT OF STATISTICS

POSTGRADUATE PROGRAM

EFFICIENT
BAYESIAN MARGINAL LIKELIHOOD ESTIMATION
IN GENERALISED LINEAR LATENT TRAIT MODELS

By

Silia Vitoratou

A THESIS
Submitted to the Department of Statistics
of the Athens University of Economics and Business
in partial fulfillment of the requirements for
the degree of Doctor of Philosophy in Statistics

Athens, Greece
2013

ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ
ΑΘΗΝΑΣ
ΤΜΗΜΑ ΣΤΑΤΙΣΤΙΚΗΣ

ΑΠΟΤΕΛΕΣΜΑΤΙΚΗ ΕΚΤΙΜΗΣΗ
ΠΕΡΙΘΩΡΕΙΑΣ ΠΙΘΑΝΟΦΑΝΕΙΑΣ ΚΑΤΑ BAYES
ΣΕ ΓΕΝΙΚΕΥΜΕΝΑ ΓΡΑΜΜΙΚΑ ΜΟΝΤΕΛΑ
ΛΑΝΘΑΝΟΥΣΩΝ ΜΕΤΑΒΛΗΤΩΝ

Σίλια Βιτωράτου

ΔΙΑΤΡΙΒΗ
Που υποβλήθηκε στο Τμήμα Στατιστικής
του Οικονομικού Πανεπιστημίου Αθήνας
ως μέρος των απαιτήσεων για την απόκτηση
Διδακτορικού Διπλώματος στη Στατιστική

Αθήνα, Ελλάδα
2013

to all volunteers, students and families of the education solidarity network ***#tutorpool*** *and to my nephews* ***Spyros*** *and* ***Andreas****, with all my love.*

#tutorpool

Δίκτυο Εθελοντισμού και Αλληλεγγύης για την Παιδεία

Synopsis

The term latent variable model (LVM) refers to a broad family of models which are used to capture abstract concepts (*unobserved* / *latent* variables or *factors*) by means of multiple indicators (*observed* variables or *items*). The key idea is that all dependencies among p observed variables are attributed to k unobserved ones, where $k << p$. That is, the LVM methodology is a multivariate analysis technique which aims to reduce the dimensionality, with as little loss of information as possible. Most importantly, the LVMs account for constructs that are not directly measurable, as for instance individuals' emotions, traits, attitudes and perceptions. In the current thesis, the LVMs are studied within the Bayesian paradigm, where model evaluation is conducted on the basis of posterior model probabilities. A key role in this comparison is played by the models' marginal likelihood, which is often a high dimensional integral, not available in closed form. The properties of the LVMs are implemented here in order to efficiently approximate the marginal likelihood.

In particular, Chapter 1 presents the origins and the basic ideas of the different types of latent variables models. The key aspects of the Bayesian analysis were outlined and the recent literature with respect to the LVMs is reviewed. Chapter 2 focuses on LVMs with binary data and describes the steps required in the Bayesian model assessment (prior specification, sampling from the posterior and Bayesian model comparison). In Chapter 3 two different formulations of the LVM marginal likelihood are presented. The variance components associated with each approach are specified and the factors that influence the corresponding errors are outlined. Additionally, the effect of the sample covariation on the estimators is explained and an index of the sample's divergence from independence is introduced, as a multivariate extension of covariance. Chapter 4 illustrates how the properties of the LVMs can be used to simplify often used estimators in the literature and to reduce the computational time required. Chapter 5 draws a link between Bayesian statistics and ideas initially presented in thermodynamics. It is shown that probability distribution divergences can be estimated via thermodynamic integration, while new thermodynamic marginal likelihood estimators are introduced. In Chapter 6, the methods discussed throughout in this thesis are illustrated in simulated and real life datasets. The thesis closes with a brief discussion on the current findings which prompt future research.

Σύνοψη

Ο όρος μοντέλα λανθανουσών μεταβλητών (ΜΛΜ) αναφέρεται σε μία ευρεία οικογένεια μοντέλων τα οποία χρησιμοποιούνται για να μετρήσουν αφηρημένες έννοιες (*μη παρατηρούμενες / λανθάνουσες* μεταβλητές ή *παράγοντες*) χρησιμοποιώντας πολλαπλούς δείκτες (*παρατηρούμενες* μεταβλητές ή *λήμματα*). Η κεντρική ιδέα είναι ότι οι εξαρτήσεις μεταξύ των p παρατηρούμενων μεταβλητών μπορούν να αποδοθούν σε k μή παρατηρούμενες μεταβλητές, όπου $k << p$. Κατά συνέπεια, η ΜΛΜ μεθοδολογία συνιστά μία πολυμεταβλητή ανάλυση που στόχο έχει να μειώσει τις διαστάσεις, με όσο το δυνατόν λιγότερη απώλεια πληροφορίας. Ακόμα σημαντικότερο είναι το γεγονός ότι τα ΜΛΜ μπορούν να μετρήσουν ποσότητες που δεν είναι άμεσα μετρήσιμες, όπως για παράδειγμα συναισθήματα, τάσεις, στάσεις και αντιλήψεις ατόμων.

Στην παρούσα διατριβή, τα ΜΛΜ μελετούνται σύμφωνα με τη στατιστική κατά Bayes, όπου η αξιολόγηση των μοντέλων γίνεται μέσω της εκ των υστέρων πιθανότητας. Βασικό ρόλο σε αυτό διαδραματίζει η περιθώρεια πιθανοφάνεια του εκάστοτε μοντέλου, η οποία συχνά είναι ένα πολυδιάστατο ολοκλήρωμα το οποίο δεν υπολο΄γιζεται σε κλειστή μορφή. Σε αυτή την εργασία χρησιμοποιούνται οι ιδιότητες των ΜΛΜ προκειμένου να εκτιμηθεί αποτελεσματικά η περιθώρεια πιθανοφάνεια.

Συγκεκριμένα, στο Κεφάλαιο 1 παρουσιάζονται οι απαρχές και οι βασικές ιδέες των διαφορετικών τύπων ΜΛΜ. Παρουσιάζονται επίσης τα βασικά σημεία της ανάλυσης κατά Bayes και γίνεται αναδρομή στη σύχρονη βιβλιογραφία. Το Κεφάλαιο 2 εστιάζει στα ΜΛΜ με δίτιμες μεταβλητές και περιγράφει τα βασικά σημεία της ανάλυσης κατά Bayes (επιλογή της εκ των προτέρων κατανομής, δειγματοληψία από την εκ των υστέρων κατανομή και αξιολόγηση του μοντέλου). Στο κεφάλαιο 3 παρουσιάζονται δύο εναλλακτικές μορφές της περιθώρειας πιθανοφάνειας. Υπολογίζονται οι συνιστώσες μεταβλητότας που αφορούν την κάθε μία από τις δύο προσεγγίσεις καθώς και οι παράγοντες που τις επηρεάζουν. Επιπλέον, εξηγείται ο ρόλος της δειγματικής συνδιασποράς και παρουσιάζεται ένας δείκτης απόκλισης από την ανεξαρτησία, ως πολυδιάστατο ανάλογο της συνδιασποράς. Στο Κεφάλαιο 4 οι ιδιότες των ΜΛΜ χρησιμοποιούνται για να απλοποιήσουν γνωστούς εκτιμητές της περιθώρειας πιθανοφάνειας, μειώνοντας έτσι το χρόνο που χρειάζεται για τον υπολογισμό τους. Στο Κεφάλαιο 5 παρουσιάζεται η στενή σχέση της στατιστικής κατά Bayes με τις ιδέες που έχουν αναπτυχθεί στο χώρο της Θερμοδυναμικής. Αποδεικνύεται ότι οι αποκλίσεις μεταξύ κατανομών πιθανοτήτων μπορούν να εκτιμηθούν μέσω της Θερμοδυναμικής ολοκλήρωσης, ενώ παρουσιάζονται νέοι εκτιμητές της περιθώρειας πιθανοφάνειας. Στο Κεφάλαιο 6, οι μέθοδοι που παρουσιάζονται σε αυτή την εργασία, εφαρμόζονται και συγκρίνονται σε προσομοιωμένα και σε πραγματικά δεδομένα. Η διατριβή ολοκληρώνεται με μία σύντομη συζήτηση των σημείων που χρήζουν μελλοντικής έρευνας.

Contents

List of Tables

List of Figures

Abbreviations

LTM	latent trait model
LPM	latent profile model
LCM	latent class model
NLFM	normal linear factor model
SEM	srtuctural equation model
LISREL	linear structural relations models
IRT	item response theory
GLLVM	generalized linear latent variable model
GLLTM	generalized linear latent trait model
GLM	generalized linear model
NOM	normal ogive model
UVA	underlying variable (approach)
EM	expectation-maximization algorithm
NR	Newton-Raphson algorithm
MH	Metropolis-Hastings algorithm
MG	Metropolis-within-Gibbs algorithm
MCMC	Markov chain Monte Carlo
BF	Bayes Factor
PBE	point-based estimators
BSE	bridge sampling estimators
PSE	path sampling estimators
BML	Bayesian marginal likelihood
CJ	Chib and Jeliazkov
CJ^I	independence Chib and Jeliazkov
LM	Laplace-Metropolis
GC	Gaussian copula
AM	arithmetic mean estimator
HM	harmonic mean estimator
RM	reciprocal mean estimator

BH	bridge harmonic estimator
BG	bridge geometric estimator
TI	thermodynamic integration
NTI	normalized thermodynamic integral

Notation

N	number of individuals
p	number of observed items
k	number of latent variables
$\mathbf{Y}$	observed data matrix
y_{ij}	response of the i-th $i$ndividual $(i = 1, ..., N)$ to the j-th observed variable $(j = 1, ..., p)$
$\mathbf{Z}$	column vector of the latent variables Z_ℓ $(\ell = 1, ..., k)$
f	$f(\cdot\|\cdot)$ denotes the conditional density of the data dependent on the parameter vector(s
$\pi(\cdot)$	denotes a prior density
$f(\mathbf{Y})$	marginal (integrated) likelihood
$g(\boldsymbol{\vartheta})$	importance (or reference) function based on the posterior output
η_j	linear predictor for each Y_j
$\upsilon_j(\cdot)$	link function for each Y_j
$\mu_j(\mathbf{Z})$	mean value of the item j, conditional on the latent variables, that is, $E(Y_j\|\mathbf{Z})$
$\boldsymbol{\alpha}$	difficulty parameter
$\boldsymbol{\beta}$	discrimination parameter
$\boldsymbol{\vartheta}$	vector of the item parameters $\{\boldsymbol{\alpha}, \boldsymbol{\beta}\}$
$\mathbf{X}$	sufficient statistic
c_Φ	Gaussian copula
$K(\cdot\|\cdot)$	denotes the kernel of the M-H algorithm
$a(\cdot\|\cdot)$	the M-H acceptance probability
$q(\cdot\|\cdot)$	proposal density
$\boldsymbol{\theta}_{\setminus j}$	the parameter vector $\boldsymbol{\theta}$ without $\boldsymbol{\theta}_j$
$\mathcal{A}(\cdot)$	bridge function
$q(\boldsymbol{\theta}'\|\,\boldsymbol{\theta})$	proposal density
$a(\boldsymbol{\theta}\|\,\boldsymbol{\theta})$	acceptance probability
$\boldsymbol{\Sigma}$	covariance matrix
$\mathcal{I}(\mathbf{b})$	information matrix
$\mathbf{I}_{(k\times k)}$	identity matrix
$KL(p_1 \parallel p_0)$	Kullback - Leibler divergence

$J(p_1, p_0)$	Kullback - Leibler divergence
$cH(p_1 \parallel p_0)$	cross entropy
$H(p_1)$	differential entropy
$\mathcal{KL}_t$	functional KL-divergence of order t
$C_t(p_1 \parallel p_0)$	Chernoff divergence of order t
$\mu(t)$	Chernoff coefficient
$Bh(p_1, p_0)$	Bhattacharyya distance
ρ_B	Bhattacharyya coefficient
$He(p_1, p_0)$	Hellinger-Bhattacharyya distance
$R_t(p_1 \parallel p_0$	Rényi t-divergence
$T_t(p_1 \parallel p_0$	Tsallis t-relative entropy
$\mathcal{H}_t(\boldsymbol{\theta})$	Hamiltonian (energy function)
$\mathcal{T}$	temperature schedule
$q_t^{PP}(\boldsymbol{\theta})$	prior-posterior path
$q_t^{IP}(\boldsymbol{\theta})$	importance-posterior path
$q_t^{MS}(\boldsymbol{\theta})$	model switch path
$(Q \circ q)_t^{IP}(\boldsymbol{\theta})$	importance-posterior BF quadrivial
PP_T	thermodynamic prior-posterior identity
PP_S	stepping-stone prior-posterior identity
IP_T	thermodynamic importance-posterior identity
IP_S	stepping-stone importance-posterior identity
MS_T	thermodynamic model-switch identity
MS_S	stepping-stone model-switch identity
Q_{PP_T}	thermodynamic prior-posterior quadrivial identity
Q_{PP_S}	stepping-stone prior-posterior quadrivial identity
Q_{IP_T}	thermodynamic importance-posterior quadrivial identity
Q_{IP_S}	stepping-stone importance-posterior quadrivial identity

Chapter 1

Latent variable models: classical and Bayesian approaches

> *‘‘Two variable organs are said to be co-related when the variation of the one is accompanied on the average by more or less variation of the other, and in the same direction. [...] co-relation must be the consequence of the variations of the two organs being partly due to common causes. If they were wholly due to common causes, the co-relation would be perfect, as is approximately the case with the symmetrically disposed parts of the body."*
>
> Francis Galton, 1888.*

*Francis Galton (1822-1911), among other things, was an anthropologist, geographer, inventor, meteorologist, proto-geneticist, psychometrician, and statistician. Bartholomew, 2011 states that this quote was brought to his attention by J.Aldrich (University of Southampton) and is the oldest one reported, that describes the core idea behind latent variable models.

1.1 Origins of the latent variable models

Depending on the nature of the observed data and on the type assumed for the latent variables, the LVMs are classified in four basic categories. The most popular and often used model is the normal linear factor model (NLFM). Factor analysis refers to metrical data and it also assumes metrical (continuous or discrete) latent variables. If, instead, the data at hand are categorical, the corresponding model is the latent trait model (LTM), which is often referred to as factor analysis model with binary/ordinal/polytomous or ranked data. However, the latent variables can be categorical as well; if the data are metrical, the latent profile model (LPM) is implemented, otherwise the suitable model is the latent class model (LCM). The four types of LVMs were motivated by different problems and therefore their origins vary. A thorough historical review can be found in Bartholomew et al. (2011) while some of interesting points are briefly reported below.

The idea of "common causes of variability" in a set of variables dates back to Galton (1888), as shown on his quote on the top of this chapter. It was Spearman however, who in his path-breaking paper (Spearman, 1904) first conceptualized the idea that the score of an individual in a test, can be partially attributed to a *common* factor (ability) that underlies the manifest items, and to a *specific* factor which represents error. Spearman's work inspired Thurstone and his colleagues in the early '30s, where the idea of *psychometrics* began to gain increased interest in psychology. Thurstone (1931, 1947) generalized Spearman's model to allow for more than one common factors and his work is considered as a milestone in factor analysis. Motivated also by problems that occur in psychometrics, Jöreskog (1970) generalized the factor analysis model within the context of the srtuctural equation model (SEM), which allows for linear relationships among the latent variables, namely the linear structural relations models (LISREL). In SEM's terminology, the latent variables are called *endogenous* variables, as opposed to the manifest items that are considered *exogenous*.

Beyond psychometrics, researchers on educational testing (for instance Birnbaum 1968), developed also very important techniques in order to measure a responder's ability on a topic of interest (for a historical review see Hambleton et al. 1991). As opposed to the psychometricians' point of view, in educational testing the interest lays mostly on each individual's position on the latent scale rather than the factor structure itself. The data are typically binary ("right" or "wrong") as for instance in the case of the Rasch model (Rasch, 1960), known for its simplicity and appealing statistical properties (see also Andersen, 1980). A more advanced class of models in educational testing is based on the item response theory (IRT). The origins of the IRT are traced back to Binet et al.'s (1916) and the early readings Richardson (1936), Lord (1952) and Lord and Novick (1968). An inclusive literature review on the IRT models can be found in Bock (1997).

Significant contribution on the LVMs was held also by researchers in Sociology. In this case the models developed assumed categorical latent variables and until recently were considered entirely separate from factor analysis (Bartholomew et al., 2011). The work of Lazarsfeld and Henry (1968) put the foundations on the latent class and latent profile models, followed by the work Everitt (1984), Langeheine and Rost (1988) and Heinen (1996), among others. The idea of a latent variable is present also in the literature of other areas of statistics, often under alternative names. Lee and Nelder (2009) list the terms *random effects*, *latent processes*, *factor*, *missing data* among others. Bartholomew (2011) highlights the resemblance of the LVMs with the *mixture models*, the *hidden variables* (implemented in discrete time series) and even the so-called *unobserved heterogeneity* in econometrics.

Nowadays, the LVMs are broadly used in most social sciences, in econometrics, in educational testing and generally whenever a theoretical construct lacks of direct measurement. Due to the advances in computing, a plethora of sophisticated methods have been developed in the recent years and the LVMs' use became a common practise in applied research of various scientific fields. The basic aspects of the LVM framework of all types are presented in the next section.

1.2 Basic aspects and key ideas

A latent variable model uses the information available from the observed data to extract information for the unobserved construct of interest. To rephrase this formally, the joint distribution of the manifest variables is implemented in order to assess the distribution of the latent ones. Even in the classical approach, this is achieved via the Bayes theorem

$$\pi(\mathbf{Z}|\mathbf{Y}) = \frac{f(\mathbf{Y}|\mathbf{Z})\,\pi(\mathbf{Z})}{f(\mathbf{Y})}, \tag{1.1}$$

where $\mathbf{Y}_{N\times p}$ (for simplicity $\mathbf{Y}$) denotes a data matrix whose elements y_{ij} correspond to the response of the i-th $i$ndividual ($i = 1, ...,N$) to the j-th observed variable ($j = 1, ...,p$). In principle, the LVMs of all types assume that the dependencies among the Y_j s are due to the existence of k latent variables, denoted hereafter with $\mathbf{Z}$= $(Z_1, ..., Z_k)$. As only the items Y_j s can be observed, any inference must be based on their joint distribution whose density may be expressed as

$$f(\mathbf{Y}) = \int f(\mathbf{Y}|\mathbf{Z})\,\pi(\mathbf{Z})\,d\,\mathbf{Z}, \tag{1.2}$$

where $\pi(\mathbf{Z})$ is the *prior distribution* of the latent variables and $f(\mathbf{Y}|\mathbf{Z})$ is the *conditional distribution* of the observed variables given the latent. These two distributions are the core features of a latent variable model and completely determine its type. They are considered as two distinct parts, each playing an important role, as discussed below.

To begin with, the prior distribution of the latent variables reflects whether they are assumed to be metrical (FA and LTM) or categorical (LCM and LPM). For the factor analysis and latent trait models, the choice of the prior $\pi(\mathbf{Z})$ is arbitrary and a matter of convention and convenience. In most cases, it is assumed to be the standard normal distribution and any shifts in the location and/or the scale are considered to be absorbed by the model (see Heinen 1996, p. 105 for a counter example). When categorical latent variables are assumed, the prior is not completely arbitrary since it consists of probability masses, located at each category of the latent variable(s), whose sum is unit. Either way, $\pi(\mathbf{Z})$ affects the derived estimators but it does not affect the dimensionality reduction endorsed by the LVM, as will be shown in the next section.

The second part of a LVM, namely the conditional distribution of the manifest variables $f(\mathbf{Y}|\mathbf{Z})$, reflects the nature of the data. For instance, in models with binary data a Bernoulli distribution is considered and in the case of polytomous data a Multinomial distribution is the natural choice. It turns out that the members of the exponential family are particulary useful in order to describe the $f(\mathbf{Y}|\mathbf{Z})$ at each of the four types of LVMs. Beyond the type of data, the LVMs assume that given the latent variables, the observed ones are independent and therefore the conditional distribution is given by

$$f(\mathbf{Y}|\mathbf{Z}) = \prod_{j=1}^{p} f(Y_j|\mathbf{Z}), \tag{1.3}$$

since all data dependencies are attributed to the existence of the latent variables. This assumption is fundamental in the LVMs and is often refer to as the *local independence assumption* or the *axiom of the conditional independence*. The local independence assumption rises the question whether $f(\mathbf{Y})$ admits the presentation

$$f(\mathbf{Y}) = \int \pi(\mathbf{Z}) \prod_{j=1}^{p} f(Y_j|\mathbf{Z}) \, d\,\mathbf{Z}. \tag{1.4}$$

In other words, if the hypothesised k latent vectors are sufficient to explain all dependencies among the p manifest variables, which is practically a question of dimensionality reduction. If so, then the information on the latent variables available from the data at hand, is assessed via their *posterior* distribution

$$\pi(\mathbf{Z}|\mathbf{Y}) = \frac{\pi(\mathbf{Z}) \prod\limits_{j=1}^{p} f(Y_j|\mathbf{Z})}{f(\mathbf{Y})}, \tag{1.5}$$

according to the Bayes theorem and (1.3).

The basic question imposed in the LVMs is whether the k latent variables assumed are sufficient to explain the observed dependencies among the p observed items. In the next section a general model that unifies the different types of LVM and addresses successfully this question is presented.

1.3 Generalised linear latent variable models

Bartholomew and Knott (1999) described a broad model setting that unifies the LVMs, based on the framework of the generalized linear model (GLM). The model is called the generalized linear latent variable model (GLLVM) and it includes the four categories of the LVMs discussed in the previous section as special cases (see also Skrondal and Rabe-Hesketh, 2004). As in the case of the GLM (McCullagh and Nelder, 1989), the GLLVM assumes that the response variables are linear combinations of the latent ones and it consists of three components:

(a) the multivariate *random component*: where each observed variable Y_j, $(j = 1, ..., p)$ has a distribution from the exponential family (Bernoulli, Poisson, Multinomial, Normal, Gamma),

(b) the *systematic component*: where the latent variables Z_ℓ, $\ell = 1, ..., k$, produce the linear predictor η_j for each Y_j

$$\eta_j = \alpha_j + \sum_{\ell\ =1}^{k} \beta_{\ell j}\, Z_\ell \ \text{ and}$$

(c) the link function $\upsilon_j(\cdot)$, that connects the previous two components

$$\upsilon_j\Big\{\mu_j\,(\mathbf{Z}\,)\Big\} = \eta_j, \text{ for } j = 1, \ldots p \text{ and } \mu_j\,(\mathbf{Z}) = \mathrm{E}(\mathrm{Y}_j|\mathbf{Z}).$$

As opposed to the GLMs, the random component in (a) is always multivariate and the regressors in the systematic component (b) are unobserved. The link function in (c) can be any monotonic differentiable function and may be different for each of the Y_j s, implying that the GLLVM includes models with mixed type data. Hereafter, the term *item parameters* will be used for the parameter vector $\boldsymbol{\vartheta}$=$\{\boldsymbol{\alpha}, \boldsymbol{\beta}\}$. The parameters α_j and $\boldsymbol{\beta}_j$ are often referred to as the *difficulty* and the *discrimination* parameters (respectively) of the item j, names that are inherited by the educational testing literature. In factor analysis, the model parameters are referred to as the *loadings* of the observed items on the latent factors and they reflect the contribution of each item in the construction of a particular factor.

As an example, consider the case of a latent trait model with p binary items and k latent variables. The conditionals $f(Y_j|\mathbf{Z})$ in (1.3) are in this case Bernoulli($\mu_j(\mathbf{Z})$), where $\mu_j(\mathbf{Z})$ is the conditional probability of a positive response to the observed item j. The logistic model is used for the response probabilities

$$\mathrm{logit}\{\mu_j\,(\mathbf{Z})\} = \alpha_j + \sum_{\ell\ =1}^{k}\beta_{\ell j}\, Z_\ell, \quad (j = 1, \ldots p). \tag{1.6}$$

In the special case where $k = 1$ (one factor is assumed), (1.6) corresponds to the two-parameter model (2-PL) which is presumably the most well studied model in IRT. This model is referred to as the item characteristic curve (ICC) or item response function (IRF) in the psychometric literature. If additionally the *loadings* β_j are considered equal ($\beta_1 = \beta_2 = ... = \beta_p$) then (1.6) coincides with the one-parameter model (1-PL) in IRT, that is, the Rasch model (Rasch, 1960).

It turns out that the latent trait models with binary data that were mentioned in Section 1.2 are easily derived from (1.6). With regard to the other types of LVMs, their derivation from the GLLVM model can be found in Bartholomew et al. (2011). For the case of the generalised latent trait models (GLTM) in particular, which account for several types of categorical data, a complete study can be found in Moustaki and Knott (2000). Additionally, models with mixed types of data (Moustaki, 1996) can be also handled within the framework of the GLLVM (Bartholomew et al., 2011) as well as models with non-linear terms (Rizopoulos and Moustaki, 2008).

1.3.1 Sufficient statistics and component scores

In order to reduce the dimensionality of the observed variables, the objective is to find k functions of $\mathbf{Y}$, say $(X_1, X_2, ..., X_k) = \mathbf{X}$, such that the conditional distributions of the observed variables given $\mathbf{X}$ not to be dependent on $\mathbf{Z}$. Then $\mathbf{X}$ is a sufficient statistic (for necessary and sufficient conditions, subject to weak regularity conditions, see Barankin and Maitra, 1963, Theorems 5.1, 5.2 and 5.3). Bartholomew et al. (2011) illustrate that this is the case in the GLLVM, where each Y_j follows a distribution from the exponential family and therefore the conditional distributions (1.3) are written as follows

$$f(Y_j|\mathbf{Z}) = F_j(Y_j)\, G_j(n_j)\, \exp\big\{n_j\, h_j(Y_j)\big\}, \tag{1.7}$$

where n_j is a linear combination the latent vectors. Substituting (1.7) in (1.5) yields

$$\pi(\mathbf{Z}|\mathbf{Y}) = \frac{\pi(\mathbf{Z})\big\{\prod_{j=1}^{p} F_j(Y_j)G_j(n_j)\big\}\, \exp\big\{\sum_{j=1}^{p} n_j\, h_i(Y_j)\big\}}{\int \big\{\prod_{j=1}^{p} F_j(Y_j)G_j(n_j)\big\}\, \exp\big\{\sum_{j=1}^{p} n_j\, h_j(Y_j)\big\}\, d\mathbf{Z}}$$

$$\propto \pi(\mathbf{Z})\prod_{j=1}^{p} G_j(n_j)\, \exp\sum_{\ell=1}^{k} Z_\ell\, X_\ell,$$

where

$$X_\ell = \sum_{j=1}^{p} \beta_{\ell j} h_j(Y_j), \quad \ell = 1, ..., k. \tag{1.8}$$

As Bartholomew et al. (2011) denote, the first important thing to notice here is that the posterior distribution of the latent variables depends on $\mathbf{Y}$ only through the k-dimensional vector $\mathbf{X}$. Hence, the latter is a Bayesian sufficient statistic that represents the dimensionality reduction effected by the GLLVM. In principle however, a sufficient statistic does not include the model parameters, as opposed to (1.8). In fact, this is true only in the case of the Rasch model mentioned previously, where all information regarding the latent variables is included in the total score (sum-score) of the observed variables (Andersen 1973,1977). However, Bartholomew et al. (2011, chapter 2, page 20) clarify that in the construction of the *components* X_ℓ, $(\ell = 1, ..., k)$, the model parameters are considered fixed and they explain that *the term "sufficient" has exactly the meaning we wish to convey and so we will use it in this extended sense.* The second important thing to notice in (1.8) is that the prior distribution of the latent variables does not play role in the dimensionality reduction. Hence, the arbitrariness in the selection of the prior, especially in the case where metrical latent variables are considered, does not affect the information we derive a-posteriori with regard to the number of the latent variables involved.

The GLLVM framework makes it therefore possible to avoid the calculation of the posterior mean of the latent variables, and thus the numerical integrations involved, by using instead the component scores. That is particularly useful in applied research where the score of each individual is used in subsequent analyses.

Returning to the example of the LTMs with binary data (1.6), the Bernoulli distribution may be written in the form (1.7) by setting $F_j(Y_j) = 1$, $G_j(n_j) = 1 - \mu_j(\mathbf{Z})$, $h_j(Y_j) = Y_j$ and $n_j = \text{logit}[\mu_j(\mathbf{Z})]$ and therefore

$$X_\ell = \sum_{j=1}^{p} \beta_{\ell j} Y_j, \quad (\ell = 1, ..., k).$$

Hence, the components are weighted sums of the manifest variables, in proportion to the latter's contribution to the common factor(s). The expression (1.9) holds only for the logit link function but not for the probit. When the probit link is used instead, the corresponding model is often referred to as the normal ogive model (NOM) and is a special case of the underlying variable (approach) (UVA), which is alternative to the GLLVM for categorical data.

1.3.2 Rotation

The GLLVM presented in Section 1.3 can be described in a matrix form as follows

$$\upsilon\{\mu(\mathbf{Z})\} = \mathbf{A} + \mathbf{BZ}$$

The matrix representation implies that the model does not have a unique solution. In particular, for any orthogonal matrix $\mathbf{U}$ such as $\mathbf{U} \times \mathbf{U}' = \mathbf{I}$ (1.9) becomes

$$\upsilon\{\mu(\mathbf{Z})\} = \mathbf{A} + \mathbf{BUU'Z}.$$

Hence, there are infinite possible solutions with respect to the model's slope parameters, $\mathbf{B}^* = \mathbf{B} \times \mathbf{U}$. Steiger (1979) provides a historical review from the early 20th century with respect to factor analysis and factor indeterminacy per ce. It occurs that at the early years the factor indeterminacy was not actually addressed as an issue. Wilson (1928) first illustrated that different sets of factor scores could fit the model for the same data. Ever since, a number of different approaches (geometrical or numerical) have been followed in order to describe, interpret and solve factor indeterminacy, leading even to publications against factor analysis (see for example Guttman, 1955).

The problem is addressed with constraints, which are imposed to the model to ensure a unique solution. One of the first readings is Anderson and Rubin (1956) who give sufficient or necessary conditions to stop rotation. Ever since, there have been proposed solutions that are based on imposing zero or non-zero constraints on the loadings matrix (see for instance Lawley and Maxwell, 1963; Jöreskog, 1969, Dunn, 1973; Algina, 1980 among others).

The indeterminacy of the latent variable models has no affect on the dimensionality reduction, that is, the determination of the number of the latent variables. Moreover, the infinite solutions become a feature of the model, rather than a drawback, when it comes to interpretation. It gives the advantage to be able to choose the solution that is more meaningful, for the problem at hand. For that purpose, orthogonal and oblique rotation techniques have been developed in the literature (for details see Bartholomew et al., 2011, Section 2.11). For further readings on identifiability, for either oblique or orthogonal solutions, see among others Williams (1978), Elffers et al. (1978), Bekker (1986), Sato (1991), Browne (2001) and references within.

1.3.3 Parameter estimation and goodness of fit

In principle, all models can be fitted using the maximum likelihood method (ML). For the general case, the joint distribution of the data (1.4) using the notation of the exponential family (1.7) becomes

$$f(\mathbf{Y}) = \int \{\prod_{j=1}^{p} F_j(Y_j) G_j(n_j)\} \exp\{\sum_{j=1}^{p} n_j \, h_j(Y_j)\} \, d\mathbf{Z},$$

according to which, the marginal distribution for each Y_j is then given by

$$f(Y_j) = F_j(Y_j) \, E_{n_j}\left[G_j(n_j) \exp\left\{ \sum_{j=1}^{p} n_j \, h_j(Y_j) \right\} \right].$$

The direct maximization of the marginal likelihood is possible yet rigorous (Bartholomew et al., 2011). Instead, Dempster et al.'s (1977) expectation-maximization algorithm (EM) is implemented in most cases in order to carry out the parameter estimation. The EM was first employed in the IRT context by Bock and Aitkin (1981) and in factor analysis by Rubin and Thayer (1982). A version suitable for the latent class model was also given by Goodman (1978). The algorithm refers to incomplete data and in the case of the LVMs the latent variables are treated as such. Initially, the formulation of the joint likelihood $f(\mathbf{Y}, \mathbf{Z})$ is required. At the E-step the log likelihood is replaced by its expected value conditional on the latent variables, using initial values for the model parameters $\boldsymbol{\theta}$. At the M-step the log likelihood is maximised to give new values for the parameters and the whole procedure iterated until convergence. It should be noted that the Newton-Raphson algorithm (NR) can be also used for direct maximisation. The NR can be faster than the EM yet more sensitive to the initial values. A combination of the two approaches often provides the best solution. The parameter estimation procedure is described in detail in Bartholomew et al. (2011) as it applies in the different categories of LVMs.

With respect to the goodness of fit, chi-square based tests have been developed in the literature of the LVMs (Bartholomew et al., 2008). In the case of the factor model a plethora of fit indices can be found in commercial software that may or not assume normality of the observed variables. The covariance of the correlation matrix of the metrical observed variables is implemented and compared with the one suggested by the model. In the case of categorical data, on the other hand, the goodness of fit indices are focused on the differences between the frequencies of the observed response patterns as compared to the ones implied by the model. An additional difficulty occurs with respect to the sparseness of the data. For instance, for p binary items there are 2^p possible response patterns and therefore often occur empty cells in the corresponding contingency tables. This phenomenon is even more intense in the case of polytomous data. To address these difficulties, tests that employ two and three way margins have been developed in order to evaluate the goodness of fit.

1.4 The Bayesian approach

The Bayesian and classical approaches address statistical challenges by completely different perspectives. Extended bibliography exists that outlines their differences both in philosophy and practise, which cannot be summarized in a few pages. Intuition is benefited however by historical reviews on the statistical thinking, which include the milestone ideas for both approaches such as Fienberg (2006). Chronicles on the Bayesian statistics are also provided by Stigler (1983) and Dale (1999) which colorfully describe the methodological paths followed during the past century until the so-called Bayesian explosion, at the early 90s. Gelman and Shalizi (2013) moreover, review the two schools of thinking from a purely philosophical point of view.

While a number of methodological and philosophical differences exist between the two approaches, at the end of the day they can be summarized in the way each approach *views* the data and the unknown parameters. That is, while the classical approach aims to describe the variability of the data for fixed parameters, the Bayesian approach aims to the exact opposite: it describes the variability of the parameters, considering the data as fixed. That quite explains the term "inverse probability" that was used to describe the Bayesian framework in the early years.

In Bayesian statistics, the parameters are stochastic and distributed according to the prior, $\pi(\cdot)$. The prior reflects the researcher's subjective opinion or previously obtained knowledge for the parameters. In the absence of such information, the prior can be non-informative (vague). Once the data have been observed, the objective of the Bayesian approach is to update the prior knowledge conditional on the data at hand, which is literally the posterior distribution (*Bayesian learning*). The Markov chains Monte Carlo (MCMC) revolution combined with the modern achievements in computer science, made this transition from the prior to the posterior knowledge feasible where used to be rigorous. The posterior output is then implemented in order to describe the parameters, as well as to evaluate the imposed model. The three steps of the Bayesian assessment, namely the *prior specification*, the *derivation of the posterior output* and finally the *model assessment* are discussed in this section, focused on applications in the various types of LVMs.

1.4.1 Latent variable models and the Bayesian paradigm

In Section 1.2 it was denoted that the Bayes theorem is used in the LVMs' framework in order to assess the latent variables. Clearly, the latent variables are handled in a Bayesian manner, starting from a prior density and concluding to their posterior. Hence, one could argue that there is no purely classical approach for the LVMs and in that sense, the LVMs are actually seen in either a semi or a fully Bayesian perspective. The difference lies on the way the item parameters $\boldsymbol{\vartheta}$ are treated. Specifically, a fully Bayesian approach requires $\boldsymbol{\vartheta}$ to be also stochastic, associated with a prior probability. That approach gained increased interest the last decades in the LVM field. In the following sections, the main features of the Bayesian approach are discussed and recent literature is reviewed with regard to their implementation in latent variable models.

1.4.1.1 On prior specification

According to the Bayes theorem, the posterior knowledge is by definition given, up to a constant, by *prior times the likelihood.* Hence, any inference is based on the relative proportion of the information provided by the prior and the data. The prior can be informative (subjective) to represent existing knowledge with regard to the parameters or, otherwise, non-informative (vague). Large literature is available, concerned with sensitivity of the posterior quantities related to the prior specification of a model (see for instance Gustafson, 1996; Berger, 1996 and references within). With regard to the non informative priors, flat or diffuse distributions are used, with large variances that reflect our uncertainty for the parameters. In such cases, the influence of the data becomes relatively stronger, meaning that the likelihood contributes more in the formation of the posterior that the prior and the derived estimates tend to be closer to the maximum-likelihood ones. For example, if the parameters are equally likely then a uniform prior is preferred (*principle of insufficient reason*). If the parameter space is continuous then a generalization is obtained by a flat prior.

Several families of priors can be found in the Bayesian literature, each employing certain strategies to account for the prior uncertainty. Kass and Wasserman (1996) use the term *reference prior* to describe Jeffreys's (1961) idea on using a standard of reference, in analogy with other scientific fields. Rather than representing our ignorance, reference priors are chosen by convention under some objective rules for a particular situation. Berger et al. (2009) recently provided a formal definition of reference priors (see also references within). Berger and Pericchi (1996) explore the idea of training samples, in order to overcome the arbitrariness arising when an informative prior is *improper* (non-integrable). According to the authors, the initial data set may be divided in two subsets and use the first as a *training sample* in order to convert the improper non-informative prior into a proper one. Then the second dataset is implemented in the main analysis with the posterior playing the role of the prior. Priors that emerge under this methodology are called *intrinsic priors*. The training sample is said to be *proper* if the corresponding integrated likelihood under the compiling models is finite and *minimal* if it is the smallest possible subset with this property. This can be expanded by implementing *imaginary data* (Spiegelhalter and Smith 1982) instead of a subset of the data at hand. Closely related to the concept of training samples are also the *expected posterior* priors Perez and Berger (2002). The method is based on the concept that a prior can be an average of previously assessed posteriors, over a suitable measure m. Finally, *historical data* are implemented in the case of the *power priors*, introduced by Ibrahim and Chen (2000) and Chen et al. (2000). The prior elicitation is based on the availability of historical data and a scalar quantity quantifying the uncertainty. As Neuenschwander et al. (2009) describe in a recent note, the power prior raises the likelihood of the historical data to the power parameter (scalar) which quantifies the discounting of the historical data due to heterogeneity between trials.

For the LVMs, the prior plays yet another important role: it is implemented in order to ensure a unique solution. Constant or truncated priors are assigned to selected item parameters, in a similar manner that constrains are imposed in the case of the classical approach in order to stop the rotation. With regard to the latent variables, the choice of independent standard normal distributions is also global when it comes for the metrical case. Such restrictions in the prior are always present in the corresponding literature and therefore no explicit mention will be made hereafter.

In factor analysis models, normal prior distributions are typically employed for the model parameters (loadings) and inverse gamma prior distributions for the residual variances (Lopes and West, 2004; Lee and Song, 2001). These choices are quite popular since they lead to conditionally-conjugate forms. Other options are recently discussed, as for instance in Aguilar and West (2000) who explore the use of reference priors on Bayesian dynamic factor modelling, in Polasek (2000) who implements Wishart priors on the covariance matrix and in Lopes (2003) who comments on the use of the expected posterior priors. Ghosh and Dunson (2009) denote that the posterior distribution is improper (in the limiting case) as the prior variance for the normal and inverse-gamma components increases and they suggest the use of heavily tailed default priors. Related to the choice of the prior is also the recent work of Frühwirth-Schnatter and Lopes (2010, see also references within) where the issues of the model identifiability and label switching are also investigated.

In the more general area of the SEM, the conjugate choice is also the typical one (Scheines et al., 1999;Lee and Song, 2003; Dunson et al., 2005 and Mutheén and Asparouhov (2012) among others), avoiding however high variance priors. The interested reader may also refer to a fairly recent book by Lee (2007) and the references within. In LVMs with categorical latent variables, Bayesian approaches can be found for the latent class models. Evans et al. (1988) consists an early reading on the field were Dirichlet priors are considered for the model parameters. Weakly informative prior distributions (normal and gamma priors for the class means and precision parameters respectively) are considered in Ghosh et al. (2011).

In the literature of the IRT and the more general LTMs, there are two schools. The first implements the probit response (1-PNO, 2-PNO and 3-PNO models; for instance Mislevy, 1986) and the second implements the logistic one (1-PL, 2-PL and 3-PL models, for instance Patz and Junker, 1999a,b). For the first approach a conjugate choice exists that facilitates the implementation. In particular, normal priors are used for the difficulty parameters, truncated normal priors for the discrimination parameters (restricted to be positive) and finally beta priors are implemented for the guessing parameter (see for instance Sahu, 2002). Similar priors are used in models with a multilevel structure either on the ability parameters (Fox and Glas, 2001) or on the item parameters (Janssen et al., 2000), on person fit analysis IRT models (Glas and Meijer, 2003), on multidimensional IRT models (Beguin and Glas, 2001) and on hierarchical multidimensional IRT (Sheng, 2008). Beguin and Glas (2001) also examine the effects of different prior distributions on parameter recovery. In the case of the logistic IRT models, there are no priors that lead to conjugate forms. Typically, $N(0, \sigma^2_{\alpha_j})$ priors are implemented for the $\boldsymbol{\alpha}$ parameter and $LN(0, \sigma^2_{\beta_j})$ for the $\boldsymbol{\beta}$ parameter with close choices for the prior variances (Kang and Cohen, 2007; Kim and Bolt, 2007; Patz and Junker, 1999a,b; Sinharay, 2005, 2006 among others). The effect of priors on the parameter estimation of the logistic IRT models is assessed in Gifford and Swaminathan (1990).

In the next section the MCMC techniques that are employed in order to sample from the posterior are briefly reviewed.

1.4.1.2 Sampling from the posterior

Any Bayesian inference is based solely on the posterior distribution of the model parameters. The posterior however is rarely known in closed form, as for instance in the *conjugate* case. The term was first appeared in Raïffa and Schlaifer (1961) to indicate that the prior and the posterior belong to the same family (for instance when they are both Gaussian) and the posterior is available analytically. Most often we refer to the conjugate priors, that is, priors that lead to closed form posteriors. In the majority of the cases however, the posterior is known up to a constant, that is, $f(\boldsymbol{\theta}|\, \mathbf{Y}) = f(\mathbf{Y}|\, \boldsymbol{\theta})\pi(\boldsymbol{\theta})/z \propto f(\mathbf{Y}|\, \boldsymbol{\theta})\pi(\boldsymbol{\theta})$. In order to obtain posterior samples, Markov chain Monte Carlo (MCMC) are nowadays employed.

The MCMC date back to the onset of computers and the literature on the topic is extensive (see for instance Carlin and Chib, 1995, Brooks, 1998 and references within). The key idea is attributed to Metropolis et al. (1953) who proposed to construct an aperiodic and irreducible Markov chain, whose stationary (or invariant) distribution is the distribution of interest (target distribution). If the chain is run for sufficiently long time, then the derived simulated values belong to the target distribution (see Meyn and Tweedie, 2009 for details on regularity conditions).

A number of MCMC samplers have been developed to account for more or less general problems. In a sense, they can all be considered as special cases or extensions of the initial Metropolis algorithm, briefly reviewed here. In principle we assume that the process starts at a value $\boldsymbol{\theta}$ and we wish to sample values in a way that eventually will lead us to the posterior distribution. According to the Metropolis algorithm, a candidate value $\boldsymbol{\theta}'$ is sampled from a symmetrical proposal density q, such as $q(\boldsymbol{\theta}'|\, \boldsymbol{\theta}){=}q(\boldsymbol{\theta}|\, \boldsymbol{\theta}')$ and is accepted with probability

$$a(\boldsymbol{\theta}|\, \boldsymbol{\theta}) = \left\{1, \frac{f(\boldsymbol{\theta}'|\, \mathbf{Y})}{f(\boldsymbol{\theta}|\, \mathbf{Y})}\right\} = \left\{1, \frac{f(\mathbf{Y}|\, \boldsymbol{\theta}')\, \pi(\boldsymbol{\theta}')}{f(\mathbf{Y}|\, \boldsymbol{\theta})\, \pi(\boldsymbol{\theta})}\right\}.$$

The *acceptance probability* does not include the unknown normalizing constant z, since it cancels out. It is therefore possible to sample from the posterior, even though it is not analytically available. A special case of the Metropolis algorithm is the random walk

Metropolis, where the proposal is of the form $q(\boldsymbol{\theta}|\,\boldsymbol{\theta}') = r(\boldsymbol{\theta} - \boldsymbol{\theta}')$, for some arbitrary density r.

The Metropolis algorithm became more widely known with the work of Hastings (1970), who generalised the algorithm by allowing for non symmetrical proposal densities (for a detailed review see Chib and Greenberg, 1995). This generally applicable method is the Metropolis-Hastings algorithm (MH) and the corresponding probability of acceptance is now given by

$$a(\boldsymbol{\theta}, \boldsymbol{\theta}') = \left\{1, \frac{f(\mathbf{Y}|\,\boldsymbol{\theta}')\,\pi(\boldsymbol{\theta}')\,q(\boldsymbol{\theta}', \boldsymbol{\theta})}{f(\mathbf{Y}|\,\boldsymbol{\theta})\,\pi(\boldsymbol{\theta})\,q(\boldsymbol{\theta}, \boldsymbol{\theta}')}\right\}.$$

In other applications, the posterior distribution is not available but we can sample from the conditional posterior distribution of each parameter component $\boldsymbol{\theta}_j$ given the rest of the parameters $\boldsymbol{\theta}_{\backslash j}$. The conditional densities $f(\boldsymbol{\theta}_j|\,\mathbf{Y}, \boldsymbol{\theta}_{\backslash j})$, are often called the *full conditionals* and this case is referred to as the *conditional conjugate case*, to denote that, under certain priors, the full conditionals are analytically available. The sampling algorithm is the celebrated Gibbs sampler which was popularised by Geman and Geman (1984). In Gibbs sampling, the candidate points are sampled from the full conditionals instead of using a proposal density. In a sense, the Gibbs sampler is a special case of the MH, where the acceptance probability is always 1 (see also Casella and George, 1992).

Finally, another important special case is when the full conditionals are known only up to a constant. In that case, the sampler is often called the Metropolis-within-Gibbs algorithm (MG). Here, the candidate values for each $\boldsymbol{\theta}_j$, are sampled from a proposal density (symmetrical or not) and the acceptance probability is given by

$$a(\boldsymbol{\theta}_j, \boldsymbol{\theta}_j') = \min\left\{1, \frac{f(\mathbf{Y}_j|\,\boldsymbol{\theta}_j')\,\pi(\boldsymbol{\theta}_j')\,q(\boldsymbol{\theta}_j', \boldsymbol{\theta}_j)}{f(\mathbf{Y}_j|\,\boldsymbol{\theta}_j)\,\pi(\boldsymbol{\theta}_j)\,q(\boldsymbol{\theta}_j, \boldsymbol{\theta}_j')}\right\}$$

To summarize, the posterior sample is directly available in the conjugate case. If the posterior is not available but the full conditionals are, the Gibbs algorithm is employed. If in turn the full conditionals are available only up to a constant, the Metropolis-within-Gibbs algorithm is used to sample from the posterior. In any other case, the more general Metropolis-Hastings algorithm can be used, which in the special case of symmetrical proposals coincides with the Metropolis MCMC algorithm.

Within the GLLVM, the conditional conjugate case is the most frequent, as already noted in Section 1.4.1.1. One exception is in the case of the logistic (1-PL, 2-PL and 3-PL) models. In that case, the Metropolis-within-Gibbs algorithm is used and it will be described in detail in the following chapter.

1.4.1.3 Bayesian model assessment

Bayesian model assessment can be conducted in a variety of ways. For instance, one may implement *measures of surprise* which quantify the degree of disagreement between the data and a hypothesised model, without specifying alternative models; for details see Bayarri and Berger (1997) and references within. Such measures of surprise are the traditional p-values, which within the Bayesian paradigm can be formed using several modifications, namely prior predictive p-values, posterior predictive p-values, conditional predictive p-values and partial posterior predictive p-values, all discussed in detail in Bayarri and Berger (1999). With regard to prediction within the Bayesian framework, future observables are considered either by using the *prior predictive distribution*

$$f(\mathbf{Y}) = \int f(\mathbf{Y}|\,\boldsymbol{\theta})\,\pi(\boldsymbol{\theta})\,d\boldsymbol{\theta}, \tag{1.9}$$

which is the likelihood averaged over the prior, or via the *posterior predictive distribution* (or marginal likelihood)

$$f(\mathbf{Y}'|\,\mathbf{Y}) = \int f(\mathbf{Y}'|\,\boldsymbol{\theta})\,f(\boldsymbol{\theta}|\,\mathbf{Y})\,d\boldsymbol{\theta}, \tag{1.10}$$

which is the likelihood of the future data, averaged over the posterior distribution (see for details Ntzoufras, 2011, chapters 10 and 11). In the LTM literature, model assessment is mostly conducted using the posterior predictive checks. Hoijtink and Molenaar (1997) and Hoijtink (1998) focus on the posterior predictive checks for LCM. Bayesian model selection for LCM is additionally discussed in van Onna (2002), Dean and Raftery (2010) and Ghosh et al. (2011). Sinharay (2005) and Sinharay et al. (2006) provide good illustrations of ways in which posterior predictive checks can be used with latent variable models for categorical responses (item response models). Glas and Meijer (2003) also implement posterior predictive checks to Person Fit Analysis and Beguin and Glas (2001) to estimate the three-parameter normal ogive model. Stone (2000) describes a fit statistic that closely approximates a scaled chi-squared random variable.Kang and Cohen (2007) compared model selection results, using the likelihood ratio test and information criteria.

Bayesian model comparison via the prior predictive distribution (1.9) is implemented for instance in Hoijtink (2001). Bolt and Lall (2003) also use (1.9) to compare compensatory versus non-compensatory modeling of the data. Sahu (2002) further proposes a predictive approach and compares it with another decision theoretic method which minimizes an expected loss function on the predictive space. Finally, Lopes and West (2004) looked at the problem of estimating the number of factors in the classical factor analysis model using reversible jump MCMC algorithms (Green, 1995). A recent review of the techniques used so far within LVM context can be also found in Kim and Bolt (2007).

In this thesis, the prior predictive distribution is implemented in order to compute the Bayes Factor (BF) which was initially mentioned in Jeffreys (1961) and popularised by Kass and Raftery (1995). The BF is defined as the ratio of the posterior odds of two competing models (say m_1 and m_2) multiplied by their corresponding prior odds, that is

$$BF_{10} = \frac{f(\mathbf{Y}|m_1)}{f(\mathbf{Y}|m_0)}\frac{\pi(m_1)}{\pi(m_0)} = \frac{f(m_1|\mathbf{Y})}{f(m_0|\mathbf{Y})}. \tag{1.11}$$

Kass and Raftery (1995) 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. When both models have an equal prior likelihood, the BF is defined as the ratio of the corresponding prior predictive distributions (1.9). In this context, (1.9) is often referred to as the marginal or intergraded likelihood (integrating over all model parameters) of the data matrix $\mathbf{Y}$ under each model. The integrated likelihood is the normalizing constant of the posterior probability of each model and thus requires the computation of multiple integrals. Therefore, in order to compute the BF, the challenging step is to assess the integrated likelihood.

1.5 Discussion

The introductory chapter of this thesis intended primarily to present the origins and the basic concepts of the different kinds of latent variables models. Moreover, the unifying GLLVM approach was described in detail. In addition, the key aspects of the Bayesian analysis were outlined and the recent literature with respect to the Bayesian implementation of the LVMs was reviewed.

Chapter 2

Fully Bayesian latent trait models with binary responses

> *"He remarks that, while the individual man is an insoluble puzzle, in the aggregate he becomes a mathematical certainty. You can, for example, never foretell what any one man will do, but you can say with precision what an average number will be up to. Individuals vary, but percentages remain constant. So says the statistician"*
>
> Sherlock Holmes quotes Winwood Reade *

*The quote is from the book *The Sign of the Four* (1890) by Arthur Conan Doyle. William Winwood Reade (1838 - 1875) was a British historian, explorer, and philosopher. His two best-known books are *The Martyrdom of Man* (1872) and *The Outcast* (1875).

2.1 Introduction

The majority of the theoretical findings presented in this thesis, apply to the general family of GLLVM models. However, the simulated examples presented hereafter involve in particular multidimensional latent trait models with binary data. In this chapter, a general prior is proposed for these models (Section 2.2), which also accounts for the model identification (see Section 1.3.2 for details). In order to sample from the posterior, a Metropolis-within-Gibbs is presented in detail (Section 2.3), which extends, for the multivariate case, the algorithm presented in Patz and Junker (1999b). Finally, Bayesian model comparison is considered here using the prior predictive distribution (1.9). Details on the estimation of (1.9) are presented at the end of the chapter (Section 2.4).

2.2 Prior specification and identification

A latent trait model with N individuals, p items and k factors, entails $(N+p)\times k+p$ parameters. All parameters are considered a-priori independent leading to a prior with the general structure

$$\pi(\boldsymbol{\vartheta},\mathbf{Z})=\prod_{i=1}^{N}\prod_{\ell=1}^{k}\pi(Z_{i\ell})\times\prod_{j=1}^{p}\pi(\alpha_j)\times\prod_{j=1}^{p}\prod_{\ell=1}^{k}\pi(\beta_{j\ell}). \tag{2.1}$$

The latent variables are assumed to be a-priori distributed as independent standard normal distributions, that is

$$\pi(Z_{i\ell})=N(0,1), \tag{2.2}$$

in accordance with the arguments made in Section 1.2. With regard to the choice of the prior for the item parameters, non informative priors are employed (see Section 1.4.1.1). In particular, normal priors are assigned to the difficulty parameters ($\boldsymbol{\alpha}$) and log normal to the discrimination parameters ($\boldsymbol{\beta}$). The prior means are typically set equal to zero while there is no consensus with regards to the prior variances. Here, four criteria were considered in the construction of the item priors. In particular, the item priors should

a) be non informative, in analogy with the literature on the field,

b) impose constraints in order to achieve unique solution,

c) be suitable for Bayesian model comparison and

d) be potentially generalised to other members of the GLLVM.

Non informative priors were consider here with the additional restriction to be proper, in order to be able to compute the Bayes factor (1.11). With regard to the identification problem, the positivity constrain which is often imposed on the IRT discrimination parameters (see Section 1.4.1.1), was partially relaxed. This constrain is suitable for the univariate case, where the rotation occurs as a change of item parameters' sign. In the multivariate case considered here, in order to achieve a unique solution, the loadings matrix was 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 $\beta_{jj} > 0$. Hence, positive loadings with log normal priors were considered only for the diagonal elements. For the unconstrained discrimination parameter as well as for the difficulty parameters normal priors were considered.

The final step was to choose the prior means and variances. The former is reasonable to be zero, in order to reflect the prior ignorance. The choice of the prior variances, on the other hand, is less trivial. In particular, non informative priors are associated with large variances yet, within the Bayesian model comparison framework, large variances increase the posterior probabilities of the simpler models (see for instance Kass and Raftery, 1995; Sinharay and Stern, 2002 and references within). To address this issue, the ideas presented in Ntzoufras et al. (2000) and further explored in the context of GLMs by Fouskakis et al. (2009, equation 6) were implemented. Specifically, consider a model m that belongs to the GLM family, with k independent variables represented by $\mathbf{X} = (X_{i\ell};\ i = 1, ..., N;\ \ell = 1, ..., k)$. For the model parameter vector $\mathbf{b} = \{b_0, b_1, .., b_k\}'$ Fouskakis et al. (2009) suggest a normal prior distribution of the general form

$$\pi(\mathbf{b}) = N(0, \mathbf{\Sigma}), \tag{2.3}$$

where $\mathbf{\Sigma}= N\,[\mathcal{I}(\mathbf{b})]^{-1}$ is the prior covariance matrix, N is the total sample size and $\mathcal{I}(\mathbf{b})$ is the information matrix

$$\mathcal{I}(\mathbf{b}) = \mathbf{X}'\mathbf{W}\mathbf{X}.$$

The matrix $\mathbf{W}$ is diagonal and its form depends on the link function. In the special case of binary responses where the probability of correct response is given by p_i, the matrix $\mathbf{W}$ takes the form $\mathbf{W} = \text{diag}\{p_i(1 - p_i)\}$ (McCullagh and Nelder, 1989). In that case, (2.3) coincides with the unit information prior introduced by Kass and Wasserman (1996), which corresponds to adding one data point to the data. Fouskakis et al. (2009) state that in the absence of prior information, the probability of correct response can be set a-priori equal to 1/2. With this choice (2.3) becomes

$$\pi(\mathbf{b}) = N(0, 4N[\mathbf{X}'\mathbf{X}]^{-1}). \tag{2.4}$$

For a generalized linear latent trait model (GLLTM) with binary data, $\mathbf{X}_\ell = \mathbf{Z}_\ell = \{Z_{1\ell}, ..., Z_{N\ell}\}'$, $\ell = 1, ..., k$ and

$$\mathbf{Z}'\,\mathbf{Z} = \begin{pmatrix} \mathbf{Z}_1' \\ \mathbf{Z}_2' \\ \vdots \\ \mathbf{Z}_k' \end{pmatrix} \begin{pmatrix} \mathbf{Z}_1 & \mathbf{Z}_2 & \cdots & \mathbf{Z}_k \end{pmatrix} = \begin{pmatrix} \mathbf{Z}_1'\mathbf{Z}_1 & \cdots & \mathbf{Z}_1'\mathbf{Z}_k \\ \vdots & \ddots & \vdots \\ \mathbf{Z}_k'\mathbf{Z}_1 & \cdots & \mathbf{Z}_k'\mathbf{Z}_k \end{pmatrix} = \mathrm{N}\; \mathbf{I}_{(k\times k)},$$

where $\mathbf{I}_{(k\times k)}$ is the identity matrix, which occurs due to the fact that the latent vectors are considered a-priori independent. Hence, according to (2.4), the prior variance for the unconstrained discrimination parameters equals 4. For the positive discrimination parameters, the prior variance of the log-transformed loadings was set equal to one which corresponds to a variance close to 4 for the initial parameters while it is in accordance with the corresponding literature (for instance see Patz and Junker, 1999b and references within).

The prior for the discrimination parameters is summarised as follows:

$$\pi(\beta_{j\ell}) = \begin{cases} \quad 1 & \text{if} \;\; j < \ell \text{ (constrained elements } \beta_{j\,\ell} = 0) \\ LN(0,1) & \text{if} \;\; j = \ell \\ N(0,4) & \text{if} \;\; j > \ell \end{cases} \tag{2.5}$$

where $X \sim LN(\mu, \sigma^2)$ is the log-normal distribution with the mean and the variance of $\log X$ being equal to μ and σ^2, respectively. For diagonal elements β_{jj}, 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. Similarly, the prior for the difficulty parameters is as follows

$$\pi(\alpha_j) = N(0,4), \;\; j = 1, ..., p. \tag{2.6}$$

The priors (2.2), (2.5) and (2.6) are used in (2.1) and this completes the prior specification and identification of the model.

2.3 Metropolis within Gibbs algorithm

In this thesis, the parameter estimation is implemented following the work of Patz and Junker (1999b). In particular, a Metropolis-within-Gibbs algorithm with stationary distribution $f(\boldsymbol{\vartheta}, \mathbf{Z}|\, \mathbf{Y})$ is constructed. Before presenting the algorithm in detail, we may first outline some points of interest.

In order to achieve an efficient algorithm that converges quickly, the model parameters can be grouped in blocks. This strategy is typical in high dimensional problems and minimizes the computational time required (for details see Chib and Greenberg, 1995).

In the construction of blocks, the general rule is to group together parameters that are expected to be a posteriori dependent. Hence, Patz and Junker (1999b) create one block for each item and one for each individual. For a k-factor model, the parameter components that are updated (accepted or rejected) simultaneously are the p components $\boldsymbol{\vartheta}_j = \{\alpha_j, \boldsymbol{\beta}_j\} = \{\alpha_j, \beta_{j1}, ..., \beta_{jk}\}$ and the N components $\mathbf{Z}_i = \{Z_{i1}, ..., Z_{ik}\}$.

The second important aspect of the algorithm is the choice of the proposal density. Patz and Junker (1999b) generate future (candidate) points from distributions centered at the current state. In particular, they use normal proposal distributions for the latent variables:

$$q(\mathbf{Z}_i'|\, \mathbf{Z}_i) = \prod_{\ell=1}^{k} q(Z_{i\ell}'|\, Z_{i\ell}), \text{ where } q(Z_{i\ell}'|\, Z_{i\ell}) = N(\, Z_{i\ell}, c_{\mathbf{Z}}^2), \; i = 1, ..., N, \tag{2.7}$$

and for the item difficulties:

$$q(\alpha_j'|\, \alpha_j) = N(\alpha_j, c_{\boldsymbol{\vartheta}}^2), \; j = 1, ..., p. \tag{2.8}$$

Log-normal proposal distributions are used finally for the item discriminations in random walk steps. In the multivariate case considered here ($k > 1$), log-normal proposals are assigned to the diagonal elements of the loadings matrix and normal proposal distributions for the unconstrained elements, that is:

$$q(\boldsymbol{\beta}_j'|\, \boldsymbol{\beta}_j) = \prod_{\ell=1}^{k} q(\beta_{j\ell}'|\, \beta_{j\ell}), \text{ where } q(\beta_{j\ell}'|\, \beta_{j\ell}) = \begin{cases} LN(\log \beta_{j\ell}, c_{\boldsymbol{\vartheta}}^2) & \text{if } \; j = \ell, \\ N(\log \beta_{j\ell}, c_{\boldsymbol{\vartheta}}^2) & \text{if } \; j > \ell. \end{cases} \tag{2.9}$$

The variance of the proposal density is often called the *tuning parameter* since it affects the acceptance rate of the MCMC algorithm. Gelman et al. (1996) recommend acceptance rates of about 50% for univariate parameter draws and of about 25% for higher dimensional blocks. The variances $c_{\mathbf{Z}}^2$ and $c_{\boldsymbol{\vartheta}}^2$ were therefore properly tuned at each example presented in this thesis, in order to achieve these rates.

Finally, one convenient aspect of the GLLTMs is that due to the prior and local independence assumptions, the acceptance probabilities are simplified in a direct way. In the case of the item parameters, the acceptance probability is given by

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

where $q(\boldsymbol{\vartheta}_j', \boldsymbol{\vartheta}_j) = q(\alpha_j'|\, \alpha_j) q(\boldsymbol{\beta}_j'|\, \boldsymbol{\beta}_j)$. In addition, the normal densities are symmetrical and the proposal term cancels out for the unconstrained item parameters. In the case of the diagonal loadings, where the log normal distribution was employed, the proposals are substituted by a factor $\beta_{\ell\ell}/\beta'_{\ell\ell}$. Therefore, the acceptance probability simplifies to

$$a\big(\boldsymbol{\vartheta}_j, \boldsymbol{\vartheta}'_j \,|\, \mathbf{Y}, \mathbf{Z}\big) = \min\left\{1, \frac{\beta_{\ell\ell}\,\pi(\boldsymbol{\vartheta}'_j) \prod_{i=1}^{N} f(y_{ij}|\,\boldsymbol{\vartheta}_j, \mathbf{Z}_i)}{\beta'_{\ell\ell}\,\pi(\boldsymbol{\vartheta}_j) \prod_{i=1}^{N} f(y_{ij}|\,\boldsymbol{\vartheta}_j, \mathbf{Z}_i)}\right\}. \tag{2.10}$$

With similar arguments, the acceptance probability for each $\mathbf{Z}'_i$ is given by

$$a\big(\mathbf{Z}_i, \mathbf{Z}'_i \,|\, \mathbf{Y}, \boldsymbol{\vartheta}\big) = \min\left\{1, \frac{\pi(\mathbf{Z}'_i) \prod_{j=1}^{p} f(y_{ij}|\mathbf{Z}'_i, \boldsymbol{\vartheta}_j)}{\pi(\mathbf{Z}_i) \prod_{j=1}^{p} f(y_{ij}|\mathbf{Z}_i, \boldsymbol{\vartheta}_j)}\right\}. \tag{2.11}$$

Based on the remarks made in this section, the MCMC algorithm is summarised as follows:

1. For $j = 1, \ldots, p$, attempt to sample $\boldsymbol{\vartheta}'_j$ from $f(\boldsymbol{\vartheta}_j|\,\mathbf{Y}, \mathbf{Z})$.
   (a) When $\boldsymbol{\vartheta}_j = \{\alpha_j, \boldsymbol{\beta}_j\}$ is the current parameter value, propose each component for $\boldsymbol{\vartheta}'_j$ from the proposals (2.6) and (2.5).
   (b) Accept the proposed move with probability $a\big(\boldsymbol{\vartheta}_j, \boldsymbol{\vartheta}'_j|\,\mathbf{Y}, \mathbf{Z}\big)$, according to (2.10).
2. For $i = 1, \ldots, N$, attempt to sample $\mathbf{Z}'_i$ from $f(\mathbf{Z}_i|\,\mathbf{Y}, \boldsymbol{\vartheta})$.
   (a) When $\mathbf{Z}_i$ is the current parameter value, propose $\mathbf{Z}'_i$ from the proposal (2.7).
   (b) Accept the proposed move with probability $a\big(\mathbf{Z}_i, \mathbf{Z}'_i|\,\mathbf{Y}, \boldsymbol{\vartheta}\big)$, according to (2.11).

2.4 Estimating the Bayesian Marginal likelihood

Within the Bayesian framework, model comparisons via the Bayes factor, posterior model probabilities and odds (Kass and Raftery, 1995) require the computation of the integrated (marginal) likelihood

$$f(\mathbf{Y}|\,m) = \int f(\mathbf{Y}|\,\boldsymbol{\theta}, m)\,\pi(\boldsymbol{\theta}|\,m)\,d\boldsymbol{\theta}, \tag{2.12}$$

where $\mathbf{Y}$ denotes the observed data, m stands for the hypothesized model and $\pi(\boldsymbol{\theta}|\,m)$ is the prior density of the model specific parameter vector $\boldsymbol{\theta}$ (m will be dropped hereafter for simplicity). The term Bayesian marginal likelihood (BML) will be used hereafter for (2.12) in order to avoid confusion with the same term occurring in the literature of the LVMs under the classical approach, used to describe the model likelihood when the latent variables have been marginalized out and the item parameters are considered fixed.

The BML often involves high dimensional integrals making the analytic computation infeasible, as in the case of the GLLVMs. A plethora of approximation methods have been developed the last decades for the efficient estimation of the Bayesian marginal likelihood. In this work (2.12) is assessed for latent trait models with binary items by implementing some of the most often used methods. In particular, the estimators considered here can be classified in three general categories (families), namely a) point-based estimators (PBE), b) bridge sampling estimators (BSE) and c) path sampling estimators (PSE).

The PBE and BSE are presented in detail in the current section. With regard to the PSE, existing and newly derived identities are discussed in detail in Chapter 5. However, the basic features of the PSE are presented briefly below as well, for completeness.

2.4.1 Point based estimators

The point-based estimators are derived via the *candidate's identity* (Besag, 1989)

$$f(\mathbf{Y}) = \frac{f(\mathbf{Y}|\,\boldsymbol{\theta})\,\pi(\boldsymbol{\theta})}{\pi(\boldsymbol{\theta}|\,\mathbf{Y})} \tag{2.13}$$

Based on (2.13), the problem of estimating the BML is relocated at the assessment of the posterior density. However, (2.13) holds for every point in the parameter space and therefore the posterior density can be estimated using a specific point $\boldsymbol{\theta}^*$, which justifies the term point-based estimators used herein. The corresponding identity for the BML in log scale is

$$\log f(\mathbf{Y}) = \log f(\mathbf{Y}|\,\boldsymbol{\theta}^*) + \log \pi(\boldsymbol{\theta}^*) - \log \pi(\boldsymbol{\theta}^*|\,\mathbf{Y}). \tag{2.14}$$

The different PBE differ in the methodology implemented for the assessment of $\pi(\boldsymbol{\theta}^*|\,\mathbf{Y})$. In this work, three methods are considered, namely a) the Laplace-Metropolis (LM) estimator proposed by Lewis and Raftery (1997), b) the Gaussian copula (GC) estimator proposed by Nott et al. (2008) and c) the Chib and Jeliazkov (CJ) estimator proposed by Chib and Jeliazkov (2001). The corresponding identities are described in detail below.

a) ***The Laplace-Metropolis method***

The LM method is a MCMC variant of the Laplace approximation (Tierney and Kadane, 1986). It applies to integrals of the form $I = \int e^{h(\mathbf{u})}\, d\mathbf{u}$, by using a Taylor series expansion of $h(\mathbf{u})$ round the p-dimensional vector $\mathbf{u}$ (Lewis and Raftery, 1997). In particular, the approximation is given by

$$I \approx (2\pi)^{\frac{p}{2}} |\,\mathbf{H}^*|^{1/2} e^{h(\mathbf{u}^*)}, \tag{2.15}$$

where $\mathbf{u}^*$ is the value where h attains its maximum and $\mathbf{H}^*$ is minus the inverse Hessian of h evaluated at $\mathbf{u}^*$. The Laplace approximation can be directly used to derive the BML (2.12), by substituting the unnormalised posterior in the place of h, which yields the identity

$$\log f(\mathbf{Y}) = \frac{p}{2}\log\{2\pi\} + \frac{1}{2}\log|\,\mathbf{H}^{*}| + \log\pi(\boldsymbol{\theta}^{*}) + \log f(\mathbf{Y}|\,\boldsymbol{\theta}^{*}). \tag{2.16}$$

The identity (2.16) implies that $\log\pi(\boldsymbol{\theta}^{*}|\,\mathbf{Y})$ in this case is approximated by the term $-\frac{p}{2}\log\{2\pi\} - \frac{1}{2}\log|\,\mathbf{H}^{*}|$. If $\boldsymbol{\theta}^{*}$ and $\mathbf{H}^{*}$ can be found analytically, then the approximation is straightforward. Otherwise, the Metropolis output can be used for that purpose (Raftery, 1996; Lewis and Raftery, 1997). In the latter case, reasonable choices for $\boldsymbol{\theta}^{*}$ are the argmax of the unnormalised posterior and the componentwise posterior mean or median. On the other hand, the Hessian is asymptotically equal to the posterior variance matrix, and can be estimated by the sample covariance matrix of the Metropolis output. However, Lewis and Raftery (1997) suggest to use instead a weighted variance matrix estimate with weights based on the minimum volume ellipsoid estimate of Rousseeuw and van Zomeren (1990).

The *LM* method is theoretically restricted to real valued functions h which are smooth, bounded and unimodal with a single dominant mode at $\mathbf{u}^{*}$, as in the case of the classical Laplace (Tierney and Kadane, 1986). However, Lewis and Raftery (1997) state that the method works well in practice even if these conditions are not fully satisfied.

b) ***The Gaussian copula method***

The *GC* identity is closely related to the *LM*. Nott et al. (2008) consider their method as a generalization of the Laplace approximation where $\pi(\boldsymbol{\theta}^{*}|\,\mathbf{Y})$ is now assessed by a Gaussian copula (Joe, 1997). In particular, let $\boldsymbol{\theta} = \{\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_p\}'$ be a p-dimensional vector constructed by the marginals $F_j(\boldsymbol{\theta}_j)$ $(j = 1, ..., p)$ and $\mathbf{L} \sim N_p(0, \boldsymbol{\Gamma})$, where $\boldsymbol{\Gamma}$ is a correlation matrix. It holds that $\Phi(L_j)$ is uniform and therefore $\boldsymbol{\theta}_j = F_j^{-1}\{\Phi(L_j)\}$. The corresponding p-dimensional Gaussian copula (c_{Φ}) is given by

$$c_{\Phi}\bigg(\Phi^{-1}\{F(\boldsymbol{\theta})\}\,|\,\boldsymbol{\Gamma}\bigg) = |\,\boldsymbol{\Gamma}|^{-1/2}\exp\left\{\frac{1}{2}\,\mathbf{q}(\boldsymbol{\theta})'(\mathbf{I}_p - \boldsymbol{\Gamma}^{-1})\,\mathbf{q}(\boldsymbol{\theta})\right\}, \tag{2.17}$$

according to Song (2000), where $\mathbf{I}_p$ is the identity matrix, $\mathbf{q}(\boldsymbol{\theta}) = (q_1, q_2, ..., q_p)'$ is the vector of the normal scores $q_j = \Phi^{-1}\{F_j(\boldsymbol{\theta}_j)\}$ $(j = 1, ..., p)$ and $\boldsymbol{\Gamma}$ denotes the correlation matrix of the normal scores. Nott et al. (2008) implement (2.17) to derive the *GC* identity for the BML at a point $\boldsymbol{\theta}^{*}$ as follows

$$f(\mathbf{Y}) = \frac{c_{\Phi}\bigg(\Phi^{-1}\{F(\boldsymbol{\theta}^{*})\}\,|\,\boldsymbol{\Gamma}\bigg)}{\prod\limits_{j=1}^{p} f_j(\boldsymbol{\theta}_j)}\,\pi(\boldsymbol{\theta}^{*})\,\log f(\mathbf{Y}|\,\boldsymbol{\theta}^{*}), \tag{2.18}$$

where f_j is the marginal density that corresponds to the distribution F_j. In the special case where $\boldsymbol{\theta}^{*}$ is the median, the normal scores become zero and (2.18) is simplified. Specifically, at the median the log BML becomes

$$\log f(\mathbf{Y}) = -\frac{1}{2}\log|\mathbf{\Gamma}| + \sum_{j=1}^{p}\log f_j(\theta_j^*) + \log\pi(\boldsymbol{\theta}^*) + \log f(\mathbf{Y}|\,\boldsymbol{\theta}^*). \tag{2.19}$$

The identity (2.19) is implemented in the current work, where $\boldsymbol{\theta}^*$ is the componentwise median derived by the posterior output. The identity (2.19) implies that $\log\pi(\boldsymbol{\theta}^*|\,\mathbf{Y})$ in this case is approximated by the term $\frac{1}{2}\log|\mathbf{\Gamma}| - \sum\limits_{j=1}^{p}\log f_j(\theta_j^*)$. The posterior output is implemented in order to estimate the quantities needed to approximate the posterior at $\boldsymbol{\theta}^*$. Initially, the marginals f_j are approximated by kernel density estimators based on the posterior output. With regard to the covariance matrix $\mathbf{\Gamma}$, the authors suggest the following procedure. Let us suppose that R points $\boldsymbol{\theta}^{(r)} = \{\theta_1^{(r)}, ...\theta_p^{(r)}\}$ are available from the posterior output and r_{jr} is the rank of the r-th draw for the parameter $\boldsymbol{\theta}_j$, $(j-1,...,p)$. Each variable L_j is then constructed as $L_j = \Phi^{-1}(\frac{r_{ir}-0.5}{R})$ and $\widehat{\mathbf{\Gamma}}$ is derived as their estimated correlation matrix by Rousseeuw and van Zomeren's (1990) ellipsoid method.

Density estimation using the Gaussian copulaes is sensible in the cases where we expect that the posterior distribution is close to normal. Hence, the *GC* method is more restricted than the *LM* and identical results may occur only in the case where c_Φ is a multivariate normal density.

c) ***The Chib and Jeliazkov method***

As opposed to the former two point-based estimators (*LM* and *GC*) the *CJ* method is generally applicable without imposing distributional assumptions. The method is presented by Chib and Jeliazkov (2001) and extends Chib's (1995) initial 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. Both methods employ (2.13) at $\boldsymbol{\theta}^*$. Assuming that the parameter space is divided into p blocks, the posterior at $\boldsymbol{\theta}^*$ can be decomposed to

$$\pi(\boldsymbol{\theta}^*|\,\mathbf{Y}) = \pi(\theta_1^*, \boldsymbol{\theta}_2^*, \cdots, \boldsymbol{\theta}_p^*|\,\mathbf{Y}) = \pi(\boldsymbol{\theta}_1^*|\,\mathbf{Y})\pi(\boldsymbol{\theta}_2^*|\,\mathbf{Y}, \boldsymbol{\theta}_1^*)\cdots\pi(\boldsymbol{\theta}_p^*|\,\mathbf{Y}, \boldsymbol{\theta}_1^*, \boldsymbol{\theta}_2^*, \cdots, \boldsymbol{\theta}_{p-1}^*). \tag{2.20}$$

The marginal likelihood is calculated in a straightforward manner when (2.20) is analytically available. If the full conditionals are known, Chib (1995) presented an algorithm that uses the output from the Gibbs sampler to estimate them by Rao-Blackwellization. Otherwise, Chib and Jeliazkov (2001) implement for that purpose the kernel of the M-H algorithm, $K(\cdot|\cdot)$, which denotes the transition probability of sampling $\boldsymbol{\theta}_j^*$ given that $\boldsymbol{\theta}_j$ has been already generated, given by:

$$K(\boldsymbol{\theta}_j, \boldsymbol{\theta}_j^*|\,\mathbf{Y}, \boldsymbol{\theta}_{\backslash j}) = a(\boldsymbol{\theta}_j, \boldsymbol{\theta}_j^*|\,\mathbf{Y}, \boldsymbol{\theta}_{\backslash j})\, q(\boldsymbol{\theta}_j, \boldsymbol{\theta}_j^*|\,\mathbf{Y}, \boldsymbol{\theta}_{\backslash j}), \quad j = 1, \cdots, p, \tag{2.21}$$

where $\boldsymbol{\theta}_{\backslash j}$ is the parameter vector $\boldsymbol{\theta}$ exclunding $\boldsymbol{\theta}_j$, $a(\cdot|\cdot)$ is the M-H acceptance probability and $q(\cdot|\cdot)$ is the proposal density. Employing the local reversibility condition (Chib and Jeliazkov, 2001), each of the posterior ordinates appearing in (2.20) can be written as

$$\pi(\boldsymbol{\theta}_j^*|\,\mathbf{Y},\boldsymbol{\theta}_1^*,\cdots,\boldsymbol{\theta}_{j-1}^*) = \frac{E_1\Big\{a\big(\boldsymbol{\theta}_j,\boldsymbol{\theta}_j^*|\,\mathbf{Y},\psi_{j-1}^*,\psi^{j+1}\big)\,q\big(\boldsymbol{\theta}_j,\boldsymbol{\theta}_j^*|\,\mathbf{Y},\psi_{j-1}^*,\psi^{j+1}\big)\Big\}}{E_2\Big\{a\big(\boldsymbol{\theta}_j^*,\boldsymbol{\theta}_j|\,\mathbf{Y},\psi_{j-1}^*,\psi^{j+1}\big)\Big\}}, \quad (2.22)$$

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.

Chib and Jeliazkov (2001) use (2.22) in order to estimate each posterior ordinate in (2.20). Therefore, in high dimensional models, such as the LVMs, the implementation of the CJ method becomes infeasible. In Chapter 4 it is shown that in the presence of local independence, the posterior ordinates of the p-blocks in (2.20) can be derived with a single MCMC run. The modified estimator is referred to hereafter as the independence Chib and Jeliazkov (CJ^I) estimator.

2.4.2 Bridge sampling estimators

Meng and Wong (1996) initially presented the method of bridge sampling in order to approximate ratios of normalizing constants z_1 and z_0. Specifically, let p_1 and p_0 denote two densities with supports Ω_1 and Ω_0 such as

$$p_i(\boldsymbol{\theta}) = \frac{q_i(\boldsymbol{\theta})}{z_i},\ \boldsymbol{\theta}\in\Omega_m,\ m=0,1.$$

The bridge sampling identity for the ratio z_1/z_0 requires the existence of an arbitrary function $\mathcal{A}(\cdot)$ defined on $\Omega_1\cap\Omega_0$, such as

$$0 < \Bigg|\int\limits_{\Omega_1\cap\Omega_0}\mathcal{A}(\boldsymbol{\theta})p_1(\boldsymbol{\theta})p_0(\boldsymbol{\theta})d\boldsymbol{\theta}\Bigg| < \infty. \quad (2.23)$$

Such a function exists if and only if

$$\int\limits_{\Omega_1\cap\Omega_0} p_1(\boldsymbol{\theta})p_0(\boldsymbol{\theta})d\boldsymbol{\theta} > 0, \quad (2.24)$$

implying that the common support of p_1 and p_0 is non-trivial (Meng and Wong, 1996). For any function $\mathcal{A}(\boldsymbol{\theta})$ where (2.23) and (2.24) hold, we have

$$\frac{\int_{\Omega_0}\mathcal{A}(\boldsymbol{\theta})q_1(\boldsymbol{\theta})p_0(\boldsymbol{\theta})d\boldsymbol{\theta}}{\int_{\Omega_1}\mathcal{A}(\boldsymbol{\theta})q_0(\boldsymbol{\theta})p_1(\boldsymbol{\theta})d\boldsymbol{\theta}} = \frac{z_1}{z_0}\times\frac{\int_{\Omega_1\cap\Omega_0}\mathcal{A}(\boldsymbol{\theta})p_1(\boldsymbol{\theta})p_0(\boldsymbol{\theta})d\boldsymbol{\theta}}{\int_{\Omega_1\cap\Omega_0}\mathcal{A}(\boldsymbol{\theta})p_1(\boldsymbol{\theta})p_0(\boldsymbol{\theta})d\boldsymbol{\theta}}, \quad (2.25)$$

which yields the bridge sampling identity

$$\frac{z_1}{z_0} = \frac{E_{p_0}\{q_1(\boldsymbol{\theta})\mathcal{A}(\boldsymbol{\theta})\}}{E_{p_1}\{q_0(\boldsymbol{\theta})\mathcal{A}(\boldsymbol{\theta})\}}. \tag{2.26}$$

Directly from (2.26) we may derive the general bridge sampling identity for the Bayesian marginal likelihood, as follows

$$f(\mathbf{Y}) = \frac{\int f(\mathbf{Y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})\mathcal{A}(\boldsymbol{\theta})g(\boldsymbol{\theta})d\boldsymbol{\theta}}{\int f(\boldsymbol{\theta}|\,\mathbf{Y})\mathcal{A}(\boldsymbol{\theta})g(\boldsymbol{\theta})d\boldsymbol{\theta}}, \tag{2.27}$$

where $p_1(\boldsymbol{\theta}) = f(\boldsymbol{\theta}|\,\mathbf{Y})$, $q_1(\boldsymbol{\theta}) = f(\mathbf{Y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})$, $z_1 = f(\mathbf{Y})$ and $q_0(\boldsymbol{\theta}) = g(\boldsymbol{\theta})$ and $g(\cdot)$ is a proper density ($z_0 = 1$). The general BMI identity (2.27) can be used to derive specific BML identities, for different choices of the bridge function $\mathcal{A}(\cdot)$. Many of these identities pre-existed in the literature while others are considered in Meng and Wong (1996), Gelman and Meng (1998) and Meng and Schilling (2002). The identities employed in the current thesis are shown in detail below.

a) ***The arithmetic and the harmonic mean***

The simplest identity for the BML occurs from (2.26) by taking $\mathcal{A}(\boldsymbol{\theta}) = g(\boldsymbol{\theta})$, namely

$$f(\mathbf{Y}) = \int f(\mathbf{Y}|\,\boldsymbol{\theta})\,\pi(\boldsymbol{\theta})\,d\boldsymbol{\theta}, \tag{2.28}$$

which is actually the definition (2.12). The identity (2.28) corresponds to the default (naive) BML estimator that occurs as the average of the model likelihood over parameter values drawn from the prior. Since Markov chains are not implemented in (2.28), the corresponding estimator is often referred to as the *Monte Carlo estimator* or the arithmetic mean estimator (AM). The AM was introduced in the early '90s by Raftery and Banleld (1991) and further explored by McCulloch and Rossi (1992). However, it was quickly abandoned since sophisticated MCMC methods began to appear in the literature at the same period. The main problem with the arithmetic mean is that simulated points $\boldsymbol{\theta}^{(r)}$ with large likelihood can dominate the estimator (Kass and Raftery, 1995). Additionally, the prior is overdispersed with respect to the likelihood and thus requiring millions of iterations for the estimator to converge (see, for instance, Lewis, 1994).

Several years later, Newton and Raftery (1994) presented the harmonic mean estimator (HM), which they derived using simple algebraic manipulations in (2.28). The harmonic mean identity can be considered as a member of the bridge family, since it occurs from (2.26) by taking $\mathcal{A}(\boldsymbol{\theta}) = g(\boldsymbol{\theta})/f(\mathbf{Y}|\,\boldsymbol{\theta})$, leading to the identity:

$$f(\mathbf{Y}) = \left\{\int \frac{1}{f(\mathbf{Y}|\,\boldsymbol{\theta})} f(\boldsymbol{\theta}|\,\mathbf{Y})\,d\boldsymbol{\theta}\right\}^{-1}. \tag{2.29}$$

The HM is presumably the most popular MCMC identity among practitioners in many scientific fields, due to its simplicity. However, the corresponding estimator suffers methodologically by weak points. In particular, as opposed to the AM, the HM estimator can be dominated by simulated points $\boldsymbol{\theta}^{(r)}$ with small likelihood. Similarly, it can be associated with infinite variance that in many cases can remain undetected (see Raftery et al., 2007, in discussion by Robert and Chopin). A third weak point occurs due to the fact that the identity (2.29) does not include the prior. According to Neal (Newton and Raftery, 1994, in discussion) "if the posterior is insensitive to the priors", then the HA will fail to distinguish between the different priors and it will not provide distinct estimates for the BML. Raftery et al. (2007) suggest potential solutions, among which is the *stabilized* harmonic mean which occurs by marginalizing out a subset of the parameters.

b) ***The reciprocal mean***

Gelfand and Dey (1994) independently proposed a generalised version of the HM (see also DiCiccio et al., 1997), namely the reciprocal mean estimator (RM). The RM identity may be derived from (2.26) by taking $\mathcal{A}(\boldsymbol{\theta}) = \{f(\mathbf{Y}|\,\boldsymbol{\theta})\,\pi(\boldsymbol{\theta})\}^{-1}$, that is

$$f(\mathbf{Y}) = \left\{\int \frac{g(\boldsymbol{\theta})}{f(\mathbf{Y}|\,\boldsymbol{\theta})\,\pi(\boldsymbol{\theta})} f(\boldsymbol{\theta}|\,\mathbf{Y})\,d\boldsymbol{\theta}\right\}^{-1}. \tag{2.30}$$

The RM requires to define $g(\boldsymbol{\vartheta})$ which in this context is often called the *importance* or *reference* function. 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 most cases the importance function is constructed from the posterior moments available from the MCMC output.

c) ***The bridge harmonic and geometric means***

The four bridge sampling identities for the BML that were presented so far, pre-dated the method of Meng and Wong (1996). On the contrary, the bridge harmonic estimator (BH) and the bridge geometric estimator (BG) considered in this section, are identities proposed by Meng and Wong (1996). The BH occurs by implementing in (2.26) the bridge function $\mathcal{A}(\boldsymbol{\theta}) = \{f(\mathbf{Y}|\,\boldsymbol{\theta})\,\pi(\boldsymbol{\theta})\,g(\boldsymbol{\theta})\}^{-1}$, yielding the identity

$$f(\mathbf{Y}) = \frac{\int g(\boldsymbol{\theta})^{-1} g(\boldsymbol{\theta}) d\boldsymbol{\theta}}{\int \{f(\mathbf{Y}|\,\boldsymbol{\theta})\,\pi(\boldsymbol{\theta})\}^{-1} f(\boldsymbol{\theta}|\,\mathbf{Y})\,d\boldsymbol{\theta}}. \tag{2.31}$$

For the BG, the bridge function is $\mathcal{A}(\boldsymbol{\theta}) = \{f(\mathbf{Y}|\,\boldsymbol{\theta})\,\pi(\boldsymbol{\theta})\,g(\boldsymbol{\theta})\}^{-1/2}$, leading to the identity:

$$f(\mathbf{Y}) = \frac{\int \{f(\mathbf{Y}|\,\boldsymbol{\theta})\,\pi(\boldsymbol{\theta})/g(\boldsymbol{\theta})\}^{1/2}\, g(\boldsymbol{\theta})\,d\boldsymbol{\theta}}{\int \{f(\mathbf{Y}|\,\boldsymbol{\theta})\,\pi(\boldsymbol{\theta})/g(\boldsymbol{\theta})\}^{-1/2}\, f(\boldsymbol{\theta}|\,\mathbf{Y})\,d\boldsymbol{\theta}}. \tag{2.32}$$

The two identities also require to construct an importance function. If that is feasible, then the two identities correspond to very efficient estimators.

2.4.3 Path sampling estimators

The path sampling (Gelman and Meng, 1998) method is based on the construction of a continuous and differentiable *path* $q_t(\boldsymbol{\theta}) = h(q_1, q_0, t)$ that links the unnormalised identities (2.23). The ratio of the corresponding normalizing constants $\lambda = z_1/z_0$ is estimated by implementing the thermodynamic integration (TI) identity

$$\log \lambda = \int_0^1 \int_{\boldsymbol{\theta}} \frac{d \log q_t(\boldsymbol{\theta})}{dt} p_t(\boldsymbol{\theta})\, d\boldsymbol{\theta}\, dt = \int_0^1 E_{p_t}\big\{U(\boldsymbol{\theta})\big\} dt, \tag{2.33}$$

where $U(\boldsymbol{\theta}) = \frac{d \log q_t(\boldsymbol{\theta})}{dt}$ and $E_{p_t}\big\{U(\boldsymbol{\theta})\big\}$ stands for the expectation over the sampling distribution $p_t(\boldsymbol{\theta})$. The scalar $t \in [0,1]$ is often referred to as the *temperature* parameter, since the TI has its origins in thermodynamics and specifically in the calculation of the difference in *free energy* of a system; for details see in Neal (1993, Section 6.2). A closely related approach is the stepping-stone sampling (Fan et al., 2011; Xie et al., 2011), where the ratio of interest is estimated by implementing the identity

$$\lambda = \prod_{i=0}^{n-1} \int_{\boldsymbol{\theta}} \left\{ \frac{q_1(\boldsymbol{\theta})}{q_0(\boldsymbol{\theta})} \right\}^{\Delta(t_i)} p_{\,t_i}(\boldsymbol{\theta})\, d\boldsymbol{\theta} \tag{2.34}$$

where $\Delta t = t_{i+1} - t_i$. The path and stepping-stone sampling methods have been recently implemented in the literature, in order to estimate the Bayesian marginal likelihood and the BF. In particular, for the BML the developed identities are:

a) the power posteriors method (Friel and Pettitt, 2008; Lartillot and Philippe, 2006)

$$\log f(\mathbf{Y}) = \int_0^1 E_{p_t}\Big[\log f(\mathbf{Y}|\,\boldsymbol{\theta})\Big]\, dt, \tag{2.35}$$

b) the stepping stone method Xie et al.'s (2011)

$$f(\mathbf{Y}) = \prod_{i=1}^{n} E_{p_t}\Big[f(\mathbf{Y}|\,\boldsymbol{\theta})^{\Delta t}\Big] \quad \text{and} \tag{2.36}$$

c) the generalised stepping stone method Fan et al.'s (2011)

$$f(\mathbf{Y}) = \prod_{i=1}^{n} E_{p_t}\left[\left\{ \frac{f(\mathbf{Y}|\,\boldsymbol{\theta})\, \pi(\boldsymbol{\theta})}{g(\boldsymbol{\theta})} \right\}^{\Delta t} \right]. \tag{2.37}$$

All methods are discussed in detail in Chapter 5.

2.5 Discussion

The purpose of this chapter was to describe the three steps (prior specification, sampling from the posterior and Bayesian model comparison) that are required in the Bayesian model assessment, for latent trait models with binary items. With regard to the prior information, in this chapter a non-informative prior was proposed that is suitable for Bayesian model comparison. The prior proposed here can be generalised to account for other types of LVMs, in accordance with the ideas presented in Ntzoufras et al. (2000). Regarding the posterior samples, the Metropolis algorithm proposed by Patz and Junker (1999b) was expanded to account for multivariate IRT models. Finally, an introduction to the methods of estimating the Bayesian marginal likelihood was presented.

Chapter 3

The behavior of joint and marginal Monte Carlo estimators in multi-parameter latent variable models

"Had some data ready to inspect,
I modeled the relation as a random effect,
The number of parameters just grew and grew,
I had to get some help from you-know-who!

Run run -- Markov chain run,
Programming you was fun,
but I'll be happy when you're done

I knew that my algorithm was no joke,
When my computer started spewing smoke.
My plan wasn't working so I had to sub,
I drowned my MC sorrows at the local pub."

Gareth Roberts and Jeffrey S. Rosenthal [*]

[*] Part of the lyrics of the song "An MCMC Saga", written by Roberts and Rosenthal, for the Valencia International Meeting on Bayesian Statistics 7.

3.1 Introduction

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)

$$f(\mathbf{Y}|\,\boldsymbol{\vartheta},\mathbf{Z}) = \prod_{i=1}^{N} f(\mathbf{Y}_i|\,\boldsymbol{\vartheta},\mathbf{Z}_i) \tag{3.1}$$

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

$$f(\mathbf{Y}|\,\boldsymbol{\vartheta}) = \prod_{i=1}^{N} \int f(\mathbf{Y}_i|\,\boldsymbol{\vartheta},\mathbf{Z}_i)\,\pi(\mathbf{Z}_i)\,d\mathbf{Z}_i, \tag{3.2}$$

where $\boldsymbol{\vartheta} = (\boldsymbol{\alpha},\boldsymbol{\beta})$ and $\mathbf{Y}_i = (Y_{i1}, Y_{i2}, .., Y_{ip})$. In analogy with (3.1) and (3.2), under the local independence assumption there are two equivalent formulations of the BML, namely

$$f(\mathbf{Y}) = \int \prod_{i=1}^{N} f(\mathbf{Y}_i|\,\boldsymbol{\vartheta},\mathbf{Z}_i)\,\pi(\boldsymbol{\vartheta},\mathbf{Z})\,d(\boldsymbol{\vartheta},\mathbf{Z}) \tag{3.3}$$

and

$$f(\mathbf{Y}) = \int f(\mathbf{Y}|\,\boldsymbol{\vartheta})\,\pi(\boldsymbol{\vartheta})\,d\boldsymbol{\vartheta} = \int \left[\prod_{i=1}^{N} \int f(\mathbf{Y}_i|\,\boldsymbol{\vartheta},\mathbf{Z}_i)\,\pi(\mathbf{Z}_i)\,d\mathbf{Z}_i\right] \pi(\boldsymbol{\vartheta})\,d\boldsymbol{\vartheta}\ . \tag{3.4}$$

Hereafter we refer to (3.3) with the term joint approach and to (3.4) with the term marginal approach for the BML and we compare them within the Bayesian framework. 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; Patz and Junker, 1999b).

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 of 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. In this chapter, are provided exact expressions for the variance components associated with each approach and the factors that influence the associated Monte Carlo error (MCE) are considered. In particular, it is illustrated 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, the effect of the sample covariation on the Monte Carlo estimates is outlined, which is considerably understated in the literature. In particular, for independent random variables the sample covariance is typically near but not zero. It is shown here that in high dimensions even small sample covariances reduce convergence and produce biased Monte Carlo estimates. This bias can remain undetected, due to the fact that the sample covariation causes also underestimation of the MCE.

Concerns arise also 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 Monte Carlo error (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 chapter 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 avoid being trapped in high dimensions.

The structure of this chapter is as follows. Section 3.2 presents a motivating example with regard to the estimation of the BML in a model with latent variables. Three popular Markov Chain Monte Carlo (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 3.3 reviews 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 exact MCEs under both approaches are derived in Section 3.3.1 while the factors that affect the error are considered in Section 3.3.2. 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 3.3.3, the variance reduction in the case of conditional independence is discussed. In Section 3.3.4 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.

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

Here, we focus on models with binary responses and k latent variables, 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 (1.6). For the BML formulations in (3.3) and (3.4), three BML estimators are employed namely: the reciprocal mean estimator (2.30), the bridge harmonic estimator (2.31) and the bridge geometric estimator (2.31). In order to construct the estimators using the joint approach, the parameter vector is augmented to include the latent variables, that is $\boldsymbol{\theta} = \{\boldsymbol{\vartheta}, \mathbf{Z}\} = \{\boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{Z}\}$, while for the marginal approach it holds $\boldsymbol{\theta} = \boldsymbol{\vartheta} = \{\boldsymbol{\alpha}, \boldsymbol{\beta}\}$.

The estimators require also an *importance* function $g(\boldsymbol{\theta})$. In the current example, an approximation based on the posterior moments for each parameter was implemented, with structure $g(\boldsymbol{\theta}) = g(\boldsymbol{\alpha})g(\boldsymbol{\beta}_e)$ where

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

and $\boldsymbol{\beta}_e$ refers to the non-zero components of $\boldsymbol{B}$ with elements $\log \beta_{jj}$ 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

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

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 $\pi(Z_{i\ell}|\boldsymbol{Y})$.

A simulated data set with $p = 6$ items, $N = 600$ cases and $k = 2$ factors was 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 $g(\boldsymbol{\vartheta})$. 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).

In this example, the BML (3.4) 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. 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).

3.2.1 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 3.1(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 3.1(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. That 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 3.1(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 3.1 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 3.2(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 3.1). The three joint estimators are depicted in Figure 3.2(b) and their marginal counterparts are illustrated in Figure 3.2(c).

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?

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

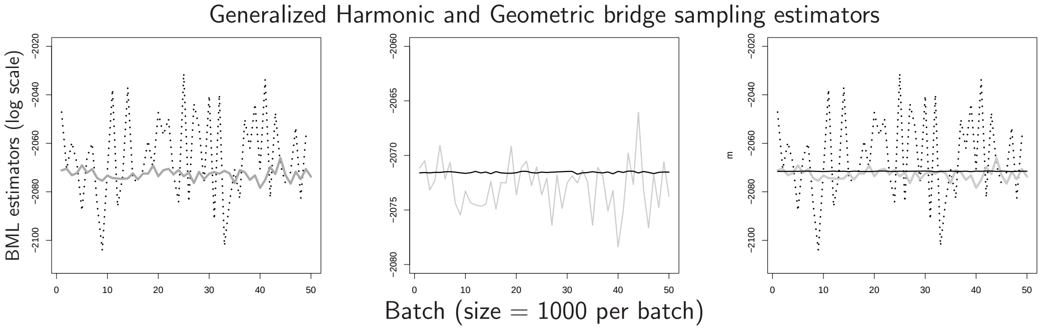
d) By increasing the number of the simulated points, do the discrepancies in the estimated values reduce? Where is this type of bias coming from?

Regarding the mechanism, we state here that is related to the model assumptions. Specifically, consider the model parameters $\boldsymbol{\vartheta}$ fixed in the BML expressions (3.3) and (3.4). It occurs that the joint expression implements the mean of the product of the independent variables $f_{\boldsymbol{\vartheta}}(\mathbf{Y}_i|\mathbf{Z}_i)$ 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.

Table 3.1: BML estimates (log scale) for the GLLTM example

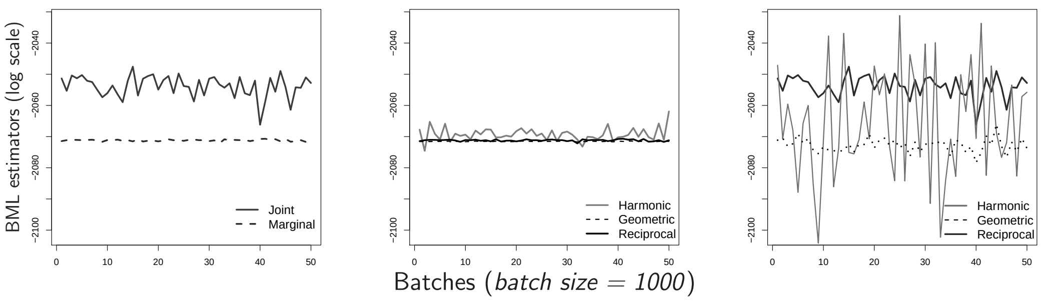
Approach	Estimator	Estimation	Batch mean	$\widehat{MCE}$
Joint	RM	-2062.3	-2053.9	3.46
	BH	-2068.8	-2065.5	17.92
	BG	-2073.3	-2072.8	2.21
Marginal	RM	-2071.3	-2071.2	0.28
	BH	-2069.6	-2069.3	2.11
	BG	-2071.6	-2071.6	0.07

The estimated BML of a GLLTM 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.



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

Figure 3.1: 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.



(a) Reciprocal estimator (b) Marginal estimators (c) Joint estimators

Figure 3.2: Joint and marginal approaches for the reciprocal (RM), generalized harmonic (BH) and 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.

3.3 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

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

The MC approximation of the integral (3.5) corresponds to the expected value of $\phi(\boldsymbol{Y})$ over $h(\boldsymbol{Y})$. Specifically, if $\mathbf{y}^R = \{y_1^{(r)}, y_2^{(r)}, ..., y_N^{(r)}\}_{r=1}^R$ is a random sample of points drawn from h, then the estimator $\widehat{I} = \overline{\phi} = \frac{1}{R}\sum_{r=1}^R \phi\big(y_1^{(r)}, y_2^{(r)}, ..., y_N^{(r)}\big)$ will approach (3.5) 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:

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

while an obvious estimator of MCE is given by $\widehat{MCE} = \widehat{\sigma}/\sqrt{R}$, provided that an estimator of the integrand's variance $\widehat{\sigma}^2$ is available. From (3.5), it occurs that the MCE directly depends on σ and R.

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

$$I = E_h[\phi(\mathbf{Y})] = E_h\left[\prod_{i=1}^{N}\phi_i(Y_i)\right] = \int \prod_{i=1}^{N}\phi_i(Y_i)h(Y_1, Y_2, ..., Y_N)\,d(Y_1, Y_2, ..., Y_N) \qquad (3.5)$$

assuming that the Y_is are independent random variables. Under this assumption, we can rewrite (3.5) as

$$I = \prod_{i=1}^{N} E_h\left[\phi_i(Y_i)\right] = \prod_{i=1}^{N} \int \phi_i(Y_i)h_i(Y_i)dY_i\ . \qquad (3.6)$$

The expressions (3.5) and (3.6) can be used to construct two unbiased Monte Carlo estimators of I, described in Definitions 3.3.1 and 3.3.2 that follow.

Definition 3.3.1 Joint estimator of I. *For any random sample* $\{y_1^{(r)}, y_2^{(r)}, ..., y_N^{(r)}\}_{r=1}^R$ *from h, the joint estimator of I is defined as*

$$\widehat{I}_J = \overline{\phi} = \frac{1}{R}\sum_{r=1}^{R}\phi\Big(y_1^{(r)}, y_2^{(r)}, ..., y_N^{(r)}\Big) = \frac{1}{R}\sum_{r=1}^{R}\left[\prod_{i=1}^{N}\phi_i\Big(y_i^{(r)}\Big)\right]\ . \qquad (3.7)$$

Definition 3.3.2 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*

$$\widehat{I}_M = \prod_{i=1}^{N} \left[\frac{1}{R} \sum_{r=1}^{R} \phi_i \left(y_i^{(r)} \right) \right] = \prod_{i=1}^{N} \overline{\phi}_i. \tag{3.8}$$

The two estimators $\widehat{I}_J$ and $\widehat{I}_M$ are asymptotically equivalent, according to the central limit theorem. In the remaining of the chapter we examine the divergencies between the two estimators in finite settings, as a result of disregarding the assumption of independence.

3.3.1 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 (3.7) is directly linked to the variance of the product of N independent variables since

$$Var(\widehat{I}_J) = Var\left[\frac{1}{R} \sum_{r=1}^{R} \Big\{ \prod_{i=1}^{N} \phi_i \big(y_i^{(r)} \big) \Big\} \right] = \frac{Var\left[\prod_{i=1}^{N} \phi_i(Y_i) \right]}{R}. \tag{3.9}$$

On the other hand, the variance of the marginal estimator (3.8) is given by the variance of the product of N univariate MC estimators, that is

$$Var(\widehat{I}_M) = Var\left[\prod_{i=1}^{N} \overline{\phi}_i \right]. \tag{3.10}$$

The difference between (3.9) and (3.10) 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

$$Var\left(\prod_{i=1}^{N} \phi_i(Y_i) \right) = \sum_{i=1}^{N} V_i \prod_{i' \neq i}^{N} E_{i'}^2 + \sum_{i_1 < i_2}^{N} V_{i_1} V_{i_2} \prod_{i' \neq i_1, i_2}^{N} E_{i'}^2 + ... + V_1 V_2 \cdots V_N, \tag{3.11}$$

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 (3.11) can be written as

$$Var\left(\prod_{i=1}^{N} \phi_i(Y_i) \right) = \sum_{k=1}^{N} \sum_{\mathcal{C} \in \binom{\mathcal{N}}{k}} \left[\prod_{i \in \mathcal{C}} V_i \prod_{j \in \mathcal{N} \setminus \mathcal{C}} E_j^2 \right], \tag{3.12}$$

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 (3.9) and (3.10) may now be expressed in terms of (3.11). Specifically, the variance of the joint estimator is directly obtained by dividing the integrand's variance in (3.11) with the simulated sample size R. For the marginal estimator, the variance (3.10) can be obtained by substituting V_i by V_i/R in (3.12). The variance components that correspond to the MCEs in each case are presented in the following lemma.

Lemma 3.3.1 *The variances of the joint (3.7) and marginal estimators (3.8) 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 3.3.1. 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.

3.3.2 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 3.3.1, the difference in their variances is given by

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

As the number of the variables increases, more positive terms are added to (3.13) 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 3.3.2 *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 (3.7) and marginal (3.8) 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

$$Var(\widehat{I}_M) = \frac{1}{R^{N_0}} \times \prod_{i\in\mathcal{N}_0} V_i \ \times \prod_{i\in\overline{\mathcal{N}}_0} E_i^2 \times \left(\prod_{i\in\overline{\mathcal{N}}_0} \left(\frac{CV_i^2}{R} + 1 \right) - I(\mathcal{N}_0 = \emptyset) \right)$$

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, $\prod\limits_{i\in\emptyset} Q_i = 1$ 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 3.3.2 is given at the Appendix. □

Based on Lemma 3.3.2, 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 3.3.2, 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 3.3.1 *The joint (3.7) and marginal (3.8) estimators achieve the same accuracy when*

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

with

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

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 3.3.1 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 3.3.2 and Corollary 3.3.1 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.

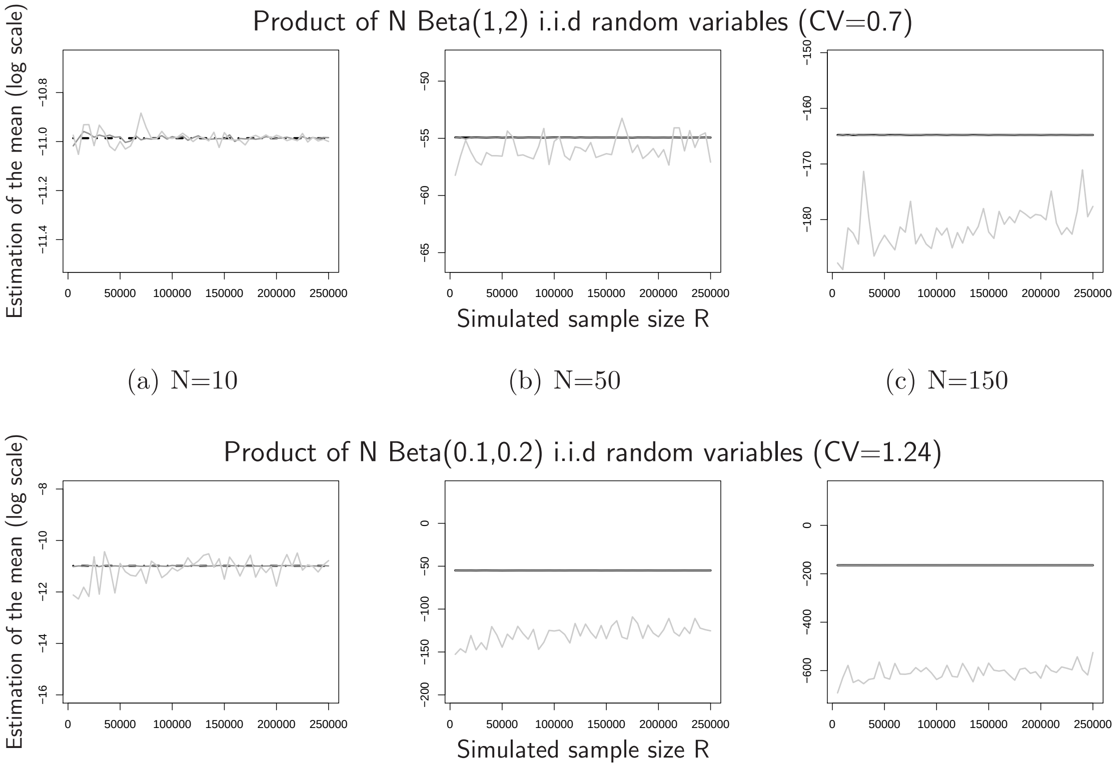
For illustration purposes, we implement a toy example of N independent and identically distributed (i.i.d) Beta random variables $Y_i \sim Beta(\lambda_1, \lambda_2)$ $(i = 1, ..., N)$. The mean of their product is given by:

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

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 3.3(a). The same procedure was repeated for $N = 50$ and $N = 150$ and is graphically represented in Figures 3.3(b) and 3.3(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 3.2.

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 3.3(d) to 3.3(e)), indicating the role of the variability of the variables involved. The estimated values and the corresponding errors are summarized in Table 3.2. Although, this example is simple assuming i.i.d random variables, the same picture can be reproduced for non identically distributed random variables.



(a) N=10

(b) N=50

(c) N=150

(d) N=10

(e) N=50

(f) N=150

Figure 3.3: The joint estimator $\widehat{I}_J$ (light grey solid line) and the marginal estimator $\widehat{I}_M$ (dark grey solid line) compared with the true mean (dashed black 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.

Table 3.2: Estimated mean of the product of i.i.d Beta variables (log scale)

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

3.3.3 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

$$\mathcal{I} = \int \Big[\prod_{i=1}^{N} \varphi_i(\boldsymbol{u}_i, \boldsymbol{v})\Big] h(\boldsymbol{u}, \boldsymbol{v})\, d(\boldsymbol{u}, \boldsymbol{v}), \tag{3.13}$$

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

$$\widehat{\mathcal{I}}_J = \frac{1}{R}\sum_{r=1}^{R}\left[\prod_{i=1}^{N}\varphi_i\big(\boldsymbol{u}_i^{(r)}, \boldsymbol{v}^{(r)}\big)\right] \tag{3.14}$$

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, (3.13) can be written as

$$\mathcal{I} = \int \left\{\prod_{i=1}^{N}\Big[\int \varphi_i(\boldsymbol{u}_i, \boldsymbol{v}) h(\boldsymbol{u}_i|\boldsymbol{v})\, d\boldsymbol{u}_j\Big]\right\} h(\boldsymbol{v})\, d\boldsymbol{v} = \int \prod_{i=1}^{N} E\big(\varphi_i\big|\boldsymbol{v}\big) h(\boldsymbol{v})\, d\boldsymbol{v}, \tag{3.15}$$

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

(3.15) we can directly obtain the corresponding marginal estimator by

$$\widehat{\mathcal{I}}_M = \frac{1}{R_1}\sum_{r_1=1}^{R_1}\Big[\prod_{i=1}^{N}\overline{\varphi}_i^{(r_1)}\Big] \ \text{ with } \ \overline{\varphi}_i^{(r_1)} = \frac{1}{R_2}\sum_{r_2=1}^{R_2}\varphi_i\big(u_i^{(r_2)}, \boldsymbol{v}^{(r_1)}\big), \tag{3.16}$$

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 3.3.3 *The variances of the joint (3.14) and marginal estimators (3.16) 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 $h(\boldsymbol{u}_i|\boldsymbol{v})$.

▷ The proof of Lemma 3.3.3 is given at the Appendix. □

Lemma 3.3.3 is an extension of Lemma 3.3.1 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 $R = R_1R_2$, 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

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

and

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

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 3.3.3 also incorporates similar expressions as in Lemma 3.3.2, 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 3.2), 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 3.1(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 nominator 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 bridge harmonic (2.31) and geometric (2.32) estimators. 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 3.3.2 we may rewrite the variances of the estimators as functions of the corresponding coefficients of variation

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

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 to 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 rational 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.

3.3.4 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 estimators themselves and how they are influenced by sample covariation which are expected to be close (but not exactly) equal to zero. These differences appear in the simulated example of Section 3.2 (see Tables 3.1 and 3.2) and cannot be attributed to the associated Monte Carlo errors of the two estimators. 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:

$$TCI(\boldsymbol{Y}) = E\Big(\prod_{i=1}^{N} Y_i\Big) - \prod_{i=1}^{N} E(Y_i), \tag{3.17}$$

which is actually the difference between the expectations under the joint and marginal approaches in their simplest forms. For instance, it coincides with the difference between the expressions in (3.5) and (3.6) if in (3.17) 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). The identity (3.17) is not useful into gaining insight on the factors that affect that difference. 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:

$$Cov_{(k)}(\boldsymbol{Y}) = Cov\Big(\prod_{i=1}^{k-1} Y_i, Y_k\Big). \tag{3.18}$$

Lemma 3.3.4 *The total covariation among N variables, is given by:*

$$TCI(\boldsymbol{Y}) = Cov_{(N)}(\boldsymbol{Y}) + \sum_{k=1}^{N-2}\left[\left(\prod_{i=N-k+1}^{N} E(Y_i)\right) Cov_{(N-k)}(\boldsymbol{Y})\right], \tag{3.19}$$

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

▷ The proof of Lemma 3.3.4 is given at the Appendix. □

The total sample covariation among the 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. In finite settings, the difference between the estimated means provided by $\widehat{I}_J$ and $\widehat{I}_M$ reflects the total sample covariation between the N variables.

When N random variables are simulated independently, even the smallest dependencies between the variables will result in non zero total sample covariation. That is, even though the N variables were sampled independently, the covariance induced by the simulation procedure cannot be ignored even for samples of several hundreds of thousands points. Therefore, if the total sample covariation is non zero, it can be considered as an index of the sample's divergence from independence. It should be noted that zero values do not ensure independence (that is, the reverse statement does not hold). By definition, the total sample covariation is accountable for and completely explains the estimation differences that were illustrated in the our examples.

Equation (3.19) implies that any divergence from the independence assumption in finite settings 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 3.3(a) to 3.3(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 for the sample covariation, according to the following lemma.

Corollary 3.3.2 *An upper bound for the absolute value of $TCI(\boldsymbol{Y})$ is given by:*

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

▷ Corollary 3.3.2 immediately follows from Lemma 3.3.4 by further implementing the Cauchy-Schwartz inequality. □

Corollary 3.3.2 provides an upper end to the total covariation therefore we cannot infer regarding the its magnitude as the various parameters increase. However, in a vise versa point of view, Lemma 3.3.2 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).

– Less variables (smaller 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 that R_0, the variance of the joint estimator is underestimated by a factor of $TCI(\boldsymbol{Y})^2$, according to the following lemma.

Lemma 3.3.5 *The variance of the product of N variables, equals their variance under assumed independence minus the square of their total covariation,*

$$Var\left(\prod_{i=1}^{N} Y_i\right) = Var\left(\prod_{i=1}^{N} Y_i \middle| Independence\right) - TCI(\boldsymbol{Y})^2 \ , \tag{3.20}$$

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

▷ The proof of Lemma 3.3.5 is given at the Appendix. □

According to Lemma 3.3.5, 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 (3.13).

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 3.2) MCMC estimators associated with low MCE lead to biased estimations and vice versa. 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 3.1). 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 nominator 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 nominator 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.

3.4 Discussion

In the presence of independence assumptions, the mean product of N variables can be either estimated by implementing the joint or the marginal approaches, as described in the current work. In finite settings the difference may be considerable, making the selection of one of the approaches crucial for the accurate estimation of specific quantities. It might seem appealing to adopt the joint approach in order to simplify the estimator and minimize the computational burden and the corresponding time required. In fact, such a gain is not obtained in practice, since the joint approach is associated with increased error and divergence from the true mean. As discussed in Section 3.3 and illustrated at the examples, the number of iterations required for the joint estimator to obtain values close to the true mean is considerably higher than the one required for the marginal estimator. In complex settings, the number of iterations might be so large, that lack of convergence may remain undetected.

Appendix

Proof of Lemma 3.3.2

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

$$Var\left(\prod_{i=1}^{N}\phi_i(Y_i)\right)=\prod_{i=1}^{N}\left(V_i+E_i^2\right)-\prod_{i=1}^{N}E_i^2. \tag{3.21}$$

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

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

which gives

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

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

The proof is completed by placing the general expression for the integrand's variance in (3.9) and (3.10) respectively. □

Proof of Lemma 3.3.3

$$
\begin{aligned}
Var(\widehat{I}_J) &= Var_{(\boldsymbol{u},\boldsymbol{v})}\left\{\frac{1}{R}\sum_{r=1}^{R}\left[\prod_{i=1}^{N}\varphi_i\big(\boldsymbol{u}_i^{(r)},\boldsymbol{v}^{(r)}\big)\right]\right\} \\
&= \frac{1}{R}Var_{(\boldsymbol{u},\boldsymbol{v})}\left[\prod_{i=1}^{N}\varphi_i\big(\boldsymbol{u}_i,\boldsymbol{v}\big)\right] \\
&= \frac{1}{R}Var_{\boldsymbol{v}}\left\{E_{\boldsymbol{u}|\boldsymbol{v}}\left[\prod_{i=1}^{N}\varphi_i\big(\boldsymbol{u}_i,\boldsymbol{v}\big)\Big|\boldsymbol{v}\right]\right\}+\frac{1}{R}E_{\boldsymbol{v}}\left\{Var_{\boldsymbol{u}|\boldsymbol{v}}\left[\prod_{i=1}^{N}\varphi_i\big(\boldsymbol{u}_i,\boldsymbol{v}\big)\Big|\boldsymbol{v}\right]\right\}
\end{aligned}
\tag{3.22}
$$

Due to conditional independence we have that

$$
E_{\boldsymbol{u}|\boldsymbol{v}}\left[\prod_{i=1}^{N}\varphi_i\big(\boldsymbol{u}_i,\boldsymbol{v}\big)\Big|\boldsymbol{v}\right] = \prod_{i=1}^{N}E_{\boldsymbol{u}|\boldsymbol{v}}\left[\varphi_i\big(\boldsymbol{u}_i,\boldsymbol{v}\big)\Big|\boldsymbol{v}\right]=\prod_{i=1}^{N}E\left(\varphi_i|\boldsymbol{v}\right). \tag{3.23}
$$

Moreover, from (3.12) we have that

$$
Var_{\boldsymbol{u}|\boldsymbol{v}}\left[\prod_{i=1}^{N}\varphi_i\big(\boldsymbol{u}_i,\boldsymbol{v}\big)\Big|\boldsymbol{v}\right] = \sum_{k=1}^{N}\sum_{\mathcal{C}\in\binom{\mathcal{N}}{k}}\left[\prod_{i\in\mathcal{C}}V\big(\varphi_i|\boldsymbol{v}\big)\prod_{j\in\mathcal{N}\setminus\mathcal{C}}E\big(\varphi_j|\boldsymbol{v}\big)^2\right] \tag{3.24}
$$

By substituting (3.23) and (3.24) in (3.22), we obtain the variance of the joint estimator of Lemma 3.3.3.

Similarly, for the marginal estimator we have

$$
\begin{aligned}
Var\left(\widehat{\mathcal{I}}_M\right) &= Var_{(\boldsymbol{u},\boldsymbol{v})}\left[\frac{1}{R_1}\sum_{r_1=1}^{R_1}\prod_{i=1}^{N}\overline{\varphi}_i^{(r_1)}\right]=\frac{1}{R_1}Var_{(\boldsymbol{u},\boldsymbol{v})}\left[\prod_{i=1}^{N}\overline{\varphi}_i\right] \\
&= \frac{1}{R_1}Var_{\boldsymbol{v}}\left\{E_{\boldsymbol{u}|\boldsymbol{v}}\Big[\prod_{i=1}^{N}\overline{\varphi}_i\Big|\boldsymbol{v}\Big]\right\}+\frac{1}{R_1}E_{\boldsymbol{v}}\left\{Var_{\boldsymbol{u}|\boldsymbol{v}}\Big[\prod_{i=1}^{N}\overline{\varphi}_i\Big|\boldsymbol{v}\Big]\right\}
\end{aligned}
\tag{3.25}
$$

Due to conditional independence we have that

$$
E_{\boldsymbol{u}|\boldsymbol{v}}\Big[\prod_{i=1}^{N}\overline{\varphi}_i\Big|\boldsymbol{v}\Big] = \prod_{i=1}^{N}E_{\boldsymbol{u}|\boldsymbol{v}}\left[\overline{\varphi}_i\Big|\boldsymbol{v}\right]=\prod_{i=1}^{N}E\left(\varphi_i|\boldsymbol{v}\right). \tag{3.26}
$$

Moreover, from Lemma 3.3.1 we have that

$$
Var_{\boldsymbol{u}|\boldsymbol{v}}\Big[\prod_{i=1}^{N}\overline{\varphi}_i\Big|\boldsymbol{v}\Big] = \sum_{k=1}^{N}\left[\frac{1}{R_2^k}\sum_{\mathcal{C}\in\binom{\mathcal{N}}{k}}\prod_{i\in\mathcal{C}}V_i\prod_{j\in\mathcal{N}\setminus\mathcal{C}}E_j^2\right], \tag{3.27}
$$

Substituting (3.26) and (3.27) in (3.25) gives the expression of the variance of the marginal estimator of Lemma 3.3.3.

Proof of Lemma 3.3.4

The proof of Lemma 3.3.4 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 3.17) for vectors $\boldsymbol{Y}$ of length equal to three.

Let us now assume that (3.19) 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

$$\mathrm{TCI}(\boldsymbol{Y}_{N+1}) = Cov_{(N+1)}(\boldsymbol{Y}) + \sum_{k=1}^{N-1}\left[\left(\prod_{i=N-k+2}^{N+1} E(Y_i)\right) Cov_{(N+1-k)}(\boldsymbol{Y})\right], \qquad (3.28)$$

is also true since

$$\begin{aligned} TCI(\boldsymbol{Y}_{N+1}) = E\left(\left[\prod_{i=1}^{N} Y_i\right] \times Y_{N+1}\right) - \left[\prod_{i=1}^{N} E(Y_i)\right] E(Y_{N+1}) & \\ = \quad Cov_{(N+1)}(\boldsymbol{Y}) + E\left(\prod_{i=1}^{N} Y_i\right) E(Y_{N+1}) - \left[\prod_{i=1}^{N} E(Y_i)\right] E(Y_{N+1}) & \\ = \quad Cov_{(N+1)}(\boldsymbol{Y}) + TCI(\boldsymbol{Y}_N)E(Y_{N+1}) & \\ = \quad Cov_{(N+1)}(\boldsymbol{Y}) + \left\{Cov_{(N)}(\boldsymbol{Y}) + \sum_{k=1}^{N-2}\left[\left(\prod_{i=N-k+1}^{N} E(Y_i)\right) Cov_{(N-k)}(\boldsymbol{Y})\right]\right\} E(Y_{N+1}) & \\ \textit{(from eq. 3.19)} & \\ = \quad Cov_{(N+1)}(\boldsymbol{Y}) + Cov_{(N)}(\boldsymbol{Y})E(Y_{N+1}) + \sum_{k=1}^{N-2}\left[\left(\prod_{i=N-k+1}^{N+1} E(Y_i)\right) Cov_{(N-k)}(\boldsymbol{Y})\right] & \\ = \quad Cov_{(N+1)}(\boldsymbol{Y}) + Cov_{(N)}(\boldsymbol{Y})E(Y_{N+1}) + \sum_{k'=2}^{N-1}\left[\left(\prod_{i=N-k'+2}^{N+1} E(Y_i)\right) Cov_{(N-k'+1)}(\boldsymbol{Y})\right] & \\ \textit{(we set } k' = k+1 \textit{)} & \\ = \quad Cov_{(N+1)}(\boldsymbol{Y}) + \sum_{k'=1}^{N-1}\left[\left(\prod_{i=N-k'+2}^{N+1} E(Y_i)\right) Cov_{(N-k'+1)}(\boldsymbol{Y})\right] & \end{aligned}$$

[for $k = 1$, the term in the summation of (3.28) is equal to $Cov_{(N)}(\boldsymbol{Y})E(Y_{N+1})$].

Proof of Lemma 3.3.5

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

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

Chapter 4

Bayesian marginal likelihood estimation using the Metropolis kernel in multi-parameter latent variable models

"Everything must be made as simple as possible. But not simpler."

Albert Einstein [*]

[*] Albert Einstein (1879 - 1955) was a theoretical physicist and humanist. Simple as that, yet not simple at all.

4.1 Introduction

One popular estimator for the Bayesian marginal likelihood is the one proposed by Chib and Jeliazkov (2001), which extends Chib's (1995) original estimator, by allowing intractable full conditional densities (denoted hereafter as the CJ estimator). The CJ estimator evaluates the posterior at a high density point $\boldsymbol{\theta}^*$, using output from sequential Metropolis-Hastings algorithms (see Section 1.4.1.2), one for each element of $\boldsymbol{\theta}$. The sequential MCMC runs, appear to be computationally demanding when the parameter space is large (see Section 2.4.1 for details). However, the method is favored by the fact that the posterior is directly obtained by the MH 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 (Newton and Raftery, 1994), or bridge family (Meng and Wong, 1996), even though very efficient, require to sample from a carefully constructed and well tuned envelope function. Quick approximation techniques, such as the Laplace Metropolis (Lewis and Raftery, 1997) or Gaussian copula (Nott et al., 2008) estimators, can be also used but they impose distributional restrictions for the posterior, such as normality or symmetry. On the contrary, Chib and Jeliazkov's (2001) approach is based on the MH kernel per ce, without any additional restrictions or assumptions.

In this chapter, the local independence assumption is employed in the construction of a multi-block MG 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 (Chib and Jeliazkov, 2001, Tanner and Wong, 1987).

The rest of the chapter is organized as follows. Section 4.2 gives a general model framework for fitting models with latent variables, where the GLLVM are derived as special case. Section 4.3 presents the CJ (Chib and Jeliazkov, 2001) estimator. Section 4.4 explains how the method can be simplified using the local independence assumption of the likelihood and compares 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 4.5.3 describes the implementation in GLLTM examples, including illustrations on simulated and real data sets. Concluding remarks are provided at the discussion section of this chapter.

4.2 Framework and model formulation

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

$$f\Big(\mathbf{Y}|\,\boldsymbol{\Theta} = (\boldsymbol{\theta}_0, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_p), \boldsymbol{L}\Big) = f\Big(\mathbf{Y}|\,\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_p), \mathbf{Z} = (\boldsymbol{\theta}_0, \boldsymbol{L})\Big), \tag{4.1}$$

where

- $\mathbf{Y} = (\mathbf{Y}_1, \ldots, \mathbf{Y}_p)$ 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 the GLLVM (Bartholomew et al., 2011). Note that in the model formulation in (4.1), the pair of parameters and the latent variables $(\boldsymbol{\Theta}, \boldsymbol{L})$ correspond to the pair $(\boldsymbol{\vartheta}, \mathbf{Z})$ with $\boldsymbol{\vartheta}$ 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 $\boldsymbol{L}$ solely refers to latent variables $\mathbf{Z}$.

4.3 The Chib and Jeliazkov marginal likelihood estimator

Both Chib's Chib (1995) and Chib and Jeliazkov Chib and Jeliazkov (2001) estimators, are based on the *candidate's identity* (Besag, 1989) presented in section 2.4.1. Following Chib (1995), let us suppose that the parameter space is divided into p blocks of parameters. Then the posterior ordinate can be decomposed to

$$f(\boldsymbol{\theta}^*|\,\mathbf{Y}) = f(\boldsymbol{\theta}_1^*, \boldsymbol{\theta}_2^*, \cdots, \boldsymbol{\theta}_p^*|\,\mathbf{Y}) = f(\boldsymbol{\theta}_1^*|\,\mathbf{Y})f(\boldsymbol{\theta}_2^*|\,\mathbf{Y}, \boldsymbol{\theta}_1^*)\cdots f(\boldsymbol{\theta}_p^*|\,\mathbf{Y}, \boldsymbol{\theta}_1^*, \boldsymbol{\theta}_2^*, \cdots, \boldsymbol{\theta}_{p-1}^*). \tag{4.2}$$

The marginal likelihood is calculated in a straightforward manner when (4.2) is analytically available. In the case when the full conditionals are known, Chib (1995) presented an algorithm that uses the output from the Gibbs sampler to estimate them by Rao-Blackwellization. In addition, Chib and Jeliazkov (2001) extended the method to deal with cases where the full conditional posterior distributions are not available and, therefore, a Metropolis–Hastings (MH) algorithm is used to generate posterior samples. The authors implement for that purpose the kernel of the MH algorithm, which denotes the transition probability of sampling $\boldsymbol{\theta}_j^*$ given that $\boldsymbol{\theta}_j$ has been already generated

$$K(\boldsymbol{\theta}_j, \boldsymbol{\theta}_j^* | \mathbf{Y}, \boldsymbol{\theta}_{\setminus j}) = a(\boldsymbol{\theta}_j, \boldsymbol{\theta}_j^* | \mathbf{Y}, \boldsymbol{\theta}_{\setminus j})\, q(\boldsymbol{\theta}_j, \boldsymbol{\theta}_j^* | \mathbf{Y}, \boldsymbol{\theta}_{\setminus j}), \quad j = 1, \cdots, p, \tag{4.3}$$

where $\boldsymbol{\theta}_{\setminus j}$ is the parameter vector $\boldsymbol{\theta}$ without $\boldsymbol{\theta}_j$, $a(\boldsymbol{\theta}_j, \boldsymbol{\theta}_j^* | \mathbf{Y}, \boldsymbol{\theta}_{\setminus j})$ is the MH acceptance probability and $q(\boldsymbol{\theta}_j, \boldsymbol{\theta}_j^* | \mathbf{Y}, \boldsymbol{\theta}_{\setminus j})$ is the proposal density. Employing the local reversibility condition, each of the posterior ordinate appearing in (4.2) can be written as

$$f(\boldsymbol{\theta}_j^* | \mathbf{Y}, \boldsymbol{\theta}_1^*, \cdots, \boldsymbol{\theta}_{j-1}^*) = \frac{E_1\Big\{a\big(\boldsymbol{\theta}_j, \boldsymbol{\theta}_j^* | \mathbf{Y}, \psi_{j-1}^*, \psi^{j+1}\big)\, q\big(\boldsymbol{\theta}_j, \boldsymbol{\theta}_j^* | \mathbf{Y}, \psi_{j-1}^*, \psi^{j+1}\big)\Big\}}{E_2\Big\{a\big(\boldsymbol{\theta}_j^*, \boldsymbol{\theta}_j | \mathbf{Y}, \psi_{j-1}^*, \psi^{j+1}\big)\Big\}}, \tag{4.4}$$

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 $f\big(\boldsymbol{\theta}_j, \psi^{j+1} | \mathbf{Y}, \psi_{j-1}^*\big)$ and $f\big(\psi^{j+1} | \mathbf{Y}, \psi_j^*\big) q\big(\boldsymbol{\theta}_j, \boldsymbol{\theta}_j^* | \psi_{j-1}^*, \psi^{j+1}\big)$ accordingly.

A Monte Carlo estimator for each ordinate can be obtained by replacing the expectations in (4.4) 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 (4.4) 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.

In particular, for the GLLVM, the posterior ordinate required to calculate the marginal likelihood includes all parameters, that is $f(\boldsymbol{\vartheta}^*, \mathbf{Z}^* | \mathbf{Y})$. Usually, the number of blocks employed for $\boldsymbol{\vartheta}^*$ 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 (2001) 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 (4.4), while the last one is calculated via a Rao-Blackwellization step as the average of $f(\boldsymbol{\vartheta}_p^* | \mathbf{Y}, \psi_{p-1}^*, \mathbf{Z})$ with respect to $f(\mathbf{Z} | \mathbf{Y}, \psi_{p-1}^*)$. This straightforward solution occurs when at least one conditional density is analytically available. The procedure is discussed in detail in Chib and Jeliazkov (2001), along with examples, as well as within the longitudinal data setting considered in Chib and Jeliazkov (2006).

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 (4.1).

4.4 Efficient estimation of the posterior ordinate in latent variable models

Within the framework given in equation (4.1), 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.

4.4.1 Models with local independence

As mentioned in Section 4.2, 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 4.4.1 *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*

$$f(\boldsymbol{Y}|\boldsymbol{\theta},\boldsymbol{L})=\prod_{j=1}^{p}f(\boldsymbol{Y}_j|\boldsymbol{\theta}_j,\boldsymbol{L}), \tag{4.5}$$

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 $\boldsymbol{L}_i$, $i\in\{1,2,...,n\}$.

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 4.4.2 which follows.

Definition 4.4.2 *For any model with likelihood given by equation (4.1), a set of parameters $\boldsymbol{\theta}$ is defined as **a-priori locally independent** if they are a-priori independent conditionally on $\boldsymbol{L}$. Therefore, the prior will satisfy the following equation*

$$\pi(\boldsymbol{\theta}|\boldsymbol{L})=\prod_{j=1}^{p}\pi(\boldsymbol{\theta}_j|\boldsymbol{L})\,. \tag{4.6}$$

Similarly we can introduce the *posterior local independence* using Definition 4.4.3.

Definition 4.4.3 *For any model with likelihood (4.1), a set of parameters $\boldsymbol{\theta}$ is defined as* ***a-posteriori locally independent*** *if they are a-posteriori independent conditionally on* $\boldsymbol{L}$*. Therefore the posterior distribution will satisfy the following equation*

$$f(\boldsymbol{\theta}|\, \boldsymbol{Y}, \boldsymbol{L}) = \prod_{j=1}^{p} f(\boldsymbol{\theta}_j|\, \boldsymbol{Y}_j, \boldsymbol{L})\,. \tag{4.7}$$

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.

4.4.1.1 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 $\mathbf{L}$ from $f(\mathbf{L}|\mathbf{Y}, \boldsymbol{\theta})$ 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 $q(\boldsymbol{\theta}_j, \boldsymbol{\theta}'_j|\mathbf{Y}, \mathbf{L})$.

   (b) Accept the proposed move with probability

$$\begin{aligned} a\big(\boldsymbol{\theta}_j, \boldsymbol{\theta}'_j|\,\mathbf{Y}, \boldsymbol{\theta}_{\setminus j}, \mathbf{L}\big) &= \min\left\{1, \frac{f(\mathbf{Y}|\boldsymbol{\theta}_{\setminus j}, \boldsymbol{\theta}'_j, \mathbf{L})\,\pi(\boldsymbol{\theta}_{\setminus j}, \boldsymbol{\theta}'_j|\mathbf{L})\,\pi(\mathbf{L})\,q(\boldsymbol{\theta}'_j, \boldsymbol{\theta}_j|\,\mathbf{Y}, \mathbf{L})}{f(\mathbf{Y}|\boldsymbol{\theta}_{\setminus j}, \boldsymbol{\theta}_j, \mathbf{L})\,\pi(\boldsymbol{\theta}_{\setminus j}, \boldsymbol{\theta}_j|\mathbf{L})\,\pi(\mathbf{L})\,q(\boldsymbol{\theta}_j, \boldsymbol{\theta}'_j|\,\mathbf{Y}, \mathbf{L})}\right\} \\ &= \min\left\{1, \frac{f(\mathbf{Y}_j|\,\boldsymbol{\theta}'_j, \mathbf{L})\,\pi(\boldsymbol{\theta}'_j|\mathbf{L})\,q(\boldsymbol{\theta}'_j, \boldsymbol{\theta}_j|\,\mathbf{Y}, \mathbf{L})}{f(\mathbf{Y}_j|\,\boldsymbol{\theta}_j, \mathbf{L})\,\pi(\boldsymbol{\theta}_j|\mathbf{L})\,q(\boldsymbol{\theta}_j, \boldsymbol{\theta}'_j|\,\mathbf{Y}, \mathbf{L})}\right\} = a(\boldsymbol{\theta}_j, \boldsymbol{\theta}'_j|\,\mathbf{Y}, \mathbf{L}), \end{aligned} \tag{4.8}$$

due to local and prior local independence defined in (4.5) and (4.6). Therefore the acceptance rate given in (4.8) depends only on the current and new (proposed) values of component $\boldsymbol{\theta}_j$ and the latent vector $\mathbf{L}$. 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 $\mathbf{L}$ 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 (4.3). Following similar arguments as in Chib and Jeliazkov (2001), we can exploit the the local reversibility condition at any point $\boldsymbol{\theta}_j^*$:

$$K(\boldsymbol{\theta}_j, \boldsymbol{\theta}_j^* | \mathbf{Y}, \mathbf{L}, \boldsymbol{\theta}_{\backslash j})\, f(\boldsymbol{\theta}_j | \mathbf{Y}, \mathbf{L}, \boldsymbol{\theta}_{\backslash j}) = K(\boldsymbol{\theta}_j^*, \boldsymbol{\theta}_j | \mathbf{Y}, \mathbf{L}, \boldsymbol{\theta}_{\backslash j})\, f(\boldsymbol{\theta}_j^* | \mathbf{Y}, \mathbf{L}, \boldsymbol{\theta}_{\backslash j}),$$

taking under consideration the posterior local independence given in equation (4.7)

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

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

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

resulting in

$$CJ_j^I = f(\boldsymbol{\theta}_j^* | \mathbf{Y}, \mathbf{L}) = \frac{\int K(\boldsymbol{\theta}_j, \boldsymbol{\theta}_j^* | \mathbf{Y}, \mathbf{L})\, f(\boldsymbol{\theta}_j | \mathbf{Y}, \mathbf{L})\, d\boldsymbol{\theta}_j}{\int K(\boldsymbol{\theta}_j^*, \boldsymbol{\theta}_j | \mathbf{Y}, \mathbf{L})\, d\boldsymbol{\theta}_j}\ , \tag{4.9}$$

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

The expression for the posterior $f(\boldsymbol{\theta} | \mathbf{Y})$ 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

$$\begin{aligned}
f(\boldsymbol{\theta}^* | \mathbf{Y}) &= \int \prod_{j=1}^{p} f(\boldsymbol{\theta}_j^* | \mathbf{Y}_j, \mathbf{L}) f(\mathbf{L} | \mathbf{Y})\, d\mathbf{L} \\
&= \int \prod_{j=1}^{p} \left[\frac{\int K(\boldsymbol{\theta}_j, \boldsymbol{\theta}_j^* | \mathbf{Y}, \mathbf{L})\, f(\boldsymbol{\theta}_j | \mathbf{Y}, \mathbf{L})\, d\boldsymbol{\theta}_j}{\int K(\boldsymbol{\theta}_j^*, \boldsymbol{\theta}_j | \mathbf{Y}, \mathbf{L})\, d\boldsymbol{\theta}_j} \right] f(\mathbf{L} | \mathbf{Y})\, d\mathbf{L} \\
&= \int \left[\frac{\prod_{j=1}^{p} K(\boldsymbol{\theta}_j, \boldsymbol{\theta}_j^* | \mathbf{Y}, \mathbf{L})}{\prod_{j=1}^{p} \int K(\boldsymbol{\theta}_j^*, \boldsymbol{\theta}_j | \mathbf{Y}, \mathbf{L})\, d\boldsymbol{\theta}_j} \right] f(\boldsymbol{\theta}, \mathbf{L} | \mathbf{Y})\, d(\boldsymbol{\theta}, \mathbf{L}) \\
&= E_{\boldsymbol{\theta}, \mathbf{L} | \mathbf{Y}} \left[\frac{\prod_{j=1}^{p} a\left(\boldsymbol{\theta}_j, \boldsymbol{\theta}_j^* | \mathbf{Y}, \mathbf{L}\right) q\left(\boldsymbol{\theta}_j, \boldsymbol{\theta}_j^* | \mathbf{Y}, \mathbf{L}\right)}{\prod_{j=1}^{p} E_{q_j} \left[a\left(\boldsymbol{\theta}_j^*, \boldsymbol{\theta}_j | \mathbf{Y}, \mathbf{L}\right) \right]} \right],
\end{aligned} \tag{4.10}$$

where $E_{\boldsymbol{\theta}, \mathbf{L} | \mathbf{Y}}$ is the posterior mean and E_{q_j} are the expectations with respect to each of the proposal densities $q(\boldsymbol{\theta}_j^*, \boldsymbol{\theta}_j | \mathbf{Y}, \mathbf{L})$. Hence, equation (4.10) can be estimated from:

$$\widehat{CJ^I} = \frac{1}{R} \sum_{r=1}^{R} \left[\frac{\prod_{j=1}^{p} a\big(\boldsymbol{\theta}_j^{(r)}, \boldsymbol{\theta}_j^* | \mathbf{Y}, \mathbf{L}^{(r)}\big)\, q\big(\boldsymbol{\theta}_j^{(r)}, \boldsymbol{\theta}_j^* | \mathbf{Y}, \mathbf{L}^{(r)}\big)}{\prod_{j=1}^{p} \left[\frac{1}{M} \sum_{m=1}^{M} a\big(\boldsymbol{\theta}_j^*, \boldsymbol{\theta}_j^{(m)} | \mathbf{Y}, \mathbf{L}^{(r)}\big) \right]} \right]. \tag{4.11}$$

The sample $\{\boldsymbol{\theta}_1^{(r)}, \boldsymbol{\theta}_2^{(r)}, \cdots, \boldsymbol{\theta}_p^{(r)}, \mathbf{L}^{(r)}\}_{r=1}^{R}$ comes from the joint posterior of $(\boldsymbol{\theta}, \mathbf{L})$ which is available from a full MCMC run. For each sampled set of latent and parameter values $\big(\boldsymbol{\theta}^{(r)}, \mathbf{L}^{(r)}\big)$, $r = 1, ..., R$, additional points $\{\boldsymbol{\theta}_j^{(m)}\}_{m=1}^{M}$ are generated from each proposal density $q(\boldsymbol{\theta}_j^*, \boldsymbol{\theta}_j | \,\mathbf{Y}, \mathbf{L}, \boldsymbol{\theta})$. These values are used to compute the expectation in the denominator of (4.10). From (4.11), 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 (4.6), 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.

4.4.1.2 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 $q(\boldsymbol{\theta}, \boldsymbol{\theta}' | \,\mathbf{Y}, \mathbf{L})$. For models with structure described by (4.1), under local and prior local independence assumptions, the acceptance probability under the one-block design is given by

$$a(\boldsymbol{\theta}, \boldsymbol{\theta}' | \,\mathbf{Y}, \mathbf{L}) = \min\left\{1, \frac{\prod_{j=1}^{p}\Big[f(\mathbf{Y}_j | \,\boldsymbol{\theta}_j', \mathbf{L})\,\pi(\boldsymbol{\theta}_j' | \mathbf{L})\Big]\,q(\boldsymbol{\theta}', \boldsymbol{\theta} | \,\mathbf{Y}, \mathbf{L})}{\prod_{j=1}^{p}\Big[f(\mathbf{Y}_j | \,\boldsymbol{\theta}_j, \mathbf{L})\,\pi(\boldsymbol{\theta}_j | \mathbf{L})\Big]\,q(\boldsymbol{\theta}, \boldsymbol{\theta}' | \,\mathbf{Y}, \mathbf{L})}\right\}. \tag{4.12}$$

Even though the properties of the local and prior local independence were also used here, the expression in (4.12) cannot be simplified further, since it requires the entire parameter vector $\boldsymbol{\theta}$, unlike the acceptance probabilities in (4.8). 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 (4.10), the expression of the posterior ordinate under the one-block design is given by

$$f(\boldsymbol{\theta}^* | \mathbf{Y}) \;=\; E_{\boldsymbol{\theta}, \mathbf{L} | \,\mathbf{Y}}\left[\frac{a\,(\boldsymbol{\theta}, \boldsymbol{\theta}^* | \,\mathbf{Y}, \mathbf{L})\,q(\boldsymbol{\theta}^*, \boldsymbol{\theta} | \,\mathbf{Y}, \mathbf{L})}{E_q\Big[a\,(\boldsymbol{\theta}^*, \boldsymbol{\theta} | \,\mathbf{Y}, \mathbf{L})\Big]}\right], \tag{4.13}$$

with draws coming from the posterior $f(\boldsymbol{\theta}, \mathbf{L} | \,\mathbf{Y})$ for the nominator and from the proposal density $q(\boldsymbol{\theta}^*, \boldsymbol{\theta} | \,\mathbf{Y}, \mathbf{L})$ for the denominator. The difference between the expressions in

(4.13) and (4.10) becomes more evident if we assume $q(\boldsymbol{\theta}', \boldsymbol{\theta}|\, \mathbf{Y}, \mathbf{L}) = \prod_{j=1}^{p} q(\boldsymbol{\theta}'_j, \boldsymbol{\theta}_j|\, \mathbf{Y}, \mathbf{L})$ that is reasonable due to the local and prior local independence. By defining the quantity A_j as

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

the acceptance probabilities involved in the posterior ordinate expressions (4.13) and (4.10) are given by $\min\left\{1, \prod\limits_{j=1}^{p} A_j\right\}$ in the case of the one-block design, and by $\prod\limits_{j=1}^{p} \min\{1,\ A_j\}$ 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 (Gilks et al., 1996, 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 (4.11) is expected to be more efficient and accurate than the alternative one-block, for the same number of iterations.

4.4.2 Models without local independence

When local independence cannot be assumed, one of the posterior ordinates in (4.2) can be exploited in order to marginalize out the latent vector $\mathbf{L}$. Chib and Jeliazkov Chib and Jeliazkov (2001) suggest to add a Rao-Blackwellization step at the end of the procedure for this purpose, provided that $f(\boldsymbol{\theta}_p^*|\, \mathbf{Y}, \mathbf{L}, \psi_{p-1}^*)$ 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 $\mathbf{L}$ from (4.4) 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:

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

The latent vector is then integrated out directly from the kernel

$$\begin{aligned} f(\boldsymbol{\theta}_p^*|\, \mathbf{Y}, \psi_{p-1}^*) &= \int \left[\frac{\int K(\boldsymbol{\theta}_p\ , \boldsymbol{\theta}_p^*|\, \mathbf{Y}, \mathbf{L}, \psi_{p-1}^*) f(\boldsymbol{\theta}_p|\, \mathbf{Y}, \mathbf{L}, \psi_{p-1}^*)\ d\boldsymbol{\theta}_p}{\int K(\boldsymbol{\theta}_p^*\ , \boldsymbol{\theta}_p|\, \mathbf{Y}, \mathbf{L}, \psi_{p-1}^*)\ d\boldsymbol{\theta}_p} \right] f(\mathbf{L}|\, \mathbf{Y}, \psi_{j-1}^*)\ d\mathbf{L} \\ &= \int \frac{K(\boldsymbol{\theta}_p\ , \boldsymbol{\theta}_p^*|\, \mathbf{Y}, \mathbf{L}, \psi_{p-1}^*)}{\int K(\boldsymbol{\theta}_p^*\ , \boldsymbol{\theta}_p|\, \mathbf{Y}, \mathbf{L}, \psi_{p-1}^*)\ d\boldsymbol{\theta}_p}\ \ f(\boldsymbol{\theta}_p\ , \mathbf{L}|\, \mathbf{Y}, \psi_{p-1}^*)\ d(\boldsymbol{\theta}_p\ , \mathbf{L}). \quad (4.14) \end{aligned}$$

The corresponding estimator of (4.14) is identical with (4.11), for $p = 1$ and conditioning upon $\{\mathbf{Y}, \mathbf{L}, \psi_{p-1}^*\}$. Naturally, the first $p-1$ ordinates in (4.2) are estimated via (4.4), while the last ordinate is used to marginalize out the latent variables. The MH 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 4.4.1.2.

4.5 Applications on GLLTM

In this section we illustrate the estimators discussed in Section 4.4 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 examples are for binary observed variables but the methodology, as already discussed, can be applied in all GLLVMs.

4.5.1 CJ^I estimator for GLLTMs

The estimate of the log-BML based on the CJ^I is given by

$$\widehat{\mathcal{L}}_{CJ^I} = \log f(\mathbf{Y}|\,\boldsymbol{\vartheta}^*) + \log f(\boldsymbol{\vartheta}^*) - \log \widehat{CJ^I}. \tag{4.15}$$

For the GLLTMs discussed here, the observed likelihood $f(\mathbf{Y}|\,\boldsymbol{\vartheta})$ was obtained by (3.4), that is, by marginalizing out the latent variables first, in accordance with the findings presented in Chapter 3. The model specification and prior identification, was conducted as presented in Chapter 2.

The Laplace-Metropolis estimator (2.15) proposed by Lewis and Raftery (1997) was used as benchmark method. The Laplace-Metropolis method was implemented on the posterior $f(\boldsymbol{\vartheta}|\,\mathbf{Y})$, therefore, the vector of the latent variables $\mathbf{Z}$ was also 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.

4.5.2 Tuning M and R

A dataset generated from a one-factor model with 4 binary items and 400 individuals ($p = 4$, $N = 400$ and $k = 1$ respectively, corresponding to 408 unknown parameters) was initially used. This rather restricted example was preferred 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, the convergence of the estimator was graphically examined by changing

a) M, that is, the number of points generated from the proposal density $q(\boldsymbol{\vartheta}, \boldsymbol{\vartheta}^* | \mathbf{Y}, \mathbf{Z})$ used for the estimation of the denominator in (4.11),

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

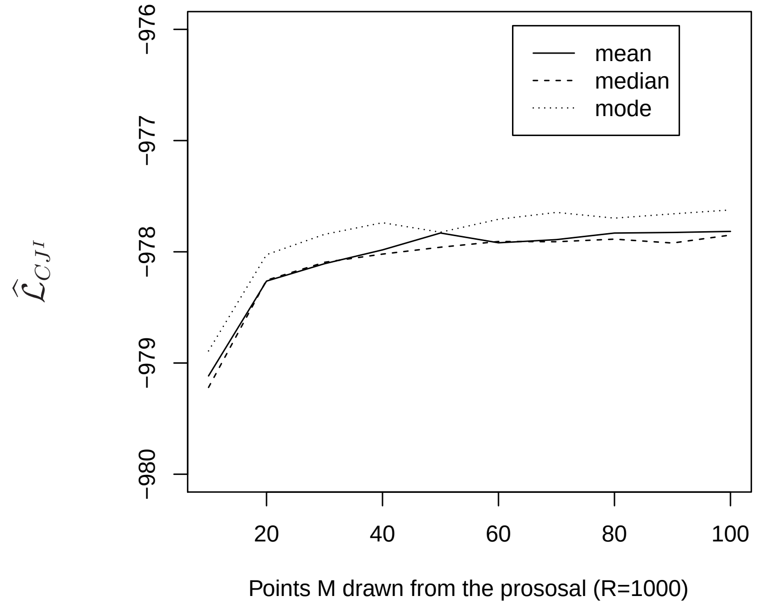
For (a), M ranged from 100 to 2000, and R was kept fixed at 1000 iterations. Figure 1(a) illustrates that all versions of $\widehat{\mathcal{L}}_{CJ^I}$ were stabilized up to a decimal point, even for $M \geq 40$. Time increased linearly, with M varying from 0.5 to 4.7 mins, which is approximately one 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 $\widehat{\mathcal{L}}_{CJ^I}$ 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.

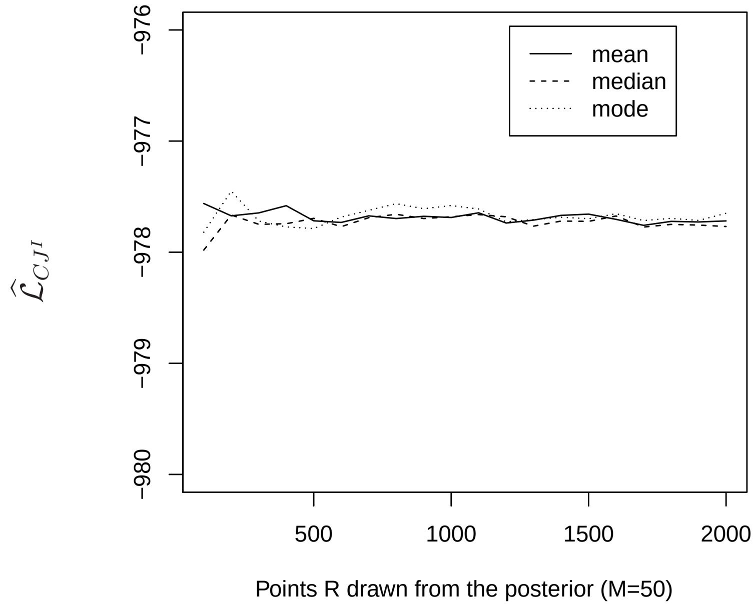
Based on Figure 1, thirty batches of size $R = 1000$ and $M = 50$ where used, to ensure convergence of the estimates. The log-BML was estimated via (4.15) at each batch. The mean over all batches, denoted by $\overline{\mathcal{L}}_{CJ^I}$, is referred to as the batch mean estimator, while the the standard deviation of the log-BML 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 $\boldsymbol{\vartheta}^*$.

Figure 2 presents the Bayesian marginal likelihood estimates based on CJ ($\widehat{\mathcal{L}}_{CJ^I}$) and LM ($\widehat{\mathcal{L}}_{LM}$) using the posterior mean, median and mode as points of central location. When using the posterior mean, $\widehat{\mathcal{L}}_{LM}$ was found equal to -977.76, while $\widehat{\mathcal{L}}_{CJ^I}$ was equal to -977.73, with estimated MCE =0.026. The estimators were quite robust, regardless of the choice of the posterior point of central location. Specifically, the $\widehat{\mathcal{L}}_{LM}$ was -977.65 at the median and -977.71 at the mode. Similarly, the $\widehat{\mathcal{L}}_{CJ^I}$ 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 examples that follow, the same tuning procedure was followed but it is not reported for brevity.

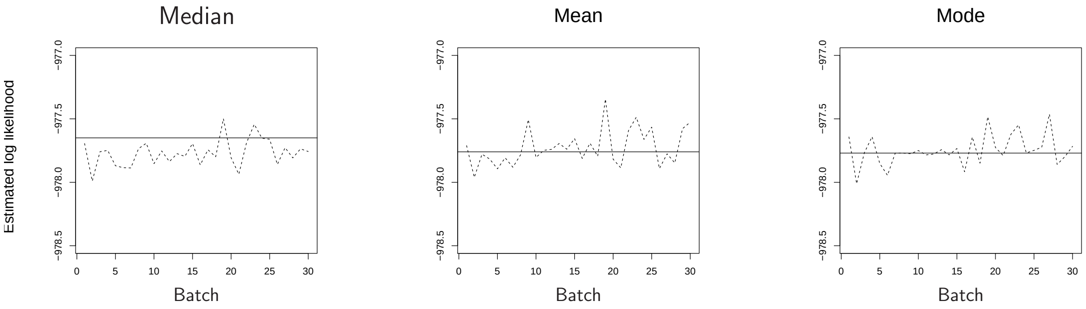


(a) Sensitivity of $\widehat{\mathcal{L}}_{CJ^I}$ on different M with $R = 1000$.



(b) Sensitivity of $\widehat{\mathcal{L}}_{CJ^I}$ on different R with $M = 50$.

Figure 4.1: Ergodic $\widehat{\mathcal{L}}_{CJ^I}$ 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.



(a) (b) (c)

Figure 4.2: The Bayesian marginal likelihood (log scale) 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.

4.5.3 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 (4.11) 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 4.5.2, 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, $\boldsymbol{\vartheta}^*$=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 $\widehat{\mathcal{L}}_{CJ^I}$ refers to the estimate of the first batch (of 1,000 iterations). The results in Table 4.1 suggest that estimates based on the independence CJ method, proposed in Section 4.4.1.1, are similar to the ones of the benchmark method, even from the first batch. Moreover, the Monte Carlo error of $\widehat{\mathcal{L}}_{CJ^I}$ 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) MH approach described in Section 4.4.1.2 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 -2200.68, with $MCE = 1.98$. In the case where two factors were assumed, the batch mean was -2066.23, with $MCE = 3.11$. In both cases the estimated log-marginal is far away from the corresponding ones reported in Table 4.1 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 4.1. 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.

Table 4.1: Simulated results: BML (log-scale) estimates

Dataset	p	N	k_{true}	k_{model}	$\widehat{\mathcal{L}}_{LM}$	$\widehat{\mathcal{L}}_{CJ^I}$	$\overline{\mathcal{L}}_{CJ^I}$	$MCE(\widehat{\mathcal{L}}_{CJ^I})$
(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; $\widehat{\mathcal{L}}_{LM}$ and $\widehat{\mathcal{L}}_{CJ^I}$: Laplace-Metropolis and Chib and Jeliazkov estimates of the marginal likelihood; $\overline{\mathcal{L}}_{CJ^I}$: Batch mean estimator of the log-marginal likelihood; $MCE(\widehat{\mathcal{L}}_{CJ^I})$: Monte Carlo error of the $\widehat{\mathcal{L}}_{CJ^I}$ obtained as the standard deviation of 30 batches of equal size.

With regards to the BF, the estimates (in log scale) reported in Table 4.2 are based on the marginal likelihood estimates presented in Table 4.1. In all three simulated datasets, the estimated Bayes factors ($\widehat{BF}$) indicated the true model. Moreover, when the indepen-

Table 4.2: Simulated results: Bayes Factor estimates

Dataset details				Comparison	$\log \widehat{BF}$		Batch summaries of $\log \widehat{BF}$			
#	p	N	k_{true}	k_1 vs. k_2	LM	CJ^I	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 $\widehat{BF}$: Summaries based on 30 batches of $\widehat{BF}$ (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).

dence 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 favor of the true generating mechanism, providing BF estimates similar to the ones obtained from the gold standard LM, in all cases.

4.5.4 Illustration on real data

We proceed with two real-data examples also analyzed in Bartholomew et al. (2008, 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 Bock and Lieberman (1970) 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

Table 4.3: Marginal Likelihood and Bayes Factor for the real data: LSAT and WIRS

	$\widehat{\mathcal{L}}_{LM}$			$\widehat{\mathcal{L}}_{CJ^I}$		
Dataset	1-factor	2-factor	$\log \widehat{BF}_{21}$	1-factor	2-factor	$\log \widehat{BF}_{21}$
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

$\widehat{\mathcal{L}}_{LM}$ and $\widehat{\mathcal{L}}_{CJ^I}$: 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; $\widehat{BF}_{21}$: Estimated Bayes factors of 2-factor versus 1-factor model.

by the computed Bayes factor (≈ 0.22 and 0.24 for the 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 4.3. In particular, the BF of the one-factor versus the two-factor model was less than 0.5 and therefore according to Kass and Raftery (1995) 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, Airey et al. (1992)). 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 4.3. As further analysis, Bartholomew et al. (2008) 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, Kass and Raftery (1995)). 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.

4.6 Discussion

This chapter focused on the CJ (Chib and Jeliazkov, 2001) 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 (Chib and Jeliazkov, 2001) 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 (Tanner and Wong, 1987; van Dyk and Meng, 2001) 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.

Chapter 5

Thermodynamic assessment of probability distribution divergencies and Bayesian model comparison

"You should call it entropy, for two reasons. In the first place your uncertainty function has been used in statistical mechanics under that name, so it already has a name. In the second place, and more important, nobody knows what entropy really is, so in a debate you will always have the advantage"

discussion between Von Neumann and Shannon [*]

[*] Claude Shannon introduced the very general concept of information entropy, used in information theory, in 1948. Initially it seems that Shannon was not particularly aware of the close similarity between his new quantity and the earlier work in thermodynamics; but the mathematician John von Neumann certainly was. The quotation first appears in: M. Tribus, E.C. McIrvine, Energy and information, Scientific American, 224, 1971. Reprint from $http://en.wikipedia.org/wiki/Talk\%3AHistory\_of\_entropy$.

5.1 Introduction

The idea of using tempered transitions has gained increased attention in Bayesian statistics as a method to improve the efficiency of the MCMC algorithms in terms of exploring the target posterior distribution. Sophisticated methods such as the Metropolis-coupled MCMC (Geyer, 1991), the simulated tempering (Marinari and Parisi, 1992; Geyer and Thompson, 1995) and the annealed sampling (Neal, 1996, 2001) incorporate transitions to overcome the slow mixing of the MCMC algorithms in multi-modal densities; see Behrens et al. (2012) for an insightful review.

Here, we focus on the ideas of path sampling (Gelman and Meng, 1994, 1998) where tempered transitions are employed in order to estimate the ratio of two intractable normalizing constants. In particular, let $q_0(\boldsymbol{\theta})$ and $q_1(\boldsymbol{\theta})$ be two unnormalized densities and z_0, z_1 be their normalizing constants leading to

$$p_t(\boldsymbol{\theta}) = \frac{q_t(\boldsymbol{\theta})}{z_t}, \quad \text{where } z_t = \int_{\boldsymbol{\theta}} q_t(\boldsymbol{\theta})\, d\boldsymbol{\theta}, \quad \text{for } t = 0, 1. \tag{5.1}$$

Gelman and Meng's (1998) method is based on the construction of a continuous and differentiable *path* $q_t(\boldsymbol{\theta}) = h(q_1, q_0, t)$ which is used to estimate the ratio of normalizing constants $\lambda = z_1/z_0$ via the *thermodynamic integration* (TI) identity

$$\log \lambda = \int_0^1 \int_{\boldsymbol{\theta}} \frac{d \log q_t(\boldsymbol{\theta})}{dt}\, p_t(\boldsymbol{\theta})\, d\boldsymbol{\theta}\, dt = \int_0^1 E_{p_t}\big\{U(\boldsymbol{\theta})\big\} dt, \tag{5.2}$$

where $U(\boldsymbol{\theta}) = \frac{d \log q_t(\boldsymbol{\theta})}{dt}$ and $E_{p_t}\big\{U(\boldsymbol{\theta})\big\}$ stands for the expectation over the sampling distribution $p_t(\boldsymbol{\theta})$. The scalar $t \in [0, 1]$ is often referred to as the *temperature* parameter, since the TI has its origins in thermodynamics and specifically in the calculation of the difference in *free energy* of a system; for details see in Neal (1993, Section 6.2). It occurs that the ideas of the thermodynamics have important applications on a variety of scientific fields, such as statistics, physics, chemistry, biology and computer science (machine learning, pattern recognition) among others. As Gelman and Meng (1998) denote, methods related to the TI have been developed by researchers from different disciplines working independently and in parallel; see, for instance, in Frenkel (1986), Binder (1986) and Ogata (1989).

A straightforward application of the path sampling refers to Bayesian model comparison. In particular, expressions for the Bayes factor (BF, Kass and Raftery, 1995) and the Bayesian marginal likelihood that employ tempered transitions have been developed by Lartillot and Philippe (2006), Friel and Pettitt (2008), Xie et al. (2011) and Fan et al. (2011). Additionally, Friel and Pettitt (2008), Calderhead and Girolami (2009), Lefebvre et al. (2010) and Behrens et al. (2012), under different motivations and scopes, outline the close relationship between the thermodynamic integration and the relative entropy, best known in statistics as the Kullback-Leibler divergence (KL; Kullback and Leibler, 1951).

All these studies, are based on specific geometric paths (Neal, 1993) of the form

$$q_t(\boldsymbol{\theta}) = q_1(\boldsymbol{\theta})^t q_0(\boldsymbol{\theta})^{1-t}, \tag{5.3}$$

for specific choices of $q_0(\boldsymbol{\theta})$ and $q_1(\boldsymbol{\theta})$. For example, Friel and Pettitt (2008) have used $q_t(\boldsymbol{\theta}) = f(\boldsymbol{y}|\boldsymbol{\theta})^t f(\boldsymbol{\theta})$ and therefore setting the unnormalized posterior as q_1 and the prior as q_0. Here, we focus on the general case of geometric paths (5.3) for any choice of q_1 and q_0. For any geometric path, (5.2) is written as

$$\log \lambda = \int_0^1 \int_{\boldsymbol{\theta}} \log \frac{q_1(\boldsymbol{\theta})}{q_0(\boldsymbol{\theta})}\, p_t(\boldsymbol{\theta})\, d\boldsymbol{\theta}\, dt. \tag{5.4}$$

since $U(\boldsymbol{\theta}) = \log q_1(\boldsymbol{\theta}) - \log q_0(\boldsymbol{\theta})$.

The identity (5.4) is implemented in order to study the connection between path sampling and entropy measures. In particular, in this thesis it is examined what happens for specific values of $t \in (0, 1)$ as well as the mechanism which eventually produces the relative entropy at the initial ($t = 1$) and at the final ($t = 0$) state, as originally discussed by Friel and Pettitt (2008) and Lefebvre et al. (2010). It is demonstrated that (5.4) can be used to compute the Chernoff information (Chernoff, 1952) as a byproduct of the path sampling procedure, which is, otherwise, a rigorous and troublesome procedure especially in multidimensional problems. Other entropy measures can be subsequently derived, such as the *Bhattacharyya distance* (Bhattacharyya, 1943) and Rényi's relative entropy (Rényi, 1961).

Based on the findings with regard to the uncertainty at the intermediated points, here the structure of the thermodynamic integration is further examined and geometrically represented . This assists us to understand the path sampling estimators in terms of error. In particular, can identify when high path-related uncertainty or large discretisation error appears and reduce it by either adopting a more efficient (in terms of error) path or tempering schedule.

Finally, attention is restricted in this chapter on the most popular implementation of TI estimation: Bayesian model evaluation. An alternative approach is considered, based on the stepping-stone identity introduced by Xie et al. (2011) and Fan et al. (2011). Existing Bayesian marginal likelihood estimators are overviewed, based on the two alternative approaches (thermodynamic and stepping-stone) by presenting recently developed TI based Bayesian marginal likelihood estimators (Friel and Pettitt, 2008; Lartillot and Philippe, 2006; Lefebvre et al., 2010) and their corresponding stepping-stone ones (Fan et al., 2011; Xie et al., 2011), based on same paths. Any blanks in the list of previously reported estimators based on the two different approaches are filled in by introducing new estimators using a identity-path selection rationality. The implementation of the two alternative approaches in the direct Bayes factor estimation is also discussed and compound paths are introduced, which can be used to efficiently switch between competing models of different dimension located at the endpoints of the path. The chapter closes with an illustration of our methods and estimators in a common regression example (previously used by Friel and Pettitt, 2008 and Lefebvre et al., 2010 for Bayesian marginal likelihood estimation) and in a latent-trait model implementation using a simulated dataset.

5.2 Entropy measures and path sampling

In Statistics, entropy is used as a measure of uncertainty which, unlike the variance, does not depend on the actual values of a random variable $\boldsymbol{\theta}$, but only on their associated probabilities. Here, we use the term *entropy measures* in a broad definition to refer to measures of divergence between probability distributions that belong to the family of f-divergencies (Ali and Silvey, 1966; Csiszár, 1963). Such measures are widely used in statistics (Liese and Vajda, 2006), information theory (Cover and Thomas, 1991) and thermodynamics (Crooks and Sivak, 2011).

The most commonly used $f-$divergence is the Kullback - Leibler (Kullback and Leibler, 1951)

$$\begin{aligned} KL(p_1 \parallel p_0) &= \int_{\boldsymbol{\theta}} p_1(\boldsymbol{\theta}) \log \frac{p_1(\boldsymbol{\theta})}{p_0(\boldsymbol{\theta})} \, d\boldsymbol{\theta} \\ &= \int_{\boldsymbol{\theta}} p_1(\boldsymbol{\theta}) \log p_1(\boldsymbol{\theta}) \, d\boldsymbol{\theta} - \int_{\boldsymbol{\theta}} p_1(\boldsymbol{\theta}) \log p_0(\boldsymbol{\theta}) \, d\boldsymbol{\theta} \\ &= -H(p_1) + cH(p_1 \parallel p_0), \end{aligned} \tag{5.5}$$

with $cH(p_1 \parallel p_0)$ being the *cross entropy* and $H(p_1)$ the *differential entropy*; see for details in Cover and Thomas (1991). The KL-divergence is always non-negative but it is not a distance or a metric with the strict mathematical definition, since neither the symmetry nor the triangle inequality conditions are satisfied. In information theory, it is mostly referred to as the *relative entropy* and is a measure of the information lost when $p_0(\boldsymbol{\theta})$ is used as an approximation of $p_1(\boldsymbol{\theta})$. Subsequently, a symmetric version of KL can naturally be defined as

$$J(p_1, p_0) = KL(p_1 \parallel p_0) + KL(p_0 \parallel p_1),$$

which dates back to Jeffreys' investigations of invariant priors (Jeffreys, 1946) and is often called as the *symmetrized KL-divergence* or *J-divergence*; see also in Lefebvre et al. (2010) for details.

The relationship between the KL-divergence and the thermodynamic integral was described by Friel and Pettitt (2008) and further studied by Lefebvre et al. (2010). In particular, the KL-divergencies between $p_1(\boldsymbol{\theta})$ and $p_0(\boldsymbol{\theta})$ can be derived by the endpoints of the expectation of $E_{p_t}\big\{U(\boldsymbol{\theta})\}$ appearing thermodynamic equation (5.4) since

$$KL(p_1 \parallel p_0) = E_{p_1}\big\{U(\boldsymbol{\theta})\big\} - \log\lambda \ \text{ and } \ KL(p_0 \parallel p_1) = -E_{p_0}\big\{U(\boldsymbol{\theta})\big\} + \log\lambda \ .$$

The findings presented by Friel et al. (2012) and Lefebvre et al. (2010) refer therefore to the endpoints of a geometric path.

The question which naturally arises here is which is the role of entropy at the intermediate points for $t \in (0,1)$. In the following, we address this issue and we illustrate how other $f-$divergencies are related to the thermodynamic integral (5.4) and how can be estimated as path sampling byproducts.

5.2.1 The normalised thermodynamic integral and $f-$divergencies

In this section, we draw attention to the normalized thermodynamic integral (NTI) given by

$$NTI = \int_0^1 \int_{\boldsymbol{\theta}} p_t(\boldsymbol{\theta}) \, \log\frac{p_1(\boldsymbol{\theta})}{p_0(\boldsymbol{\theta})} \, d\boldsymbol{\theta}\, dt. \tag{5.6}$$

The NTI is zero for any choices of p_0, p_1 and any geometric path p_t and it can be expressed via the thermodynamic integral using the identity

$$NTI = \int_0^1 \int_{\boldsymbol{\theta}} p_t(\boldsymbol{\theta}) \, \log\frac{q_1(\boldsymbol{\theta})}{q_0(\boldsymbol{\theta})} \, d\boldsymbol{\theta}\, dt - \log\lambda \ .$$

This identity will be used to link the thermodynamic integrals with $f-$divergencies at any $t \in (0,1)$, generalizing the findings of Friel et al. (2012) and Lefebvre et al. (2010) which associate the endpoints of the TI with KL divergencies. To do so, we need to rewrite (5.6) as $NTI = \int_0^1 \mathcal{KL}_t \, dt$, where $\mathcal{KL}_t$ is the *functional KL-divergence of order t* defined as

$$\mathcal{KL}_t = \int_{\boldsymbol{\theta}} p_t(\boldsymbol{\theta}) \log\frac{p_1(\boldsymbol{\theta})}{p_0(\boldsymbol{\theta})} \, d\boldsymbol{\theta} = E_{p_t}\big\{U(\boldsymbol{\theta})\big\} - \log\lambda \ . \tag{5.7}$$

Then, we can express $\mathcal{KL}_t$ as the difference between the KL divergencies of p_t with the two endpoint densities p_1 and p_0 since

$$\mathcal{KL}_t \ = \ -cH(p_t \parallel p_1) + cH(p_t \parallel p_0) = KL(p_t \parallel p_1) - KL(p_t \parallel p_0).$$

This reduces to $\mathcal{KL}_0 = -KL(p_0 \parallel p_1)$ and to $\mathcal{KL}_1 = KL(p_1 \parallel p_0)$ at the endpoints of the geometric path, which is in accordance with the findings of Friel et al. (2012) and Lefebvre et al. (2010).

The divergence $\mathcal{KL}_t$ can be interpreted as a measure of *relative location* of a density p_t relative to p_1 and p_0. Hence, for any $t \in [0,1]$, $\mathcal{KL}_t$ indicates whether p_t is closer to

p_0 (negative values) or to p_1 (positive values). The solution of the equation $\mathcal{KL}_{t^*} = 0$ defines the point t^* where p_{t^*} is equidistant (in the KL sense) from the endpoint densities. Moreover, from (5.7) it is obvious that $E_{p_{t^*}}\big\{U(\boldsymbol{\theta})\big\}$ is equal to $\log\lambda$. Therefore, in the case that t^* is known, the ratio of the normalizing constants λ can be estimated in a single MCMC run (with $t = t^*$), rather than employing the entire path using multiple simulations. However this is rarely the case and, using the inverse logic, t^* can be estimated by path sampling. Having t^* estimated, then the *Chernoff information* can be computed in straightforward manner (Parzen, 1992, Johnson and Sinanovic, 2000, Nielsen, 2011).

Following Parzen (1992), the Chernoff t-divergence (Chernoff, 1952) is given by

$$C_t(p_1 \parallel p_0) \;=\; -\log\int_{\boldsymbol{\theta}} p_1(\boldsymbol{\theta})^t p_0(\boldsymbol{\theta})^{1-t} d\boldsymbol{\theta} = -\log\mu(t), \tag{5.8}$$

where $\mu(t)$ is the *Chernoff coefficient* (Chernoff, 1952); also see Kakizawa et al. (1998) and Rauber et al. (2008). The key observation here is that when adopting geometric paths, the sampling distribution $p_t(\boldsymbol{\theta})$ embodies the Chernoff coefficient since

$$p_t(\boldsymbol{\theta}) \;=\; \frac{\big\{z_1 p_1(\boldsymbol{\theta})\big\}^t\big\{z_0 p_0(\boldsymbol{\theta})\big\}^{1-t}}{\int_{\boldsymbol{\theta}} q_1(\boldsymbol{\theta})^t q_0(\boldsymbol{\theta})^{1-t} d\boldsymbol{\theta}} = \frac{p_1(\boldsymbol{\theta})^t p_0(\boldsymbol{\theta})^{1-t}}{\mu(t)}, \tag{5.9}$$

for any $t \in [0,1]$, which is the Boltzmann-Gibbs distribution pertaining to the Hamiltonian (energy function) $\mathcal{H}_t(\boldsymbol{\theta}) = -t\log p_1(\boldsymbol{\theta}) - (1-t)\log p_0(\boldsymbol{\theta})$; see, for details, in Merhav (2010, chapter 3). In view of (5.9) the NTI becomes

$$\int_0^1\int_{\boldsymbol{\theta}} \frac{p_1(\boldsymbol{\theta})^t p_0(\boldsymbol{\theta})^{1-t}}{\mu(t)} \log\frac{p_1(\boldsymbol{\theta})}{p_0(\boldsymbol{\theta})} d\boldsymbol{\theta}\, dt = \int_0^1 \frac{d\log\mu(t)}{dt} dt = \Big[\log\mu(t)\Big]_0^1 = 0, \tag{5.10}$$

since

$$\frac{d\log\mu(t)}{dt} = \frac{1}{\mu(t)}\int \frac{d\{\, p_1(\boldsymbol{\theta})^t p_0(\boldsymbol{\theta})^{1-t}\}}{dt}\, dt.$$

From (5.10) it is straightforward to see that the NTI up to any point $t \in (0,1)$ is directly related to the Chernoff t-divergence, as described in detail in the following lemma.

Lemma 5.2.1 *The normalised thermodynamic integral (5.6) up to any point* $t \in (0,1)$ *given by*

$$NTI(t) = \int_0^t\int_{\boldsymbol{\theta}} p_t(\boldsymbol{\theta})\log\frac{p_1(\boldsymbol{\theta})}{p_0(\boldsymbol{\theta})} d\boldsymbol{\theta} \tag{5.11}$$

is equal to minus the Chernoff t*-divergence of the endpoint densities, that is*

$$NTI(t) = \log\mu(t) = -C_t(p_1 \parallel p_0). \tag{5.12}$$

The proof of Lemma 5.2.1 is obtained in straightforward manner as (5.10). □

Another interesting result can be obtained for $t = t^*$, the solution of the equation $\mathcal{KL}_t = 0$, and it is described in Lemma 5.2.2 which follows.

Lemma 5.2.2 *The Chernoff information, defined as*

$$C(p_1 \parallel p_0) = \max_{t \in [0,1]} C_t(p_1 \parallel p_0)$$

is equal to $NTI(t^*)$ *with* t^* *being the solution of equation* $\mathcal{KL}_t = 0$*, i.e.*

$$C(p_1 \parallel p_0) = NTI(t^*) \text{ with } t^* \in [0,1] : \mathcal{KL}_{t^*} = 0.$$

Proof: Consider the continuous and differentiable function $g(t) = NTI(t) = \log \mu(t)$. Then $g'(t) = d \log \mu(t)/dt = \mathcal{KL}_t$ and $g''(t) = V_{p_t}\Big\{ \log \frac{p_1(\boldsymbol{\theta})}{p_0(\boldsymbol{\theta})} \Big\} > 0$; where $V_{p_t}\Big\{ \log \frac{p_1(\boldsymbol{\theta})}{p_0(\boldsymbol{\theta})} \Big\}$ is the variance of $\log \frac{p_1(\boldsymbol{\theta})}{p_0(\boldsymbol{\theta})}$ with respect to $p_t(\boldsymbol{\theta})$. Since $g'(t^*) = \mathcal{KL}_{t^*} = 0$ and $g''(t^*) > 0$, then $g(t^*) = \min_{t \in [0,1]} \log \mu(t)$. Hence, from (5.12) we have that

$$C(p_1 \parallel p_0) = \max_{t \in [0,1]} C_t(p_1 \parallel p_0) = \min_{t \in [0,1]} NTI(t) = NTI(t^*).$$

□

The Chernoff information is often used to identify an upped bound of the probability of error of the Bayes rule in classification problems with two possible decisions including hypothesis testing; see Nussbaum,M and Szkoła, A. (2009) and Cover and Thomas (1991) for details. It has been also used in a variety of scientific fields, primarily as a measure of similarity between two distributions, as for example in cryptography (Baignères et al., 2010). The estimation of the Chernoff information is straightforward and it has been treated sporadically in problem-specific cases; see for example in Nielsen (2011) for computation in exponential families, or in Julier (2006) for Gaussian mixture models. The result of Lemma 5.2.2 can be used to construct a general algorithm for the estimation of the Chernoff information for any choice of p_1 and p_0 which is described in detail in Section 5.2.2.1.

Before proceeding any further, we may first outline the *balance property* of the NTI, which is based on the anti-symmetry property $C_t(p_1 \parallel p_0) = C_{1-t}(p_0 \parallel p_1)$, considered in Crooks and Sivak (2011).

The balance property: For any intermediate point $t \in (0,1)$ it holds that

$$NTI(t) = -\overline{NTI}(t) \text{ with } \overline{NTI}(t) = \int_t^1 \int_{\boldsymbol{\theta}} p_t(\boldsymbol{\theta}) \log \frac{p_1(\boldsymbol{\theta})}{p_0(\boldsymbol{\theta})} \, d\boldsymbol{\theta} \qquad (5.13)$$

and therefore the maximum absolute value occurs at t^* and it is equal to NTI(t^*).

Based on Lemmas 5.2.1 and 5.2.2 and the balance property, it occurs that the Chernoff $t-$divergences (either from p_1 to p_0 or in the opposite direction) can be directly computed from the NTI. Subsequently, a number of other divergencies related to Chernoff can be obtained from NTI. The *Bhattacharyya distance* (Bhattacharyya, 1943) occurs at $t = 0.5$, that is

$$Bh(p_1, p_0) = C_{0.5}(p_1 \parallel p_0) = -\log \int_{\boldsymbol{\theta}} \sqrt{p_1(\boldsymbol{\theta}) p_0(\boldsymbol{\theta})} d\boldsymbol{\theta} = -\log \rho_B.$$

The Bhattacharyya coefficient ρ_B can be implemented in turn to derive the *Bhattacharyya-Hellinger distance* (Bhattacharyya, 1943; Hellinger, 1909) since $He(p_1, p_0) = \sqrt{1 - \rho_B}$. Based on the Chernoff t-divergence we may also derive the *Rényi t-divergence*

$$R_t(p_1 \parallel p_0 = \frac{1}{t-1} \log \int_{\boldsymbol{\theta}} p_1(\boldsymbol{\theta})^t p_0(\boldsymbol{\theta})^{1-t} d\boldsymbol{\theta} = C_t(p_1 \parallel p_0)/(1-t)$$

(Rényi, 1961) and the Tsallis t-relative entropy

$$T_t(p_1 \parallel p_0 = \frac{1}{t-1} \left\{ \int_{\boldsymbol{\theta}} p_1(\boldsymbol{\theta})^t p_0(\boldsymbol{\theta})^{1-t} d\boldsymbol{\theta} - 1 \right\} = \big[\exp\big\{ - C_t(p_1 \parallel p_0) \big\} - 1 \big]/(1-t).$$

A graphical representation of the NTI is given in Figure 5.1. The cross entropy differences between p_t and the endpoint distributions (p_0 and p_1) are depicted on the vertical axis. The KL-divergencies between p_0 and p_1 are located at the endpoints of $[0, 1]$. Their difference represents the $J-$divergence. From Lemma 5.2.1, the Chernoff $t-$divergence for any $t_i \in [0, 1]$ is given by the area between the curve and the t-axis from $t = 0$ to $t = t_i$. The Chernoff information is given by the corresponding area up to $t = t^*$ while the Bhattacharyya distance is given by the corresponding area from zero up to $t = 0.5$.

To sum up, in this section it was illustrated how entropy measures are directly associated with the NTI. For this reason, all these measures can be derived using path sampling. Hence, the NTI given in (5.6) can offer another link between Bayesian inference, information theory and thermodynamics (or statistical mechanics). For instance, under the Hamiltonian $\mathcal{H}_t(\boldsymbol{\theta})$, Merhav (2010, Section 3.3) discuss the *excess* or *dissipated* work in thermodynamics and its relation to the data processing theorem in information theory, with the NTI emerging in the case of reversible processes. In a more general framework, Crooks and Sivak (2011) consider *conjugate trajectories*, that is forward (from $t = 0$ to $t = 1$) and backward processes (from $t = 1$ to $t = 0$), to derive the physical significance of the $f-$divergencies considered here, in terms of non-equilibrium dynamics. Note also that the balance property (5.13) satisfies the (recently derived) equality of Jarzynski (1997) and confirms Crooks's (1999) theorem; see, for details, in Merhav (2010) and Crooks and Sivak (2011). Further parallelism between the NTI and statistical mechanics is not attempted here, leaving this part to the experts on the field. In the next section we focus on the study of the MCMC estimators of $\log \lambda$ constructed using TI and geometric paths. We further study and analyse how the $f-$divergencies can be estimated as path sampling byproducts.

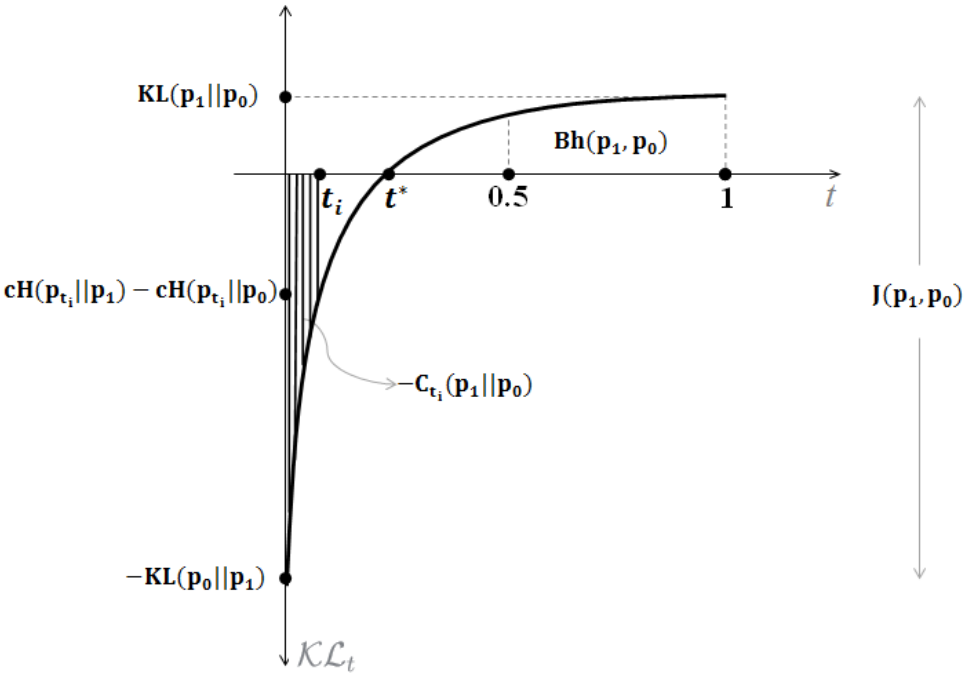


Figure 5.1: Graphical representation of the NTI: the plot of $\mathcal{KL}_t(\boldsymbol{\theta})$ over t.

5.2.2 MCMC path sampling estimators

Numerical approaches are typically used to compute the external integral of (5.2), such as the trapezoidal or Simpson's rule (Ogata, 1989; Neal, 1993; Gelman and Meng, 1998, among others). The numerical approaches require the formulation of an n-point discretisation $\mathcal{T} = \{t_0, t_1, \ldots, t_n\}$ of $[0, 1]$, such that $0 = t_0 < ... < t_{n-1} < t_n = 1$, which is called *temperature schedule*. A separate MCMC run is performed at each t_i with target distribution the corresponding $p(\boldsymbol{\theta}|\, t_i)$, $i = 0, ..., n$. The MCMC output is then used to estimate $\mathcal{E}_t = E_{p_t}\{U(\boldsymbol{\theta})\}$ by the sample mean $\widehat{\mathcal{E}_t}$ of the simulated values $\{\boldsymbol{\theta}^{(r)}\}_{r=1}^{R}$ generated from p_t for each $t \in \mathcal{T}$. The final estimator is derived by

$$\log \widehat{\lambda} \quad = \quad \sum_{i=0}^{n-1} (t_{i+1} - t_i) \frac{\widehat{\mathcal{E}}_{t_{i+1}} + \widehat{\mathcal{E}}_{t_i}}{2}; \tag{5.14}$$

see also in Friel and Pettitt (2008).

At a second step, the posterior output at each t_i and $\log \widehat{\lambda}$ can be employed to estimate t^* and the Chernoff information. Here we provide an algorithm for that purpose, which yields also the estimated Chernoff $t-$divergencies for any $t \in (0, 1)$ and subsequently the $f-$divergencies described in Section 5.2.1.

5.2.2.1 Estimation of the Chernoff $t-$ divergencies and information

Estimating the Chernoff information is generally a non-trivial and cumbersome procedure. For instance, Nielsen (2011) describe a *geodesic bisection optimization algorithm* that approximates $C(p_1 \parallel p_0)$ for multidimensional distributions which belong to the exponential family, based on Bregman divergences (named after Bregman, who introduced the concept in Bregman, 1967). Julier (2006) provides also an approximation for Gaussian mixture models. Here we introduce a TI based MCMC method for the estimation of Chernoff information which can be used for any choice of p_0 and p_1 distributions.

Following Lemma 5.2.2, the Chernoff information is given by $NTI(t^*)$. Therefore, in order to compute the Chernoff information we need first to estimate t^* for which $\mathcal{KL}_{t^*}$ is zero. The computation of t^* can be achieved by adding a number of steps in the path sampling procedure according to the following algorithm.

Step 1 Perform n MCMC runs to obtain $\widehat{\mathcal{E}}_t$ for all $t \in \mathcal{T}$ and $\log \widehat{\lambda}$ from (5.14).

Step 2 Calculate $\widehat{\mathcal{KL}}_t = \widehat{\mathcal{E}}_t - \log \widehat{\lambda}$ for all $t \in \mathcal{T}$.

Step 3 Identify interval $\big(t^-_{i^*}, t^+_{i^*+1}\big)$ where the sign of $\mathcal{KL}_t$ changes; where

$$t^-_i = \max\big(t \in \mathcal{T} : \widehat{\mathcal{KL}}_t < 0\big) \quad \text{and} \quad t^+_i = \min\big(t \in \mathcal{T} : \widehat{\mathcal{KL}}_t > 0\big) \ .$$

Note, that $\mathcal{KL}_t$ will be negative for any $t < t^*$ and positive otherwise since since $\frac{d\mathcal{KL}_t}{dt} = V_{p_t}\Big\{\log \frac{p_1(\boldsymbol{\theta})}{p_0(\boldsymbol{\theta})}\Big\} > 0$ and therefore $\mathcal{KL}_t$ it is an increasing function of t.

Step 4 Perform extra MCMC cycles by further discretising $\big(t^-_{i^*}, t^+_{i^*+1}\big)$ until the required precision is achieved.

Step 5 Update $\mathcal{T}$ and n to account for the new points $t_i \in \big(t^-_{i^*}, t^+_{i^*+1}\big)$ used in Step 5.

Step 6 Once the t^* is estimated, the MCMC output already available from the runs in Steps 1 and 4 can be used to estimate the Chernoff information. In particular, it is estimated as described in (5.14) having substituted $\widehat{\mathcal{E}}_t$ by $\widehat{\mathcal{KL}}_t$ for all $t \in \mathcal{T}$ and only accounting for $t_i \leq t^*$ in the summation. Therefore, the Chernoff information is estimated by $\widehat{NTI}(t^*)$ given by

$$\begin{aligned} \log \widehat{NTI}(t^*) &= \sum_{i \in \mathcal{I}:\, t_{i+1} \leq t^*} (t_{i+1} - t_i) \frac{\widehat{\mathcal{KL}}_{t_{i+1}} + \widehat{\mathcal{KL}}_{t_i}}{2} \\ &= \sum_{i \in \mathcal{I}:\, t_{i+1} \leq t^*} (t_{i+1} - t_i) \frac{\widehat{\mathcal{E}}_{t_{i+1}} + \widehat{\mathcal{E}}_{t_i}}{2} - t^* \log \widehat{\lambda} \ , \end{aligned} \tag{5.15}$$

where the $\mathcal{I} = \{0, 1, \ldots, n\}$ and $n = |\mathcal{T}|$.

In the special case where the path sampling is combined with output from MCMC algorithms which involve tempered transitions (see Calderhead and Girolami, 2009 for details), the estimation of the Chernoff information comes with low computational cost. This approach can be attractive and useful in the case of multi-modal densities. The same algorithm can be also implemented to compute the rest of the f-divergencies measures discussed in Section 5.2.1. In fact, their estimation is less demanding since it requires one additional MCMC run, in order to derive the estimated $\mathcal{KL}_{t_i}$ at the point of interest; for instance at t_i=0.5 we derive the $Bh(p_1,\, p_0)$ and $He(p_1,\, p_0)$ divergencies.

5.2.3 Error, temperature schedule and geometric perspective

In this section we study two important sources of error for path sampling estimators: the *path-related variance* and the *discretisation error*. The path-related variance is the error related to the choice of the path which, for geometric ones, is restricted to the selection of the endpoint densities. On the other hand, for any given path, the discretisation error is related to the choice of the temperature schedule $\mathcal{T}$ and is derived from the numerical approximation of the integral over $[0, 1]$. In order to examine these two error sources, we provide a geometric representation of TI (eq. 5.4) and NTI (eq. 5.6) identities. This leads us to a better understanding of the behaviour of the path sampling estimators.

5.2.3.1 Path-related variance

The total variance of $\log\widehat{\lambda}$ has been reported by Gelman and Meng (1998) in the case of stochastic t with an appropriate prior distribution attached to it. Further results were also presented by Lefebvre et al. (2010) for geometric paths. They have showed that the total variance is associated with the $J-$divergence of the endpoint densities and therefore with the choice of the path. Here we focus on the t-specific variances $V_t = V_{p_t}\{U(\boldsymbol{\theta})\} > 0$ of $U(\boldsymbol{\theta})$ (hereafter *local variance*) which are the components of the total variance.

Figure 5.2 is a graphical representation of TI. To be more specific, the curve represents the $\mathcal{E}_t$ values for each $t \in [0, 1]$ while the area between the t-axis and the curve gives the thermodynamic integral (5.2). In this figure, the error of the TI estimators is depicted by the steepness of the curve of $\mathcal{E}_t$. This result is based on the fact that the *partition function* z_t is the cumulant generating function of $U(\boldsymbol{\theta})$ (Merhav, 2010, section 2.4) and therefore the first derivative of $\mathcal{E}_t$ is given by the local variance V_t, that is $\mathcal{E}'_t = V_t$. It follows that the slope of the tangent of the curve at each t equals to V_t. Therefore, the graphical representation of two competing paths can provide valuable information about the associated variances of their corresponding estimators.

In the case of geometric paths particularly, $J(p_1,\, p_0)$ coincides with the slope of the secant defined at the endpoints of the curve and lays below the curve of the strictly

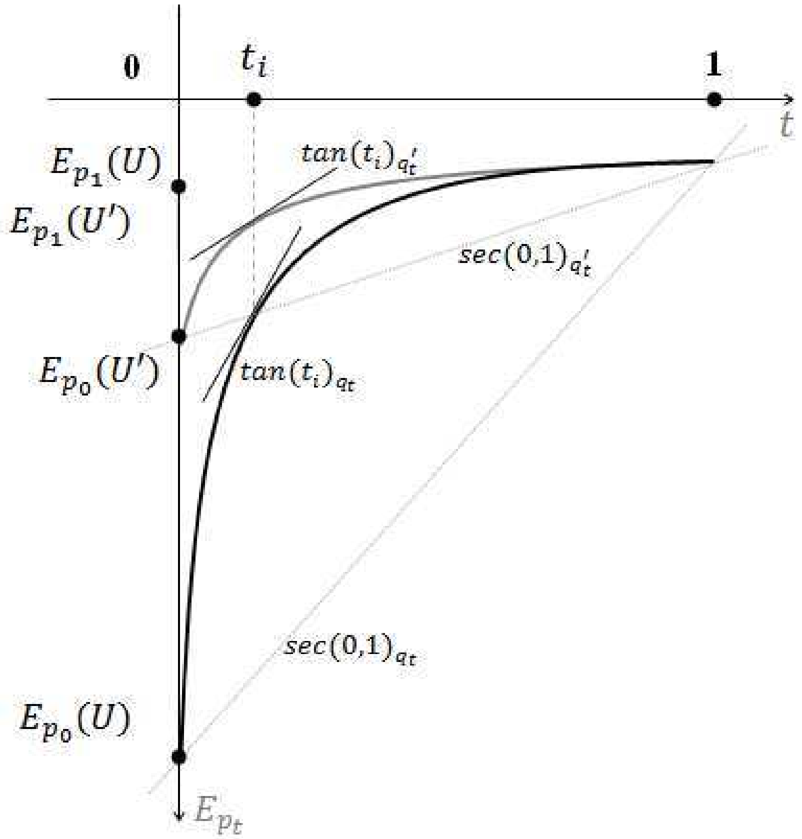


Figure 5.2: Graphical representation of the TI: the plot of the curve $\mathcal{E}_t = E_{p_t}\{U(\boldsymbol{\theta})\}$ over t, based on two paths q_t (black line) and q'_t (grey line). For each path, the $J-$distance between the endpoints coincides with the slope of the corresponding secant, $sec(0,1)$. The slope of the tangent $tan(t_i)$ equals the local variance V_{t_i}.

increasing (in terms of t) function $\mathcal{E}_t$. Therefore, it can be used as an indicator of the slope of the curve and the result of Lefebvre et al. (2010) has a direct visual realisation. The result can be generalised for any other pair of successive points, say $(t_i, \mathcal{E}_{t_i})$ and $(t_{i+1}, \mathcal{E}_{t_{i+1}})$, with the corresponding slope (or gradient) of the secant $sec(t_i, t_{i+1})$ given by

$$\nabla sec(t_i, t_{i+1}) \quad = \quad \frac{\mathcal{E}_{t_{i+1}} - \mathcal{E}_{t_{i+1}}}{t_{i+1} - t_i} = \frac{\mathcal{KL}_{t_{i+1}} - \mathcal{KL}_{t_i}}{t_{i+1} - t_i}. \tag{5.16}$$

The latter is derived from (5.7) and it reflects the fact that the slopes of the curves depicted in Figures 5.1 and 5.2 are identical. Additionally, $\mathcal{KL}_t$ can be written in terms of the KL-divergence between the successive sampling densities p_{t_i} and $p_{t_{i+1}}$ since, from (5.9) we obtain

$$\begin{aligned} KL(p_{t_i} \parallel p_{t_{i+1}}) &= \int_{\boldsymbol{\theta}} p_{t_i}(\boldsymbol{\theta}) \log \left\{ p_1(\boldsymbol{\theta})^{t_i - t_{i+1}} p_0(\boldsymbol{\theta})^{t_{i+1} - t_i} \right\} d\boldsymbol{\theta} + \log \frac{\mu(t_{i+1})}{\mu(t_i)} \\ &= -(t_{i+1} - t_i)\mathcal{KL}_{t_i} + \log \frac{\mu(t_{i+1})}{\mu(t_i)}. \end{aligned} \tag{5.17}$$

Using (5.16) and (5.17), we can associate the $J-$divergence between two successive points with the slope of the secant $sec(t_i, t_{i+1})$ since

$$\nabla sec(t_i, t_{i+1}) = \frac{J(p_{t_i}, p_{t_{i+1}})}{(t_{i+1} - t_i)^2} \tag{5.18}$$

generalizing the result of Lefebvre et al. (2010) for the endpoints of the graph where the slope of the $sec(0, 1)$ is given by $J(p_1, p_0)$. For successive points closely placed to each other (that is, for $\Delta(t_i) = t_{i+1} - t_i \to 0$) the slope of the secant approximates the corresponding slope of the tangent of the curve and therefore the local variance. Hence, the $J-$divergence between any two successive points is indicative of the slope of the curve and consequently of the associated variance. For example, in Figure 5.2 for values of t close to zero the slope of curve is very steep indicating high local variability.

The local variances of the path sampling estimators discussed here depend on the selection of the path. In the next section, we proceed with the study of the discretisation error and its effect on the path sampling estimators based on both the TI and NTI identities for any fixed geometric path.

5.2.3.2 Discretisation error

Calderhead and Girolami (2009) expressed the discretisation error in terms of differences of relative entropies of successive (in terms of t) sampling distributions. The result of Calderhead and Girolami (2009) can be written for any geometric path as follows

$$\begin{aligned} \log \widehat{\lambda} = \sum_{i=0}^{n-1} \frac{\widehat{z}_{t_{i+1}}}{\widehat{z}_{t_i}} &= \frac{1}{2} \sum_{i=0}^{n-1} (t_{i+1} - t_i) \left\{ \widehat{\mathcal{E}}_{t_{i+1}} + \widehat{\mathcal{E}}_{t_i} \right\} \\ &+ \frac{1}{2} \sum_{i=0}^{n-1} \left\{ \widehat{KL}(p_{t_i} \parallel p_{t_{i+1}}) - \widehat{KL}(p_{t_{i+1}} \parallel p_{t_i}) \right\}, \end{aligned} \tag{5.19}$$

Calderhead and Girolami (2009) consider the case for $\Delta(t_i) \to 0$ in (5.19) and outline that the first summation is equivalent to the trapezium rule used for numerical integration with the associated error expressed in terms of the asymmetries between the KL divergencies defined between p_{t_i} and $p_{t_{i+1}}$. In view of (5.17), expression 5.19 becomes

$$\log \widehat{\lambda} = \frac{1}{2} \sum_{i=0}^{n-1} \Delta(t_i) \left\{ \widehat{\mathcal{E}}_{t_{i+1}} + \widehat{\mathcal{E}}_{t_i} \right\} - \frac{1}{2} \sum_{i=0}^{n-1} \Delta(t_i) (\widehat{\mathcal{KL}}_{t_i} + \widehat{\mathcal{KL}}_{t_{i+1}}), \tag{5.20}$$

since $\sum_{i=0}^{n-1} \log \frac{\mu(t_i)}{\mu(t_{i+1})} = 0$. The second term in the left side of (5.20) is the approximation of the NTI (using the trapezoidal rule), which indeed it should be zero. According to the discussion in Section 5.2.3.1, the relative entropies in (5.19), as well as the areas above and below the t-axis which represent the Chernoff divergencies, are not expected to be zero. They both represent the path-related variance which is independent (and pre-existing) of the discretisation error. The discretisation error consists of the asymmetries that occur under any particular tempering schedule either in the TI or in NTI. The symmetry is a feature of the thermodynamic integration and it represents the trade-off between uncertainty in the forward and backward trajectories. Therefore, the error manifests as lack of symmetry in the assessment of the uncertainty due to the discretisation, as explained below.

While the path-related variance is independent from the discretisation error, the reverse argument does not hold. In fact, the discretisation error is highly influenced and dependent upon the path-related variance. Consider two pairs of successive points, located close to the zero and unit endpoints in Figure 5.1, say $t_i^{(0)}, t_{i+1}^{(0)}$ and $t_j^{(1)}, t_{j+1}^{(1)}$ respectively, for $i, j = 1, ..., n$. Further assume that the distances between the points within each pair are equal, say $\delta > 0$. For the first pair, the corresponding $\mathcal{KL}_t$s on the vertical axis are distant due to the steepness of the curve. On the contrary, for the second pair the corresponding $\mathcal{KL}_t$s are very close, due to the fact that the slope of the curve is almost horizontal. Therefore, using the trapezoidal rule, for equally spaced pairs of points we approximate a large part of the curve towards the zero end and a small part of the curve towards the unit end. In order to achieve the same degree of accuracy at both ends, the second pair of points need to be closer. In conclusion, the temperature schedule should place more points towards the end of the path where the uncertainty (slope) is higher. For instance, the powered fraction (PF) schedule (Friel and Pettitt, 2008)

$$\mathcal{T}_{PF} = \{t_i\}_{i=1}^{n} \text{ such as } t_i = (1/n)^{\mathcal{C}}, \mathcal{C} = 1/a > 1, \tag{5.21}$$

places more points towards the zero endpoint of the path. Xie et al. (2011) proposed a closely related geometric schedule where the t_is are chosen according to evenly spaced quartiles of a $Beta(a, 1)$ distribution. Recently, Friel et al. (2012) proposed an adaptive algorithm for the temperature schedule that takes under consideration the local variances in order to locate the high uncertainty points. The algorithm traces the points on the curve and assigns more t_is close to their regions. The gain in the error is then achieved with a small computational price.

5.3 Bayesian model comparison using tempered transitions

The Bayesian marginal likelihood is simply the normalizing constant of the posterior distribution $f(\boldsymbol{\theta}|\boldsymbol{y}, m_i)$ and can be estimated by path sampling. Recently, such methods have been considered for Bayesian marginal likelihood estimation by Lartillot and Philippe (2006), Friel and Pettitt (2008) and Lefebvre et al. (2010).

5.3.1 The stepping-stone identity

In this section we consider an alternative approach that is based on the stepping-stone sampling, presented by Xie et al. (2011) and Fan et al. (2011) for the estimation of the Bayesian marginal likelihood. Closely related ideas are also discussed in the context of the free energy estimation in Neal (1993, see section 6.2 and references within). The stepping-stone sampling considers finite values $t_i \in \mathcal{T}$, that are placed according to a temperature schedule as the ones discussed in Section 5.2.3. The ratio of the normalizing constants can be expressed as

$$\lambda = \frac{z_1}{z_0} = \frac{z_{t_n}}{z_{t_{n-1}}} \frac{z_{t_{n-1}}}{z_{t_{n-2}}} \cdots \frac{z_{t_1}}{z_{t_0}} = \prod_{i=0}^{n-1} \frac{z_{t_{i+1}}}{z_{t_i}}.$$

Hence, the ratio of the normalizing constants can be estimated using $z_{t_{i+1}}/z_{t_i}$ as an intermediate step that can be estimated from t specific MCMC samples based on the identity

$$\frac{z_{t_{i+1}}}{z_{t_i}} = \int_{\boldsymbol{\theta}} \frac{q_{t_{i+1}}(\boldsymbol{\theta})}{q_{t_i}(\boldsymbol{\theta})} \, p_{t_i}(\boldsymbol{\theta}) \, d\boldsymbol{\theta};$$

see Xie et al. (2011) for details. For geometric paths, the stepping-stone identity for λ is then given by

$$\lambda = \prod_{i=0}^{n-1} \int_{\boldsymbol{\theta}} \left\{ \frac{q_1(\boldsymbol{\theta})}{q_0(\boldsymbol{\theta})} \right\}^{\Delta(t_i)} p_{t_i}(\boldsymbol{\theta}) \, d\boldsymbol{\theta}. \tag{5.22}$$

Xie et al. (2011) presented the stepping-stone sampling specifically for estimating the Bayesian marginal likelihood (under a certain geometric path) while Fan et al. (2011) modified the initial Bayesian marginal likelihood estimator in order to improve its properties (both estimators are addressed later on in this section). However, as outlined here, the stepping-stone sampling can be considered as a general method, alternative to path sampling, that can be applied for the estimation of ratios of unknown normalized constants.

Hence, identities (5.4) and (5.22), are two closely related alternative tempered transition methods for the estimation of normalizing constants using geometric paths. Any estimator developed via thermodynamic integration has its corresponding stepping-stone estimator and vise versa. In the next section, we present existing methods classified by the tempered method that has been originated and the adopted path. This method-path approach allows us to further introduce new estimators based on the counterpart existing ones.

5.3.2 Bayesian marginal likelihood estimators

In order to avoid confusion, hereafter we will name each estimator based on the method (thermodynamic or stepping-stone) and on the path implemented for its derivation.

The power posteriors (Lartillot and Philippe, 2006, Friel and Pettitt, 2008) and the the stepping stone (Xie et al., 2011) Bayesian marginal likelihood estimators are using the same geometric path but they are based on different identities, approaching the same problem using a different perspective. Both methods implement the geometric *prior-posterior* path, namely

$$q_t^{PP}(\boldsymbol{\theta}) = \{f(\boldsymbol{y}|\,\boldsymbol{\theta})\pi(\boldsymbol{\theta})\}^t\,\pi(\boldsymbol{\theta})^{1-t} = f(\boldsymbol{y}|\,\boldsymbol{\theta})^t\pi(\boldsymbol{\theta}), \tag{5.23}$$

where $q_0(\boldsymbol{\theta}) = \pi(\boldsymbol{\theta})$ is a proper prior for the model parameters and $q_1(\boldsymbol{\theta}) = f(\boldsymbol{\theta}|\,\boldsymbol{y})\,\pi(\boldsymbol{\theta})$ is the corresponding unnormalized posterior density. Setting the prior-posterior in (5.4) and (5.22), yields the thermodynamic and the stepping-stone prior-posterior identities (PP$_T$ and PP$_S$ respectively) for the Bayesian marginal likelihood

$$\log f(\boldsymbol{y}) = \int_0^1 E_{p_t^{PP}}\{\log f(\boldsymbol{y}|\,\boldsymbol{\theta})\}\;dt \;\text{ and }\; f(\boldsymbol{y}) = \prod_{i=0}^{n-1}\int_{\boldsymbol{\theta}}\{\log f(\boldsymbol{y}|\,\boldsymbol{\theta})\}^{\Delta(t_i)}\,p_{t_i}^{PP}(\boldsymbol{\theta})\,d\boldsymbol{\theta}$$

where $p_t^{PP}(\boldsymbol{\theta}|\boldsymbol{y})$ is the density normalized version of (5.23).

Fan et al. (2011) modified the estimator of Xie et al. (2011) using instead the *importance-posterior* path

$$q_t^{IP}(\boldsymbol{\theta}) = \{f(\boldsymbol{y}|\,\boldsymbol{\theta})\,\pi(\boldsymbol{\theta})\}^t\,g(\boldsymbol{\theta})^{1-t}.$$

The importance posterior path was one of the paths that Lefebvre et al. (2010) considered for the estimation of the Bayesian marginal likelihood. It should be noted that the density $g(\boldsymbol{\theta})$ is required to be proper so that $z_0 = 1$. It can be constructed by implementing the posterior moments available from the MCMC output at $t = 1$. The thermodynamic and stepping-stone importance-posteriors (IP$_T$ and IP$_S$ respectively) are derived by the identities

$$\begin{aligned}\log f(\boldsymbol{y}) &= \int_0^1 E_{p_t}^{IP}\left[\log\frac{f(\boldsymbol{y}|\,\boldsymbol{\theta})\,\pi(\boldsymbol{\theta})}{g(\boldsymbol{\theta})}\right]dt \;\text{ and} \\ f(\boldsymbol{y}) &= \prod_{i=0}^{n-1}\int_{\boldsymbol{\theta}}\left\{\frac{f(\boldsymbol{y}|\,\boldsymbol{\theta})\,\pi(\boldsymbol{\theta})}{g(\boldsymbol{\theta})}\right\}^{\Delta(t_i)} p_{t_i}^{IP}(\boldsymbol{\theta})\,d\boldsymbol{\theta},\end{aligned} \tag{5.24}$$

where $p_t^{IP}(\boldsymbol{\theta})$ is the density normalized version of $q_t^{IP}(\boldsymbol{\theta})$.

The TI identity appearing in (5.24) has the attractive feature of sampling from $g(\boldsymbol{\theta})$, rather than the prior, for $t = 0$. It also retains the stability ensured by averaging in log scale according to the thermodynamic approach. Therefore, in specific model settings, the estimators based on the thermodynamic importance posteriors can perform more efficiently than estimators based on the other expressions, provided that an importance function can be formulated. It is our belief that beyond the four expressions reviewed here, others may be developed within this broad framework, by choosing the appropriate path for particular models, coming with thermodynamic and stepping-stone variants.

5.3.3 Bayes factor direct estimators

The BF is by definition a ratio of normalized constants. Therefore, (5.4) and (5.22) can be implemented to construct direct BF estimators, rather than applying the methods to each model separately. Lartillot and Philippe (2006) implemented the thermodynamic integration, in order to link two competing (not necessary nested) models, instead of densities. That was achieved by choosing the appropriate path, in a way that eventually produces directly a BF estimator. Lartillot and Philippe (2006) were motivated by the fact that lack of precision on each Bayesian marginal likelihood estimation, may alter the BF interpretation. They argue, that a simultaneous estimation of the two constants can ameliorate that to some extend. The idea is to employ a bidirectional *melting-annealing* sampling scheme, based on the *model-switch* path:

$$q_t^{MS}(\boldsymbol{\theta}) = \{f(\boldsymbol{y}|\,\boldsymbol{\theta}, m_1)\,\pi(\boldsymbol{\theta}|\,m_1)\}^t\,\{f(\boldsymbol{y}|\,\boldsymbol{\theta}, m_0)\,\pi(\boldsymbol{\theta}|\,m_0)\}^{1-t}\,.$$

Lartillot and Philippe's (2006) thermodynamic model-switch (MS$_T$) identity for the BF and its stepping-stone counterpart (MS$_S$) are as follows

$$\log BF_{10} = \int_0^1 E_{p_t^{MS}}\left[\log\left\{\frac{f(\boldsymbol{y}|\,\boldsymbol{\theta}, m_1)\,\pi(\boldsymbol{\theta}|\,m_1)}{f(\boldsymbol{y}|\,\boldsymbol{\theta}, m_0)\,\pi(\boldsymbol{\theta}|\,m_0)}\right\}\right]dt$$

and

$$BF_{10} = \prod_{i=0}^{n-1}\int_{\boldsymbol{\theta}}\left\{\frac{f(\boldsymbol{y}|\,\boldsymbol{\theta}, m_1)\,\pi(\boldsymbol{\theta}|\,m_1)}{f(\boldsymbol{y}|\,\boldsymbol{\theta}, m_0)\,\pi(\boldsymbol{\theta}|\,m_0)}\right\}^{\Delta(t_i)} p_{t_i}^{MS}(\boldsymbol{\theta}|\,\boldsymbol{y})\,d\boldsymbol{\theta},$$

where the expectation is taken over $p_t^{MS}(\boldsymbol{\theta}|\boldsymbol{y})$ which is the density obtained after normalizing the model-switch path $q_t^{MS}(\boldsymbol{\theta})$. In case where $\boldsymbol{\theta}$ is common between the two models (for instance if the method is used to compare paths under different endpoints, see Lartillot and Philippe, 2006 for an example) the method is directly applicable. Otherwise, if $\boldsymbol{\theta} = (\boldsymbol{\theta}_{m_1}, \boldsymbol{\theta}_{m_0})$, pseudo-priors need to be assigned at the endpoints of the path.

Having in mind the direct estimation of Bayes factors, more complicated estimators may be derived using *compound* geometric paths. With the term compound paths we refer to paths that consist of a *hyper* geometric path, $Q_t(\boldsymbol{\theta}) = Q_1(\boldsymbol{\theta})^t Q_0(\boldsymbol{\theta})^{1-t}$, used to link two competing models and a *nested* path $q_t(\boldsymbol{\theta}, i)$ for each endpoint function Q_i, for $i = 0, 1$. The two intersecting paths form a *quadrivial*, $(Q \circ q)_t(\boldsymbol{\theta})$ with $t \in [0, 1]$ that can be defined as

$$(Q \circ q)_t(\boldsymbol{\theta}) = \left[q_1(\boldsymbol{\theta}, 1)^t q_0(\boldsymbol{\theta}, 1)^{1-t}\right]^t \left[q_1(\boldsymbol{\theta}, 0)^t q_0(\boldsymbol{\theta}, 0)^{1-t}\right]^{1-t}.$$

The multivariate extension is discussed in detail in Gelman and Meng (1998). The endpoint target densities are given by $q_i(\boldsymbol{\theta}, i)$ for $t = 0$ and $t = 1$ respectively estimating the ratio $z_1/z_0 = \int q_1(\boldsymbol{\theta}, 1) d\boldsymbol{\theta} \times \left[\int q_0(\boldsymbol{\theta}, 0) d\boldsymbol{\theta}\right]^{-1}$. The densities $q_i(\boldsymbol{\theta}, j)$ for $i, j = 0, 1$ and $i \neq j$ serve as linking densities within each nested path. Therefore, following the importance-sampling logic, they should play the role of approximating (importance) functions for each $q_i(\boldsymbol{\theta}, i)$.

For the specific case of the Bayes factor estimation, the objective is to retrieve the Bayesian marginal likelihoods at the endpoints and therefore it is reasonable to consider as nested paths the prior-posterior and the importance-posterior paths, discussed in the previous section. The importance-posterior BF quadrivial, for instance, is as follows

$$\begin{aligned}(Q \circ q)_t^{IP}(\boldsymbol{\theta}) &= \left[\left\{f(\boldsymbol{y}|\,\boldsymbol{\theta}, m_1)\pi(\boldsymbol{\theta}|\,m_1)\right\}^t g(\boldsymbol{\theta}|\,m_1)^{1-t}\right]^t \\ &\qquad \times \left[\left\{f(\boldsymbol{y}|\,\boldsymbol{\theta}, m_0)\pi(\boldsymbol{\theta}|\,m_0)\right\}^{1-t} g(\boldsymbol{\theta}|\,m_0)^t\right]^{1-t}\end{aligned}$$

leading to the thermodynamic (Q_{IP_T}) and stepping-stone (Q_{IP_S}) expressions

$$\log BF_{10} = \int_0^1 E_{P_t}\left[\log \frac{\left\{f(\boldsymbol{y}|\,\boldsymbol{\theta}, m_1)\,\pi(\boldsymbol{\theta}|\,m_1)/g(\boldsymbol{\theta}|\,m_1)\right\}^{2t} g(\boldsymbol{\theta}|\,m_1)}{\left\{f(\boldsymbol{y}|\,\boldsymbol{\theta}, m_0)\,\pi(\boldsymbol{\theta}|\,m_0)/g(\boldsymbol{\theta}|\,m_0)\right\}^{2(1-t)} g(\boldsymbol{\theta}|\,m_0)}\right] dt$$

and

$$BF_{10} = \prod_{i=0}^{n-1} \int_{\boldsymbol{\theta}} \log \frac{\left\{f(\boldsymbol{y}|\,\boldsymbol{\theta}, m_1)\,\pi(\boldsymbol{\theta}|\,m_1)/g(\boldsymbol{\theta}|\,m_1)\right\}^{2T_i} g(\boldsymbol{\theta}|\,m_1)}{\left\{f(\boldsymbol{y}|\,\boldsymbol{\theta}, m_0)\,\pi(\boldsymbol{\theta}|\,m_0)/g(\boldsymbol{\theta}|\,m_0)\right\}^{2(1-T_i)} g(\boldsymbol{\theta}|\,m_0)}\, P_{t_i}(\boldsymbol{\theta})\, d\boldsymbol{\theta},$$

where $P_t(\boldsymbol{\theta}) = (Q \circ q)_t^{IP}(\boldsymbol{\theta}) = /Z_t$, $Z_t = \int_{\boldsymbol{\theta}} (Q \circ q)_t^{IP}\, d\boldsymbol{\theta}$, $t \in [0, 1]$. In the thermodynamic expression, t is the *melting* temperature and $1 - t$ the *annealing* one, assuming that the procedure starts at $t = 0$ and gradually increases to $t = 1$. The hyper-path ensures that while the model m_1 is melting, the model m_0 is annealing. At the same time, the importance-posterior path serving as the nested one, links the posterior with the importance at each model separately. In the stepping-stone counterpart expression the melting and annealing temperatures are given by $T_i = (t_{i+1}+t_i)/2$ for any $i = 0, 1, \ldots, n-1$.

From the expressions Q_{IP_T} and Q_{IP_S} we may derive the analogue ones for the prior-posterior quadrivial (Q_{PP_T} and Q_{PP_S}) by substituting the importance densities $g(\boldsymbol{\theta}|\,m_i)$ with the corresponding priors $\pi(\boldsymbol{\theta}|\,m_i)$, $(i = 0, 1)$. The quadrivial expressions, univariate and multivariate, are under ongoing research and it is not yet clear to the authors which applications could benefit from their complected structure. The optimal tempering scheme is also an open issue. In the next section, all estimators discussed here are applied in simulated examples.

5.4 Illustrative Examples

5.4.1 Regression modelling in the pine dataset

For the illustration of the estimators discussed in Section 5.3 we implement the Pine data set, which has been studied by Friel and Pettitt (2008) and Lefebvre et al. (2010) in the context of path sampling. The dataset consists of measurements taken on 42 specimens of Pinus radiata. A linear regression model was fitted for the specimen's maximum compressive strength (y), using their density (x) as independent variable, that is

$$y_i = \alpha + \beta(x_i - \bar{x}) + \epsilon_i,\ \epsilon_i \sim N(0, \sigma^2),\ i = 1, ..., 42. \tag{5.25}$$

The objective in this example is to illustrate how each method and path combination responds to prior uncertainty. To do so, we use three different prior schemes, namely:

$$\Pi_1 : \quad (\alpha, \beta)' \sim N\left\{(3000, 185)', (10^6, 10^4)'\right\},\ \sigma^2 \sim IG(3, 1.8 \times 10^5)\ ,$$

$$\Pi_2 : \quad (\alpha, \beta)' \sim N\left\{(3000, 0)', (10^5, 10^3)'\right\},\ \sigma^2 \sim IG(3, 1.8 \times 10^4)\ \ ,$$

$$\Pi_3 : \quad (\alpha, \beta)' \sim N\left\{(3000, 0)', (10^5, 10^3)'\right\},\ \sigma^2 \sim IG(0.3, 1.8 \times 10^4),$$

where $IG(a, b)$ denotes the inverse gamma distribution with shape a and rate b. The Bayesian marginal likelihoods were estimated over $n_1 = 50$ and $n_2 = 100$ evenly spaced temperatures. At each temperature, a Gibbs algorithm was implemented and 30,000 posterior observations were generated; after discarding 5,000 as a burn-in period. The posterior output was divided into 30 batches (of equal size of R_b=1,000 points) and all estimators were computed within each batch. The mean over all batches was used as the final estimate, denoted by $\log \widehat{\lambda}_i$ for each prior Π_i, $i = 1, 2, 3$. In order the estimators to be directly comparable in terms of error, the batch means method (Schmeiser, 1982, Bratley et al., 1987) was preferred. In particular, the standard deviation of the $\log \widehat{\lambda}$ over the 30 batches was considered as the estimated error, denoted hereafter by $\widehat{MCE}$. Lefebvre et al. (2010) used $n = 1001$ equally spaced points to compute the gold standard for $\log \widehat{\lambda}_1 = -309.9$. Following the same approach we derived $\log \widehat{\lambda}_2 = -323.3$ and $\log \widehat{\lambda}_3 = -328.2$. These values are considered as benchmarks in the current study. Finally, the importance functions for each model were constructed from the posterior means and variances at $t = 1$.

The estimations for the marginal likelihoods are presented in Table 5.1. The values that were obtained based on the importance-posterior path, reached the gold standards even when $n = 50$. The thermodynamic (IP$_T$) and the stepping–stone (IP$_S$) counterparts performed equally well and were associated with similar errors. On the contrary, the estimators that are based on the prior-posterior path yielded different values depending on the method. In particular, the stepping–stone estimator (PP$_S$) was fairly close to

Table 5.1: Marginal likelihood estimates - Pine data

n	Path/Method	$\log \widehat{\lambda}_1$	$\log \widehat{\lambda}_2$	$\log \widehat{\lambda}_3$
	PP_T	-312.9 (0.21)	-324.7 (0.19)	-352.4 (0.57)
	PP_S	-310.2 (0.06)	-322.6 (0.05)	-328.5 (0.03)
50	IP_T	-310.0 (0.02)	-323.4 (0.03)	-328.2 (0.03)
	IP_S	-310.0 (0.02)	-323.4 (0.03)	-328.2 (0.03)
100	PP_T	-311.3 (0.11)	-323.7 (0.14)	-339.0 (0.03)
	PP_S	-310.1 (0.06)	-323.5 (0.03)	-328.5 (0.03)
	IP_T	-309.9 (0.02)	-323.4 (0.02)	-328.2 (0.03)
	IP_S	-309.9 (0.02)	-323.4 (0.02)	-328.2 (0.03)

PP denotes the prior-posterior path and IP the importance posterior path. The indices T and S imply the thermodynamic and stepping–stone analogues.

the gold standards with low error, for all prior schemes. The thermodynamic estimator (PP_T) on the other hand, underestimated the Bayesian marginal likelihood and exhibited higher errors than all other methods. Logarithms of the ratios of the estimated Bayesian marginal likelihoods along with the estimated BF values directly derived by the model-switch methods are further presented in Table 5.2. The thermodynamic and stepping-stone analogues of MS, Q_{PP} and Q_{IP}, yielded estimates with similar values and errors.

In this example, we have used a uniform temperature schedule, moderate number of points n and non informative priors. It was therefore reasonable to expect that the prior-based methods would be associated with higher error. The interesting result here was that the stepping–stone estimator addressed the prior uncertainty more successfully. In fact, the thermodynamic and stepping–stone approaches coincided only when the gold standard was reached, which means that the discretisation error (5.19) was minimized. The next step in our analysis was to employ a temperature schedule that places more points towards the prior in order to reduce the uncertainty. The powered fraction (5.21) schedule (Friel and Pettitt, 2008) was used with $\mathcal{C} = 5$. For $n = 100$, the PP_T yielded the benchmark values for the Bayesian marginal likelihoods, namely $\log \widehat{\lambda}_1 = 310.0$ (0.01), $\log \widehat{\lambda}_2 = 323.5$ (0.01) and $\log \widehat{\lambda}_2 = 328.3$ (0.02). The results were almost identical for the PP_S.

Once the thermodynamic procedure yielded the benchmark values, we proceeded with the estimation of the entropy measures (see Section 5.2.1) presented in Table 5.3. The precision for the point t^* was set to the third decimal point and the extra MCMC runs costed less than a minute of computational time. The Bhattacharyya and Bhattacharyya-

Table 5.2: Estimated **log** ratio of the Bayesian marginal likelihoods

	$n = 50$		$n = 100$	
Path/Method	$\log\left(\widehat{\lambda}_2/\widehat{\lambda}_1\right)$	$\log\left(\widehat{\lambda}_3/\widehat{\lambda}_1\right)$	$\log\left(\widehat{\lambda}_2/\widehat{\lambda}_1\right)$	$\log\left(\widehat{\lambda}_3/\widehat{\lambda}_1\right)$
PP_T	-11.8 (0.21)	-39.5 (0.57)	-12.4 (0.14)	-26.0 (0.38)
PP_S	-12.5 (0.06)	-18.4 (0.73)	-12.5 (0.06)	-18.5 (0.34)
IP_T	-13.4 (0.04)	-18.2 (0.04)	-13.4 (0.03)	-18.2 (0.04)
IP_S	-13.4 (0.04)	-18.2 (0.04)	-13.4 (0.03)	-18.2 (0.01)
MS_T	-13.5 (0.01)	-18.2 (0.01)	-13.5 (0.01)	-18.2 (0.01)
MS_S	-13.5 (0.01)	-18.2 (0.01)	-13.5 (0.01)	-18.2 (0.01)
Q_{PP_T}	-13.5 (0.01)	-18.2 (0.01)	-13.5 (0.01)	-18.2 (0.01)
Q_{PP_S}	-13.5 (0.01)	-18.2 (0.02)	-13.5 (0.01)	-18.2 (0.01)
Q_{IP_T}	-13.5 (0.01)	-18.2 (0.01)	-13.5 (0.01)	-18.2 (0.01)
Q_{IP_S}	-13.5 (0.01)	-18.2 (0.01)	-13.5 (0.01)	-18.2 (0.01)

PP denotes the prior-posterior path and IP the importance posterior path. MS and Q stand for the model-switch and quadrivial (bidirectional) methods. The indices T and S imply the thermodynamic and stepping–stone analogues.

Hellinger values indicate that the priors Π_1, Π_2 and Π_3 where very distant from the corresponding posteriors. On the contrary, the importance functions were close approximations of their matching posterior densities. This fact completely explains the differences in the estimation, reflecting the increased local variances encountered by the PP_T as opposed to IP_T.

5.4.2 Bayesian marginal likelihood for latent trait models in a simulated dataset

According to the current results, the uncertainty in the pine data example was manageable under a suitable tempering schedule. This will not always be the case, especially in high dimensional problems. Here we consider also a factor analysis model with binary items. The dataset consists of $N = 400$ responses, $p = 4$ observed items and $k = 1$ latent variable and was previously considered in Chapter 4, within the context of Bayesian marginal likelihood estimation. Under the non informative prior for the 404 model parameters (see Section 2.2) the Bayesian marginal likelihood was estimated close to -977.8, based on the CJ^J estimator and the LM estimator. Using the same prior and importance functions as in Chapter 4, the PP and the IP paths were computed in order to derive the estimated Bayesian marginal likelihood. Due to the dimensionality of the model, $n = 200$ runs were used and 30,000 posterior observations from a Metropolis-within-Gibbs algorithm were

Table 5.3: Estimated $f-$divergencies

	Π_1		Π_2		Π_3	
$f-$**divergency**	PP_T	IP_T	PP_T	IP_T	PP_T	IP_T
$KL\ (p_1 \parallel p_0)$	5.6 (<0.01)	0.03 (<0.01)	16.3 (<0.01)	0.10 (<0.01)	24.8 (<0.01)	0.10 (<0.01)
$KL\ (p_0 \parallel p_1)$	414.8 (4.61)	0.06 (<0.01)	304.1 (5.71)	0.09 (<0.01)	3061.0 (53.1)	0.09 (<0.01)
$J\ (p_0, p_1)$	420.5 (4.62)	0.09 (<0.01)	320.4 (5.63)	0.20 (<0.01)	3085.0 (53.4)	0.02 (<0.01)
$Bh\ (p_0, p_1)$	2.53 (<0.01)	0.01 (<0.01)	6.68 (<0.01)	0.03 (<0.01)	11.4 (<0.01)	0.07 (<0.01)
$He\ (p_0, p_1)$	0.96 (<0.01)	0.11 (<0.01)	0.99 (<0.01)	0.17 (<0.01)	0.99 (<0.01)	0.26 (<0.01)
$C_{t^*}\ (p_0 \parallel p_1)$	3.38 (<0.01)	0.01 (<0.01)	7.24 (<0.01)	0.03 (<0.01)	15.0 (<0.01)	0.03 (<0.01)
$R_{t^*}\ (p_0 \parallel p_1)$	2.76 (<0.01)	0.01 (<0.01)	4.61 (<0.01)	0.02 (<0.01)	12.1 (<0.01)	0.02 (<0.01)
$T_{t^*}\ (p_0 \parallel p_1)$	1.19 (<0.01)	0.02 (<0.01)	1.57 (<0.01)	0.06 (<0.01)	1.24 (<0.01)	0.06 (<0.01)
t^*	0.183	0.552	0.445	0.363	0.192	0.437

$KL(\cdot \parallel \cdot)$: Kullback-Leibler relative entropy, $J(\cdot,\cdot)$: Jeffreys' divergence, $Bh(\cdot,\cdot)$: Bhattacharyya distance, $He(\cdot,\cdot)$: Bhattacharyya-Hellinger distance. Estimated at t^*: $C(\cdot \parallel \cdot)$: Chernoff information, $R(\cdot \parallel \cdot)$: Rényi relative entropy, $T(\cdot \parallel \cdot)$: Tsallis relative entropy. PP denotes the prior-posterior path and IP the importance posterior path. The indices T and S imply the thermodynamic and stepping–stone analogues.

derived at each temperature point (burn in period: 10,000 iterations, thinned by 10).

The batch means for the thermodynamic and stepping-stone importance posteriors were -978.1 and -977.9 respectively, with associated MCE errors 0.018 and 0.013. The corresponding values under the prior posterior path were -995.4 and -995.1 with associated MCE errors 0.032 and 0.027 respectively. The low MCEs indicated that the error was not stochastic but rather due to the temperature placement. Even though the powered fraction (5.21) schedule was used to place more values close to the prior ($\mathcal{C} = 5$), the uncertainty was not successfully addressed. The estimators did not improve when the process was replicated for $n = 500$. This example indicates that in high dimensional models with non informative priors, the PP_T and PP_S estimators can be deteriorated by discretisation error even for large n.

In all examples presented here, `R` was used along with the `OpenBUGS` software (version `3.2.2`; Lunn and Best, 2009). Specifically, the `R2WinBUGS` package (Sturtz et al., 2005) was used to obtain posterior samples at each temperature t_i via the `OpenBUGS`, which subsequently were employed in `R` in order to compute the final estimators.

5.5 Discussion

In this chapter the quest started from general thermodynamic approaches using geometric paths, was continued from the normalized thermodynamic integration to f-divergencies, and, finally, concluded to the BML and BF estimators.

The study through these topics offers a direct connection between thermodynamic integration and divergence measures such as Kullback-Leibler and Chernoff divergencies, Chernoff information and other divergencies emerging as special cases or functions of them. By this way, we were able to offer an efficient MCMC based thermodynamic algorithm for the estimation of the Chernoff information for a general framework which was not available in the past.

Moreover, the study of the thermodynamic identities and integrals has lead us to an understanding of the error sources of the TI estimators. All these are accompanied with detailed graphical and geometric representation and interpretation offering insight to the thermodynamic approach of estimating ratios of normalizing constants.

Finally, attention was focused on the most popular implementation of thermodynamic integration in Bayesian statistics: the estimation of the Bayesian marginal likelihood and the Bayes factors. An alternative thermodynamic approachwas presented, based on the stepping-stone identity introduced in biology by Xie et al. (2011) and Fan et al. (2011). By this way, we were able to present in parallel the available in the literature estimators under the two different approaches (thermodynamic and stepping-stone) and further introduce new appropriate estimators (based on equivalent paths) filling in the blanks in the list of the Bayesian marginal likelihood and Bayes factors estimators. The quadrivial Bayes factor estimators were also introduced, which are based on nested, more complex, paths which seem to perform efficiently when estimating directly Bayes factors instead of Bayesian marginal likelihoods.

The unified framework in thermodynamic integration presented in this article offers new highways for research and further investigation. Here we discussed only some of the possible future research directions.

The first one is the identification of a possible link between the deviance information criterion, DIC, (Spiegelhalter et al., 2002) and thermodynamic integration. It is well-known that in mixture models there are problems in estimating the number of efficient parameters. A possible connection between TI and DIC may offer alternative ways of estimating it in cases with multimodal posterior densities. The connection between TI and KL as well as the connection between AIC, DIC and KL leave promises that such a connection can be achieved.

A second research direction is the development of a stochastic TI approach where the temperature will be treated as a unknown parameter. In this case, a suitable prior should be elicitated in order to a-priori support points where higher uncertainty of $\widehat{\mathcal{E}}_t$ is located.

Such a stochastic approach will eliminate the discretisation error which is an important source of variability for TI estimators.

Finally, MCMC samplers used for Bayesian variable selection is another interesting area of implementation of the TI approach. In such cases, interest may lie on the estimation of the normalizing constants over the whole model space and the direct estimation of posterior inclusion probabilities of each covariate. This might be extremely useful in large spaces with high number of covariates where the full exploration of the model space is infeasible due to its size and due to the existence of multiple neighborhoods of local maxima placed around well-fitted models.

Chapter 6

Implementation in simulated and real life datasets

"Figures don't lie, but liars do figure"

Mark Twain [*]

[*] Mark Twain or Samuel Clemens (1835–1910) was an American author (and humorist), best known by his books *The Adventures of Tom Sawyer* (1876) and *Adventures of Huckleberry Finn* (1885).

6.1 Introduction

In this chapter, the Bayesian marginal likelihood estimators discussed previously in this thesis are illustrated in simulated and real datasets. Emphasis is given on the estimation of the Bayesian marginal likelihood and the computation of the Bayes factor, as means of comparing models with different number of latent variables. In particular, Section 6.2 briefly summarizes the points made in the previous chapters with regard to the efficient estimation of the Bayesian marginal likelihood in GLLTM models. The simulation scheme is fully described and method-specific details are given when necessary. In Section 6.3, the Bayesian marginal likelihood is estimated in an IRT (2-PL) model. In Section 6.4, latent trait models with binary data are considered in simulated (Section 6.4.1) and real life (Section 6.4.2) examples and the Bayes factors between competing models are computed. The results are discussed in Section 6.5 where the estimators are compared in terms of efficiency and computational expense.

6.2 Simulation scheme

In this section, the steps followed through out the simulation procedure are described and the findings of previous chapters are summarised. To begin with, the *prior specification* in all examples was held in accordance with Section 2.2 which is summarised as follows:

$$\pi(\boldsymbol{\vartheta}, \mathbf{Z}) = \prod_{i=1}^{N}\prod_{\ell=1}^{k} \pi(Z_{i\ell}) \times \prod_{j=1}^{p} \pi(\alpha_j) \times \prod_{j=1}^{p}\prod_{\ell=1}^{k} \pi(\beta_{j\ell}),$$

where

$$\pi(\cdot) = \begin{cases} N(0,1) & \text{for all } Z_{i\ell} \\ LN(0,1) & \text{for } \beta_{j\ell} \text{ where } j = \ell, \\ N(0,4) & \text{for } \beta_{j\ell} \text{ where } j > \ell \text{ and for all } \alpha_j. \end{cases}$$

In order to sample from the *posterior*, the Metropolis-within-Gibbs algorithm, described in Section 2.3, was employed in all cases. For each simulated example, 300,000 posterior observations were generated after discarding additional 10,000 iterations as a burn in period. A thinning interval of 10 iterations was considered in order to diminish autocorrelations, leaving a total of 30,000 values available for posterior analysis. The posterior output was divided into 30 batches (with equal sizes R_b=1,000 points) and all estimators were computed within each batch. The mean over all batches, often referred to as the *batch mean* (Schmeiser, 1982, Bratley et al., 1987), was used as the final marginal likelihood estimate, denoted hereafter by $\log \widehat{f(\mathbf{Y})}$. In order the various estimators to be directly comparable in terms of error, the batch means method was implemented in all cases. In particular, the standard deviation of the log-marginal likelihood estimates over the 30 batches was considered as the estimated error of the $\log \widehat{f(\mathbf{Y})}$, denoted by $\widehat{MCE}_f$. The same procedure was used for the estimated values of the Bayes factor (denoted by $\log \widehat{BF}_{12}$ with regard to the competing models m_1 and m_2) and its associated error $\widehat{MCE}$.

The BSE and PBE (see Section 2.4.1) identities presented in this chapter are summarised in Tables 6.1 and 6.2 respectively. All estimators were implemented on the posterior $f(\boldsymbol{\vartheta}|\,\mathbf{Y})$. That is, in accordance with the findings presented in Chapter 3, the vector of the latent variables $\mathbf{Z}$ was marginalized out, according to (3.4), before applying each of the identities presented in Section 2.4. The marginalization was held via fixed Gauss-Hermite quadrature points (see also Rabe-Hesketh et al. 2005, Schilling and Bock 2005), in order to reduce the computational burden using fairly precise approximations.

Method-specific strategies were also followed in order to increase the efficiency of the corresponding estimators. With regard to the point-based estimators LM, GC and CJ^I, the componentwise posterior median was used as $\boldsymbol{\vartheta}^*$, based on the findings presented in Chapter 4. Additionally, the normal approximations used in the ML and GC methods were applied to the original parameters for all α_j and $\beta_{j\ell}$, with $j < \ell$ and on the $\log \beta_{jj}$. The log scale was preferred for the diagonal elements in order to avoid the asymmetries introduced by the positivity constraint and, by this way, to achieve a well behaved approximation of the marginal likelihood.

Finally, the output sampled from the posterior $f(\boldsymbol{\vartheta}|\mathbf{Y})$ was also used in order to construct the *importance* distribution $g(\boldsymbol{\vartheta})$, involved in the computation of the RM, BH and BG estimators. In each case, an approximation was constructed based on the posterior moments the parameters, with structure $g(\boldsymbol{\vartheta}) = g(\boldsymbol{\alpha})g(\boldsymbol{\beta}_e)$, where

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

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.

All simulations that are presented in this chapter were held on a quad core `i5 Central Processor Unit` (CPU), at `3.2GHz` and with `4GB` of `RAM`. The Metropolis-within-GIbbs algorithm (see Section for 2.3 details) was conducted using a custom routine in `R` (`version 3.0.1`; R Core Team, 2013). The time required in order to extract the posterior outputs was approximately 1 minute per 1000 iterations for the one factor models, 3 minutes per 1000 iterations for the two factor models and 5 minutes per 1000 iterations the three factor models. Custom routines written in `R` language were used for the bridge sampling and for the point-based estimators.

Table 6.1: Bridge sampling Bayesian marginal likelihood identities

Method	BML identity	Bridge function $\mathcal{A}(\boldsymbol{\vartheta})$
Arithmetic mean	$\int f(\mathbf{Y}\|\,\boldsymbol{\vartheta})\,\pi(\boldsymbol{\vartheta})\,d\boldsymbol{\vartheta}$	$g(\boldsymbol{\vartheta})$
Harmonic mean	$\left\{\int \frac{1}{f(\mathbf{Y}\|\,\boldsymbol{\vartheta})} f(\boldsymbol{\vartheta}\|\mathbf{Y})\,d\boldsymbol{\vartheta}\right\}^{-1}$	$g(\boldsymbol{\vartheta})/f(\mathbf{Y}\|\,\boldsymbol{\vartheta})$
Reciprocal mean	$\left\{\int \frac{g(\boldsymbol{\vartheta})}{f(\mathbf{Y}\|\,\boldsymbol{\vartheta})\,\pi(\boldsymbol{\vartheta})} f(\boldsymbol{\vartheta}\|\mathbf{Y})\,d\boldsymbol{\vartheta}\right\}^{-1}$	$\{f(\mathbf{Y}\|\,\boldsymbol{\vartheta})\,\pi(\boldsymbol{\vartheta})\}^{-1}$
Bridge harmonic	$\frac{\int g(\boldsymbol{\vartheta})^{-1} g(\boldsymbol{\vartheta}) d\boldsymbol{\vartheta}}{\int \{f(\mathbf{Y}\|\,\boldsymbol{\vartheta})\,\pi(\boldsymbol{\vartheta})\}^{-1} f(\boldsymbol{\vartheta}\|\mathbf{Y})\,d\boldsymbol{\vartheta}}$	$\{f(\mathbf{Y}\|\,\boldsymbol{\vartheta})\,\pi(\boldsymbol{\vartheta})\,g(\boldsymbol{\vartheta})\}^{-1}$
Bridge geometric	$\frac{\int \{f(\mathbf{Y}\|\,\boldsymbol{\vartheta})\,\pi(\boldsymbol{\vartheta})/g(\boldsymbol{\vartheta})\}^{1/2} g(\boldsymbol{\vartheta})\,d\boldsymbol{\vartheta}}{\int \{f(\mathbf{Y}\|\,\boldsymbol{\vartheta})\,\pi(\boldsymbol{\vartheta})/g(\boldsymbol{\vartheta})\}^{-1/2} f(\boldsymbol{\vartheta}\|\mathbf{Y})\,d\boldsymbol{\vartheta}}.$	$\{f(\mathbf{Y}\|\,\boldsymbol{\vartheta})\,\pi(\boldsymbol{\vartheta})\,g(\boldsymbol{\vartheta})\}^{-1/2}$

Table 6.2: Point-based Bayesian marginal likelihood identities.

Method	BML identity (log scale)
Laplace Metropolis	$\frac{p}{2}\log\{2\pi\} + \frac{1}{2}\log\|\,\mathbf{H}^*\| + \log\pi(\boldsymbol{\vartheta}^*) + \log f(\mathbf{Y}\|\,\boldsymbol{\vartheta}^*)$
Gaussian copula	$-\frac{1}{2}\log\|\,\boldsymbol{\Gamma}\| + \sum_{j=1}^{p}\log f_j(\boldsymbol{\vartheta}_j^*) + \log\pi(\boldsymbol{\vartheta}^*) + \log f(\mathbf{Y}\|\,\boldsymbol{\vartheta}^*)$
Independence CJ	$-\log\left\{E_{\boldsymbol{\vartheta},\,\mathbf{Z}\|\mathbf{Y}}\left[\frac{\prod_{j=1}^{p} a\big(\boldsymbol{\vartheta}_j, \boldsymbol{\vartheta}_j^*\|\mathbf{Y},\mathbf{Z}\big)\,q\big(\boldsymbol{\vartheta}_j, \boldsymbol{\vartheta}_j^*\|\mathbf{Y},\mathbf{Z}\big)}{\prod_{j=1}^{p} E_{q_j}\left[a\big(\boldsymbol{\vartheta}_j^*, \boldsymbol{\vartheta}_j\|\mathbf{Y},\mathbf{Z}\big)\right]}\right]\right\}$ $+\log\pi(\boldsymbol{\vartheta}^*) + \log f(\mathbf{Y}\|\,\boldsymbol{\vartheta}^*)$

6.3 Marginal likelihood estimation in an IRT model

The first dataset considered here was generated from a one-factor model with 4 binary items and 400 individuals. Hence, it is classified under the category of the 2-PL IRT models (see for instance Patz and Junker, 1999b). The dataset has been presented also in Chapter 4, in order to evaluate the CJ^I estimator.

The estimated values that correspond to each method are presented in Table 6.3. The three point-based estimators (GC, LM and CJ^I) yielded similar values for the Bayesian marginal likelihood with fairly small errors (0.12-0.16). The smallest error however occurred in the case of the BG estimator (0.05). With regard the other bridge estimators, the RM estimator yielded also close values, with slightly higher error (0.19). On the contrary, the BH estimator was associated with more than 8 times higher error (1.64) and the estimated BML was lower by one unit, in log scale. Finally, the HM overestimated the BML by 12 units in log scale. The estimated values over the 30 batches are presented graphically in Figures 6.1 and 6.2, for the point-based and bridge families respectively.

In Table 6.3, it is shown also that the $\log \widehat{f(\mathbf{Y})}$ value obtained by the AM estimator, far exceeded the ones provided by the other estimators, with very high associated $M\widehat{C}E_f$. This result was expected yet the AM was included in this example in order to facilitate the understanding of the path sampling estimators. In particular, at $t = 0$ the PP_S estimator coincides with the AM while the PP_T implements also points drawn solely from the prior. The estimated values (-995.4 and -995.5 respectively) lay in between the AM value and the values obtained by the rest of the estimators (Table 6.3). It occurs that the divergencies between the successive sampling distributions towards the zero end were not adequately reduced, under the current sampling scheme (see Section 6.2 for details). On the contrary, the two path sampling estimators that implement the importance function $g(\boldsymbol{\theta})$ rather than the prior, yielded well acceptable results (see also Figure 6.3). However, the time elapsed for the computation of the path sampling estimators was approximately 15 hours. This is due to the fact that the time-consuming MG algorithm was implemented n times instead of one, that is, one for each temperature included in the schedule. Hence, in the following sections where higher dimensional models that are presented, the path sampling estimators were discarded along with the AM estimator.

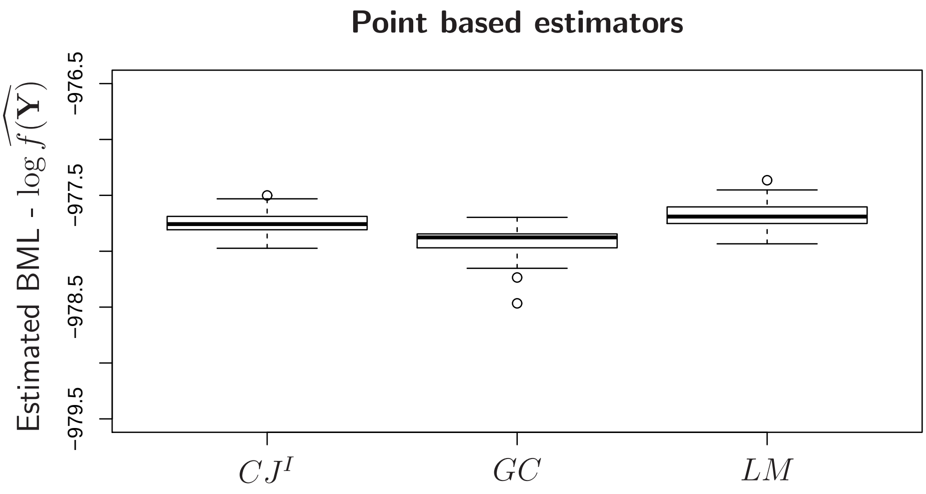


Figure 6.1: BML estimates (per batch): Point-based estimators ($p = 4$, $N = 400$, $k = 1$).

Point-based estimators: Gaussian copula (GC), independence Chib & Jeliazkov (CJ^I) and Laplace Metropolis (LM).

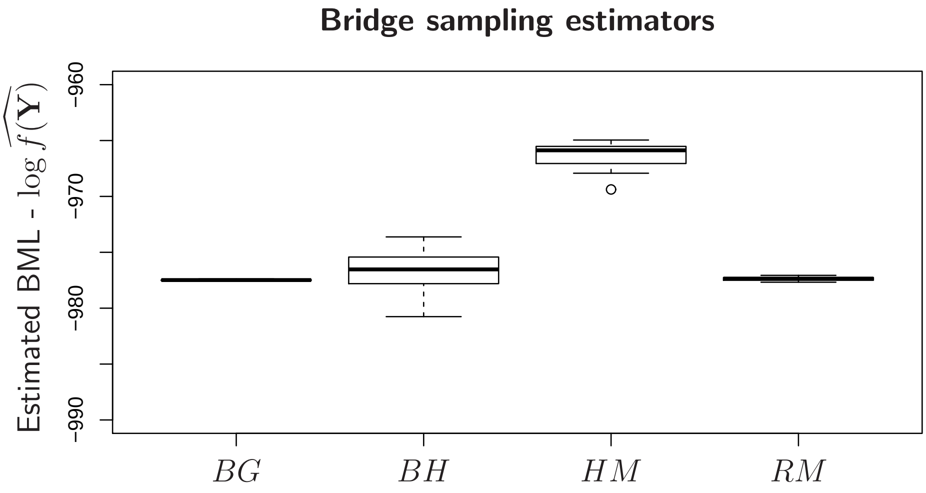


Figure 6.2: BML estimates (per batch): Bridge sampling estimators ($p = 4$, $N = 400$, $k = 1$).

Bridge sampling estimators: harmonic mean (HM), reciprocal mean (RM), bridge harmonic (BH) and geometric (BG).

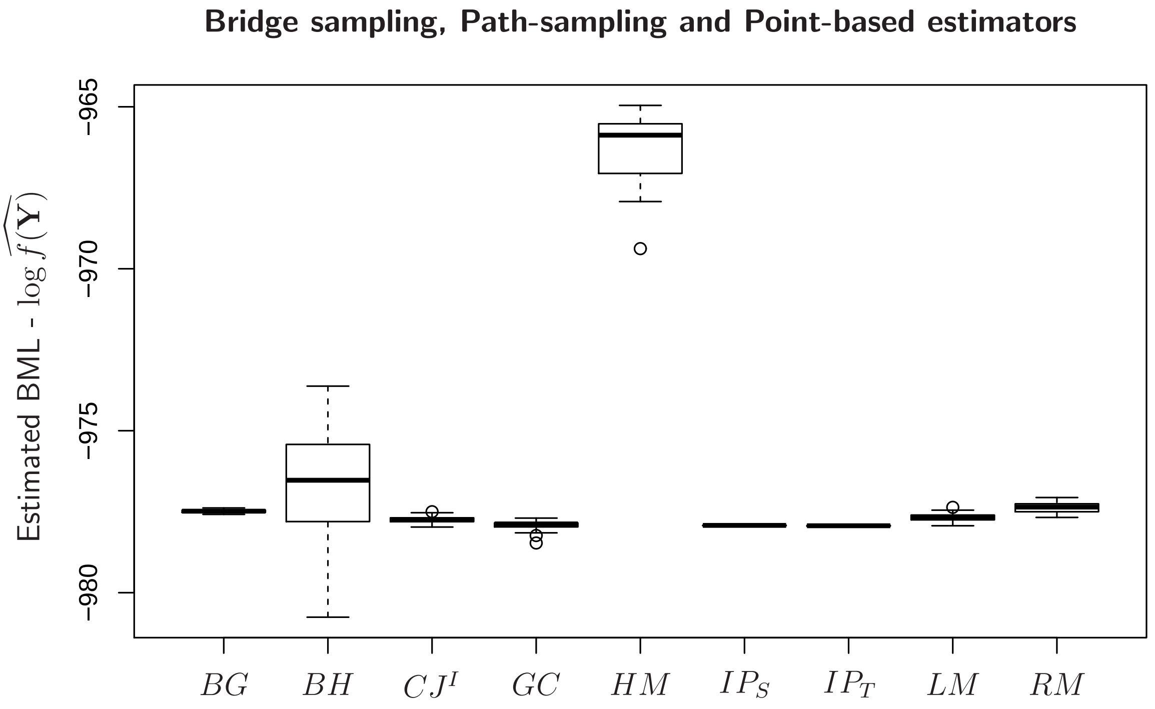


Figure 6.3: BML estimates (per batch): BSE, PBE and PSE ($p = 4$, $N = 400$, $k = 1$).

Bridge sampling estimators: harmonic mean (HM), reciprocal mean (RM), bridge harmonic (BH) and bridge geometric (BG). *Point-based* estimators: Gaussian copula (GC), independence Chib & Jeliazkov (CJ^I) and Laplace Metropolis (LM). *Path sampling* estimators: thermodynamic and stepping-stone power posteriors (PP_T and PP_S respectively), thermodynamic and stepping-stone importance posteriors (IP_T and IP_S respectively).

Table 6.3: Marginal likelihood estimates for an IRT model ($N = 400$, $p = 4$ & $k = 1$)

Method	Estimator	$\log \widehat{f(\mathbf{Y})}$	$M\widehat{C}E_f$
Bridge sampling	AM	-1030.1	39.43
	HM	-965.9	0.90
	BH	-976.5	1.64
	RM	-977.4	0.19
	BG	-977.5	0.05
Point based	LM	-977.7	0.13
	CJ^I	-977.8	0.12
	GC	-977.9	0.16
Path sampling	PP_T	-995.4	0.02
	PP_S	-995.5	0.03
	IP_T	-977.9	0.02
	IP_S	-977.9	0.01

Bridge sampling estimators: arithmetic mean (AM), harmonic mean (HM), reciprocal mean (RM), bridge harmonic (BH) and bridge geometric (BG). *Point-based* estimators: Gaussian copula (GC), independence Chib & Jeliazkov (CJ^I) and Laplace Metropolis (LM). *Path sampling* estimators: thermodynamic and stepping-stone power posteriors (PP_T and PP_S respectively), thermodynamic and stepping-stone importance posteriors (IP_T and IP_S respectively).

6.4 Implementation in latent trait models with binary data

In this section, simulated and real data sets of larger size are implemented that allow for fitted models with higher dimensions. Hence, the Bayes factor can be computed in order to evaluate the competing models.

6.4.1 Simulated data examples

Three simulated datasets are considered here, namely:

Dataset A: $N = 600$ observations with $p = 6$ items generated from a $k = 1$ model.

Dataset B: $N = 600$ observations with $p = 6$ items generated from a $k = 2$ model.

Dataset C: $N = 800$ observations with $p = 7$ items generated from a $k = 3$ model.

All model parameters were selected randomly from a uniform distribution, $U(-2, 2)$. The number of unknown parameters is equal to $k(p+N)+p$, corresponding to 606, 1218 and 2428 parameters, respectively, for each of the three situations described above.

6.4.1.1 Dataset A ($N = 600,\ p = 6\ \&\ k = 1$)

Dataset A was generated from a single factor model. In Table 6.4 are presented the estimated values for the BML, under the true model ($k = 1$) and under a model that overestimates the number of latent variables ($k = 2$). As in the case of the IRT model, the HM estimator overestimated the BML, in both models. With regard to the other estimators, their values were fairly close in the case of the true model (Figure 6.4), but discrepancies were observed when two latent variables were assumed (Figure 6.5).

Table 6.4: Marginal likelihood estimates for *Dataset A* ($N = 600$, $p = 6$ & $k = 1$)

	$k=1$		$k=2$	
Estimator	$\log \widehat{f(\mathbf{Y})}$	$M\widehat{C}E_f$	$\log \widehat{f(\mathbf{Y})}$	$M\widehat{C}E_f$
HM	-2153.4	1.14	-2154.0	1.78
RM	-2174.7	0.11	-2175.5	0.81
BH	-2174.0	1.83	-2170.0	2.37
BG	-2174.7	0.03	-2177.2	0.29
LM	-2175.2	0.16	-2178.7	0.86
CJ^I	-2175.1	0.10	-2178.2	1.39
GC	-2175.3	0.22	-2180.3	0.47

Bridge sampling estimators: harmonic mean (HM), reciprocal mean (RM), bridge harmonic (BH) and bridge geometric (BG). *Point-based* estimators: Gaussian copula (GC), independence Chib & Jeliazkov (CJ^I) and Laplace Metropolis (LM).

6.4.1.2 Dataset B ($N = 600,\ p = 6\ \&\ k = 2$)

The second dataset (Dataset B) was generated from a two-factor model. Table 6.5 presents the estimated Bayesian marginal likelihoods under the true model ($k = 2$) and

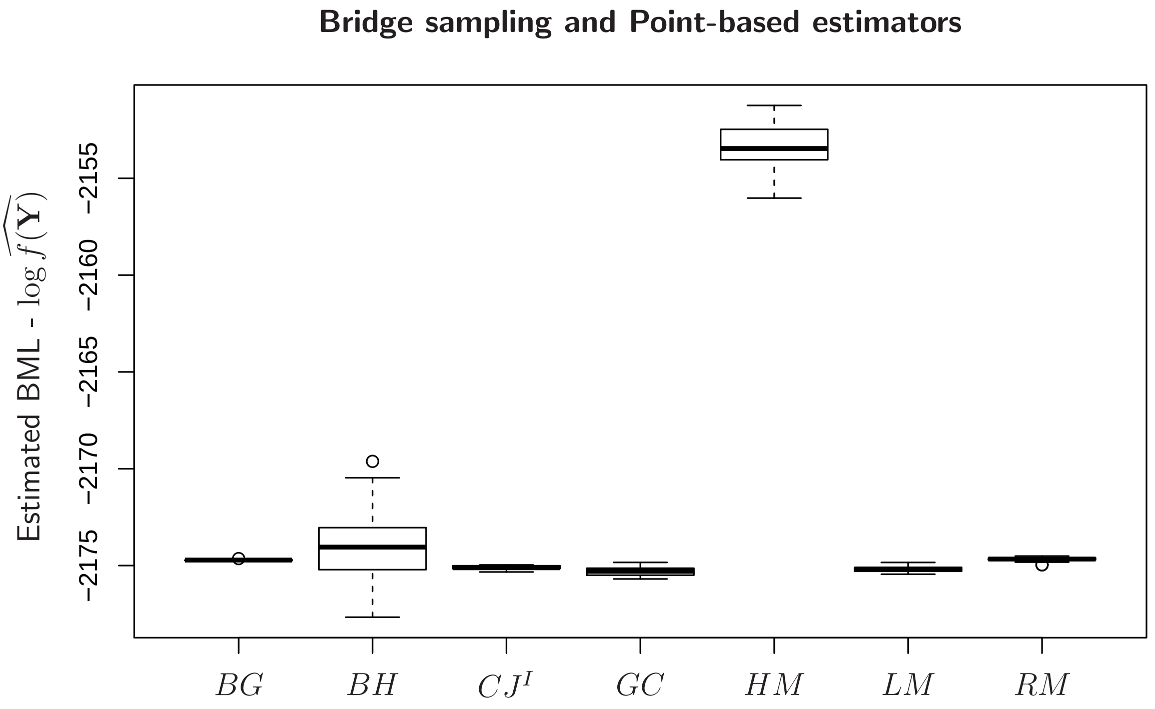


Figure 6.4: Dataset A: Bridge sampling and Point-based BML estimators ($p = 6$, $N = 600$, $k_{true} = 1$ $/k_{model} = 1$).

Bridge sampling estimators: harmonic mean (HM), reciprocal mean (RM), bridge harmonic (BH) and geometric (BG). *Point-based* estimators: Gaussian copula (GC), independence Chib & Jeliazkov (CJ^I) and Laplace Metropolis (LM).

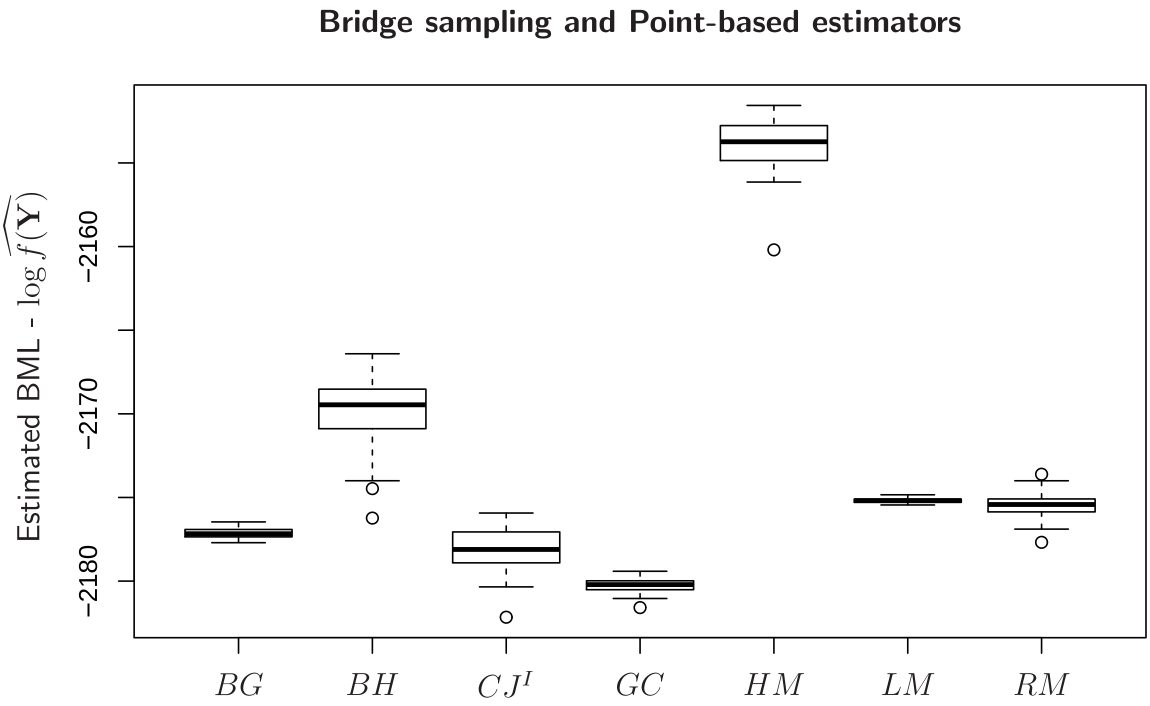


Figure 6.5: Dataset A: Bridge sampling and Point-based BML estimators ($p = 6$, $N = 600$, $k_{true} = 1$ $/k_{model} = 2$).

Bridge sampling estimators: harmonic mean (HM), reciprocal mean (RM), bridge harmonic (BH) and geometric (BG). *Point-based* estimators: Gaussian copula (GC), independence Chib & Jeliazkov (CJ^I) and Laplace Metropolis (LM).

under a model which now underestimates the number of latent variables ($k = 1$). Hence, the scenario of the previous example has been reversed here. As previously, under the bivariate (in terms of latent factors) true model, all estimators yielded similar results (Figure 6.7), with the exception of the HA estimator. Under the hypothesised univariate model, the estimated valuees were similar within each family of estimators (BSE and PBE) but not between the families. That is, the estimated values derived via the PBE were smaller by approximately 4 units in log scale (see also Figure 6.6).

Table 6.5: Marginal likelihood estimates for *Dataset B* ($N = 600$, $p = 6$ & $k = 2$)

	$k = 1$		$k = 2$	
Estimator	$\log \widehat{f(\mathbf{Y})}$	$M\widehat{C}E_f$	$\log \widehat{f(\mathbf{Y})}$	$M\widehat{C}E_f$
HM	-2161.5	1.26	-2043.9	2.14
BH	-2182.5	2.48	-2070.2	1.99
RM	-2183.7	0.23	-2071.2	0.39
BG	-2183.9	0.06	-2071.5	0.10
LM	-2187.3	0.16	-2071.2	0.27
CJ^I	-2187.5	0.18	-2071.2	0.36
GC	-2187.5	0.21	-2071.6	0.29

Bridge sampling estimators: harmonic mean (HM), reciprocal mean (RM), bridge harmonic (BH) and bridge geometric (BG). *Point-based* estimators: Gaussian copula (GC), independence Chib & Jeliazkov (CJ^I) and Laplace Metropolis (LM).

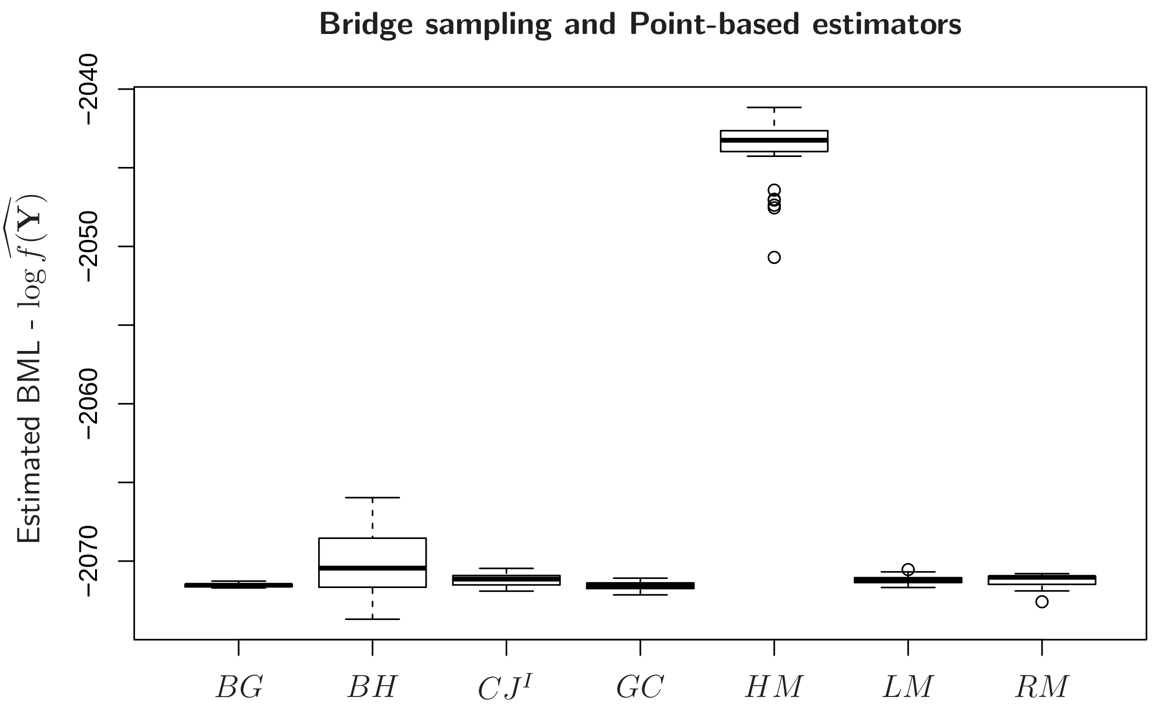


Figure 6.6: Dataset B: Bridge sampling and Point-based BML estimators ($p = 6$, $N = 600$, $k_{true} = 2$ $/k_{model} = 2$).

Bridge sampling estimators: harmonic mean (HM), reciprocal mean (RM), bridge harmonic (BH) and geometric (BG). *Point-based* estimators: Gaussian copula (GC), independence Chib & Jeliazkov (CJ^I) and Laplace Metropolis (LM).

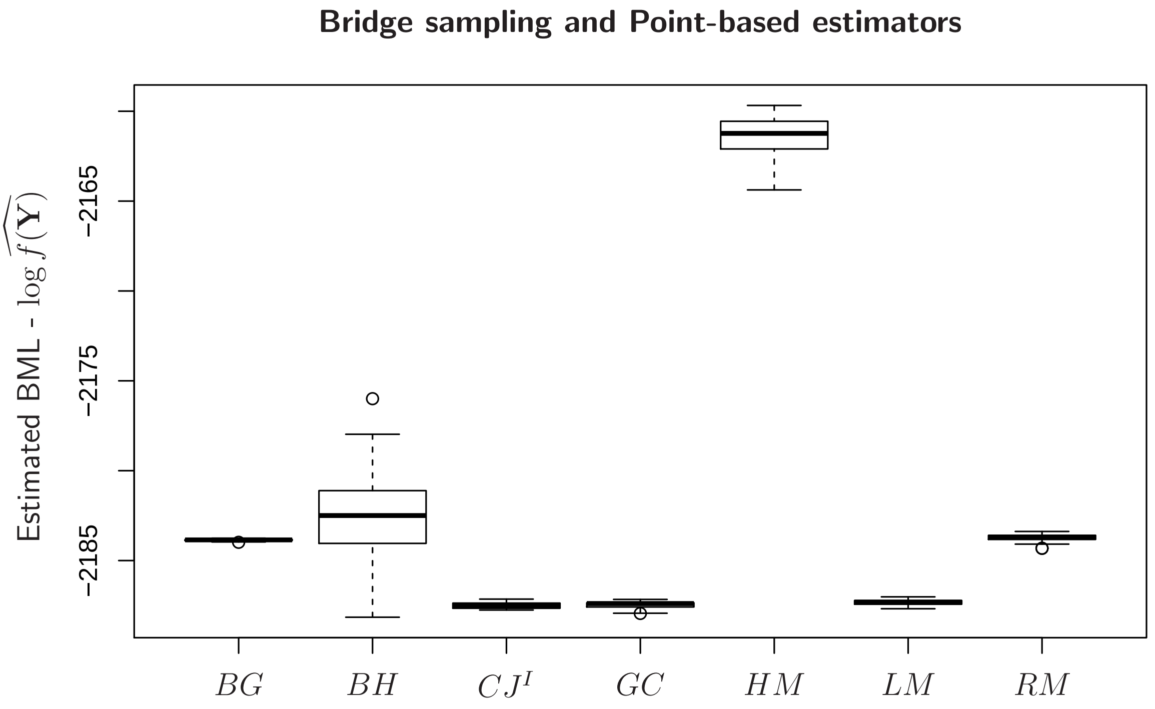


Figure 6.7: Dataset B: Bridge sampling and Point-based BML estimators ($p = 6$, $N = 600$, $k_{true} = 2$ $/k_{model} = 1$).

Bridge sampling estimators: harmonic mean (HM), reciprocal mean (RM), bridge harmonic (BH) and geometric (BG). *Point-based* estimators: Gaussian copula (GC), independence Chib & Jeliazkov (CJ^I) and Laplace Metropolis (LM).

6.4.1.3 Data set C ($N = 800,\ p = 7$ & $k = 3$)

The third dataset corresponds to a high dimensional model generated form a three-factor model. In Table 6.6 are presented the estimated marginal likelihoods under the true model ($k = 3$) and under two models that overestimate the number of latent variables ($k = 1$ and $k = 2$). The estimated BML values were fairly close in all cases, despite the high dimensionality of the model (see also Figures 6.8 to 6.10).

Table 6.6: Marginal likelihood estimates for *Dataset C* ($N = 800$, $p = 7$ & $k = 3$)

	$k = 1$		$k = 2$		$k = 3$	
Estimator	$\log \widehat{f(\mathbf{Y})}$	$M\widehat{C}E_f$	$\log \widehat{f(\mathbf{Y})}$	$M\widehat{C}E_f$	$\log \widehat{f(\mathbf{Y})}$	$M\widehat{C}E_f$
HM	-3395.1	1.38	-3338.8	2.14	-3302.6	2.10
BH	-3420.9	1.74	-3373.1	2.37	-3337.4	2.83
RM	-3421.6	0.19	-3375.3	0.42	-3341.5	0.61
BG	-3421.7	0.05	-3376.3	0.16	-3343.3	0.27
LM	-3422.4	0.18	-3374.7	0.29	-3341.3	0.35
CJ^I	-3422.5	0.16	-3375.2	0.69	-3339.3	1.82
GC	-3422.7	0.22	-3375.5	0.23	-3343.1	0.30

Bridge sampling estimators: harmonic mean (HM), reciprocal mean (RM), bridge harmonic (BH) and bridge geometric (BG). *Point-based* estimators: Gaussian copula (GC), independence Chib & Jeliazkov (CJ^I) and Laplace Metropolis (LM).

6.4.1.4 Bayes factor estimates

The estimated Bayes factors which correspond to Datasets A to C (Tables 6.4 to 6.6) are presented in Table 6.7 and depicted in Figures 6.11 to 6.14. The discrepancies observed in the estimation of the BLM via the HM estimator, resulted in substantially different BF estimated values, that at least in one case indicated the wrong model. In particular, the HM-based BF that corresponds to Dataset A was -4 (MCE=2.6), strongly suggesting the 2-factor model (Kass and Raftery, 1995) instead of the true IRT model. The rest of the estimators, resulted to non-decisive BF values (0.5 - 0.8) or suggested the correct model (BF: 2.4 to 5.0). For the other two datasets, the correct model was strongly suggested by the estimated BF, in all cases.

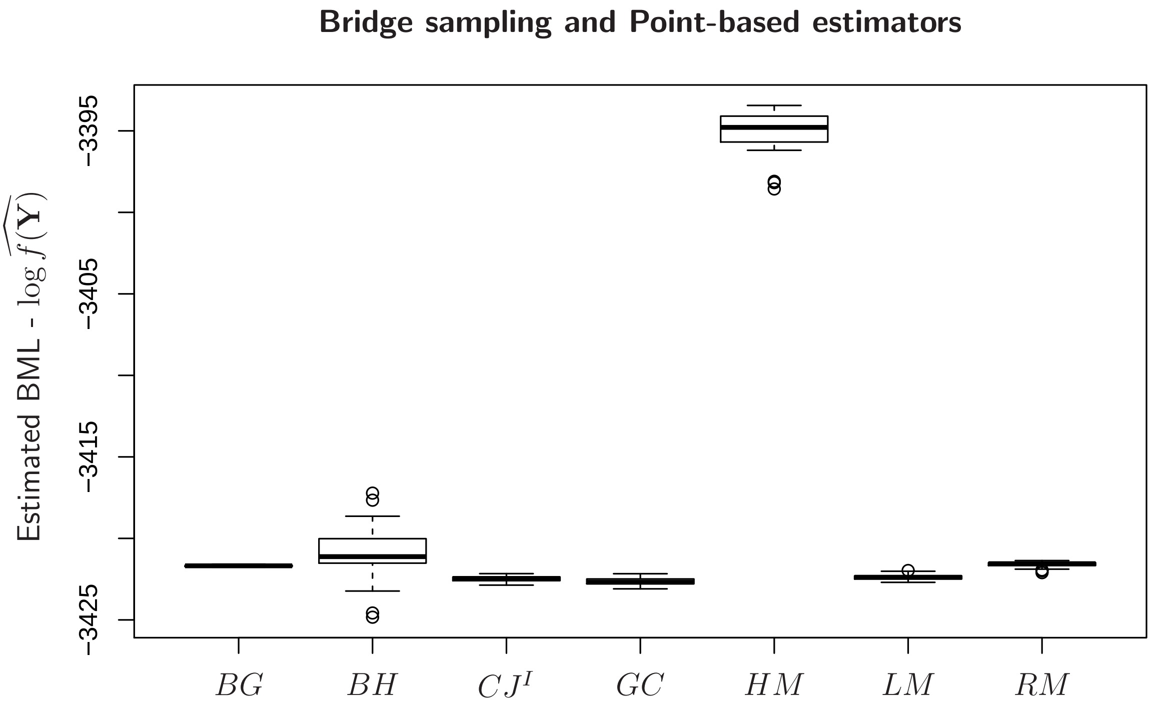


Figure 6.8: Dataset C: Bridge sampling and Point-based BML estimators ($p = 7$, $N = 800$, $k_{true} = 3$ $/k_{model} = 1$).

Bridge sampling estimators: harmonic mean (HM), reciprocal mean (RM), bridge harmonic (BH) and geometric (BG). *Point-based* estimators: Gaussian copula (GC), independence Chib & Jeliazkov (CJ^I) and Laplace Metropolis (LM).

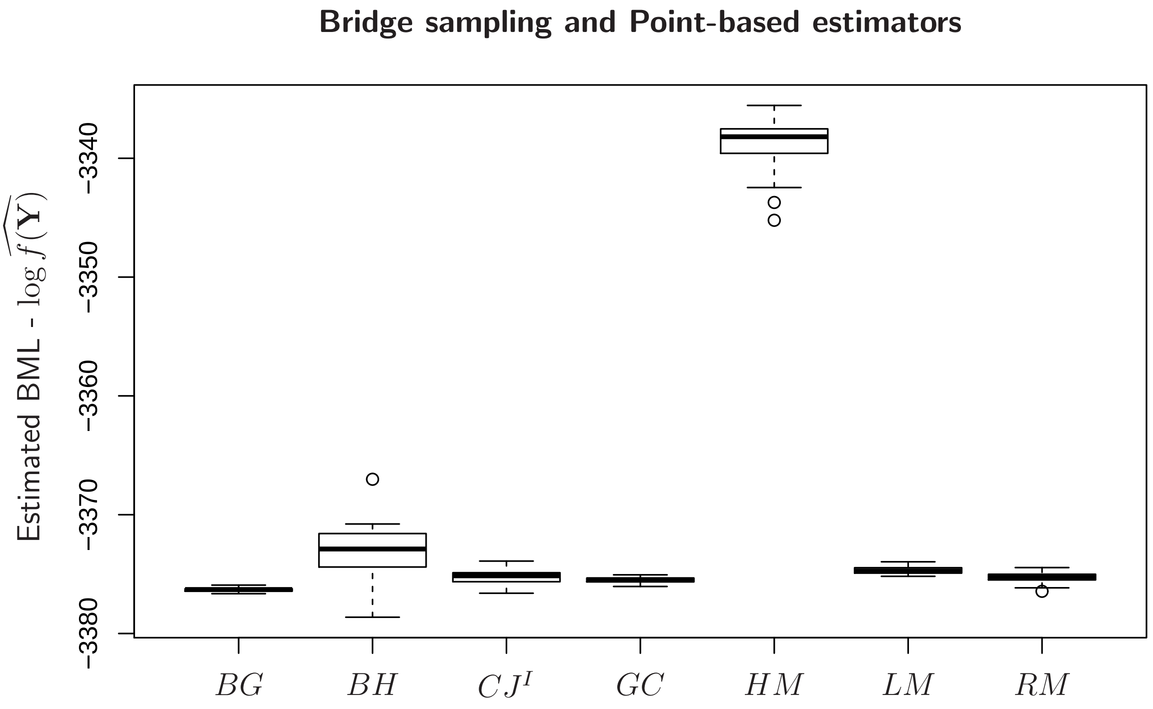


Figure 6.9: Dataset C: Bridge sampling and Point-based BML estimators ($p = 7$, $N = 800$, $k_{true} = 3$ $/k_{model} = 2$).

Bridge sampling estimators: harmonic mean (HM), reciprocal mean (RM), bridge harmonic (BH) and geometric (BG).
Point-based estimators: Gaussian copula (GC), independence Chib & Jeliazkov (CJ^I) and Laplace Metropolis (LM).

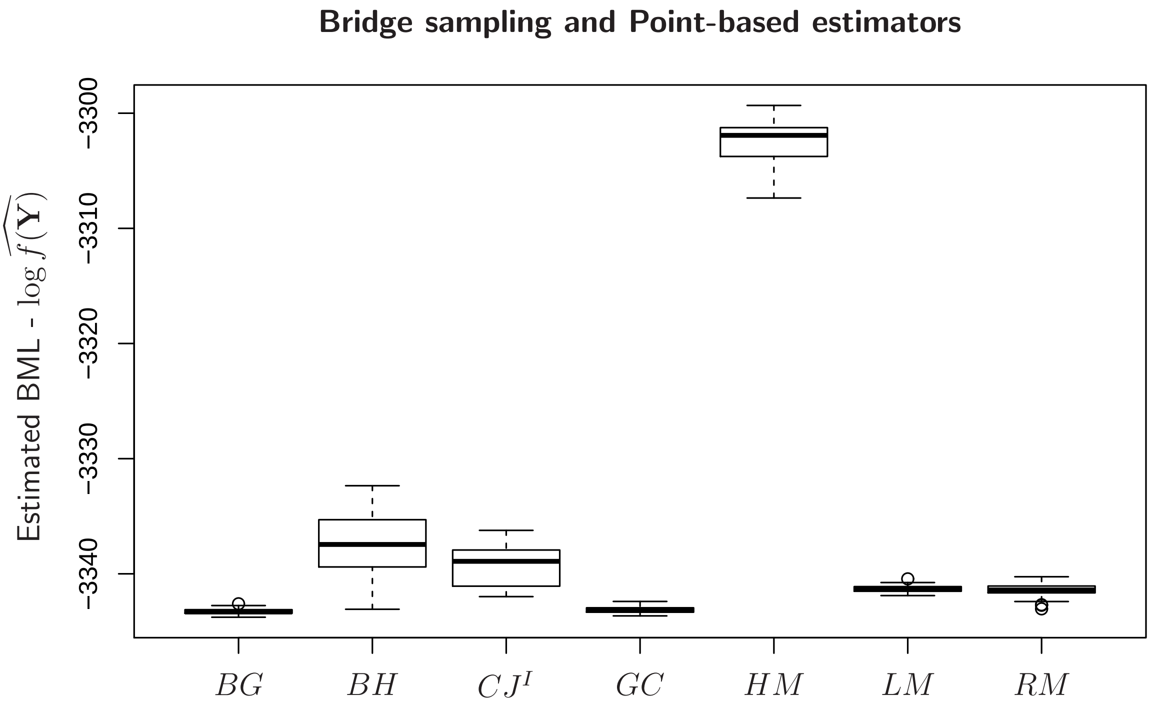


Figure 6.10: Dataset C: Bridge sampling and Point-based BML estimators ($p = 7$, $N = 800$, $k_{true} = 3$ $/k_{model} = 3$).

Bridge sampling estimators: harmonic mean (HM), reciprocal mean (RM), bridge harmonic (BH) and geometric (BG). *Point-based* estimators: Gaussian copula (GC), independence Chib & Jeliazkov (CJ^I) and Laplace Metropolis (LM).

Table 6.7: Estimated Bayes factors (log scale)

	Dataset A	Dataset B	Dataset C	
Estimator	$\log \widehat{BF}_{12}$ $(M\widehat{C}E)$	$\log \widehat{BF}_{21}$ $(M\widehat{C}E)$	$\log \widehat{BF}_{31}$ $(M\widehat{C}E)$	$\log \widehat{BF}_{32}$ $(M\widehat{C}E)$
HM	0.5 (2.0)	117.6 (2.5)	92.5 (2.7)	36.2 (3.1)
BH	-4.0 (2.6)	112.3 (3.1)	83.4 (3.4)	35.7 (3.8)
RM	0.8 (0.8)	112.5 (0.4)	80.1 (0.6)	33.8 (0.7)
BG	2.4 (0.3)	112.3 (0.1)	78.4 (0.3)	33.1 (0.3)
LM	3.5 (0.9)	116.1 (0.3)	81.1 (0.3)	33.4 (0.5)
CJ^I	3.1 (1.4)	116.3 (0.4)	83.2 (1.8)	35.9 (1.7)
GC	5.0 (0.5)	115.9 (0.4)	79.5 (0.4)	32.4 (0.4)

Bridge sampling estimators: harmonic mean (HM), reciprocal mean (RM), bridge harmonic (BH) and bridge geometric (BG). *Point-based* estimators: Gaussian copula (GC), independence Chib & Jeliazkov (CJ^I) and Laplace Metropolis (LM).

6.4.2 Applications in real data

We proceed with the two real-data examples, analyzed also in Chapter 4 and in (Bartholomew et al., 2008, chapter 8). In all examples the Bayesian marginal likelihood was estimated over samples of 10 thousand iterations (after discarding 1000 iterations as a burn in period and keeping 1 every 10 iterations to reduce autocorrelations).

6.4.2.1 Law School Admission Test (LSAT)

With regard to the the LSAT data (see Section 4.5.4 for details), most estimators did not support the 2-factor model, yielding BF estimators less than 2. Only exception was the GC estimator, which yielded $\log \widehat{BF}_{12} = -3$, strongly supporting the 1-factor model (Table 6.8).

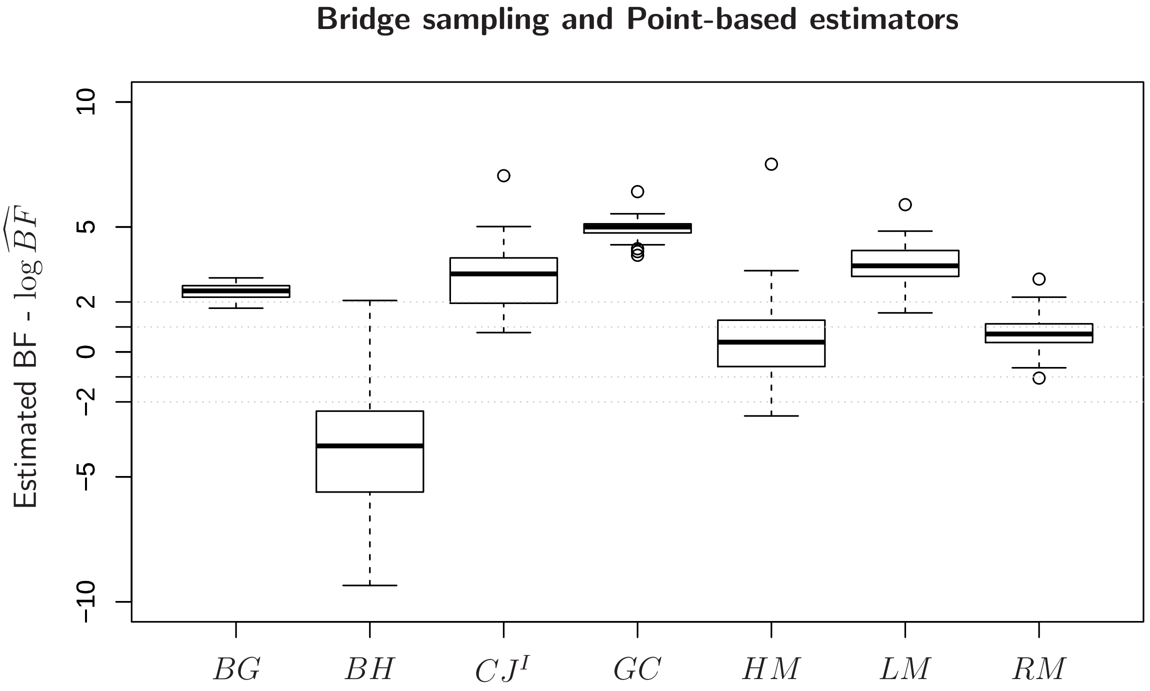


Figure 6.11: Dataset A: Bayes factor, all estimators ($p = 6$, $N = 600$, $k_{true} = 1$ vs $k = 2$).

Bridge sampling estimators: harmonic mean (HM), reciprocal mean (RM), bridge harmonic (BH) and geometric (BG). *Point-based* estimators: Gaussian copula (GC), independence Chib & Jeliazkov (CJ^I) and Laplace Metropolis (LM).

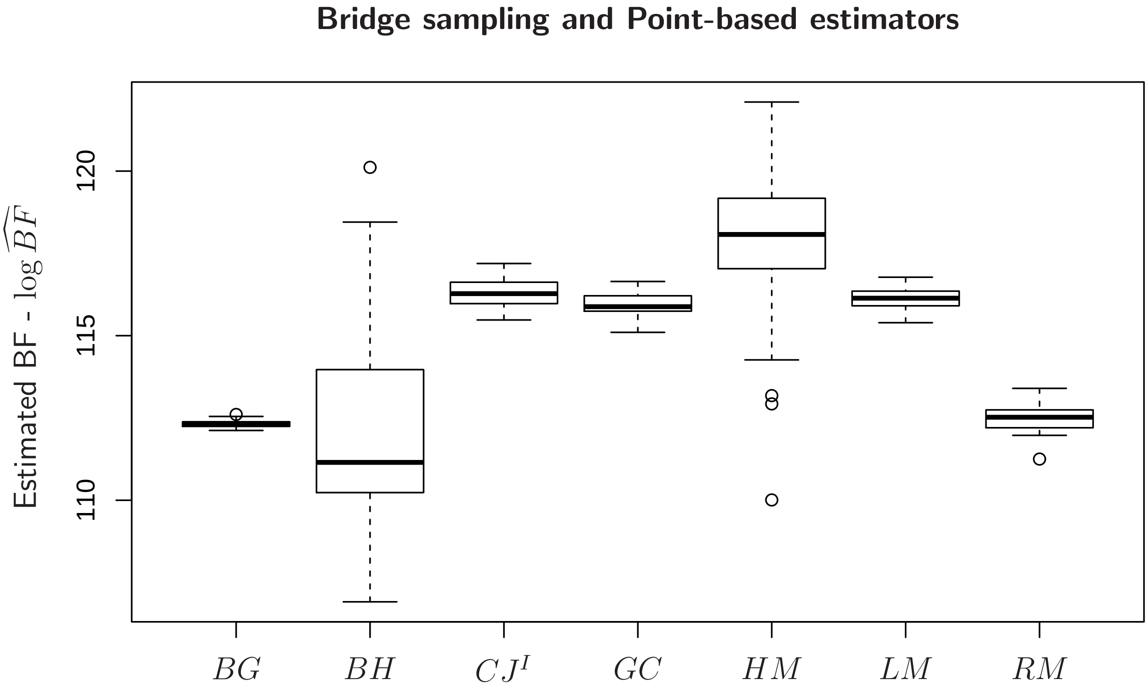


Figure 6.12: Dataset B: Bayes factor, all estimators ($p = 6$, $N = 600$, $k_{true} = 2$ vs $k = 1$).

Bridge sampling estimators: harmonic mean (HM), reciprocal mean (RM), bridge harmonic (BH) and geometric (BG). *Point-based* estimators: Gaussian copula (GC), independence Chib & Jeliazkov (CJ^I) and Laplace Metropolis (LM).

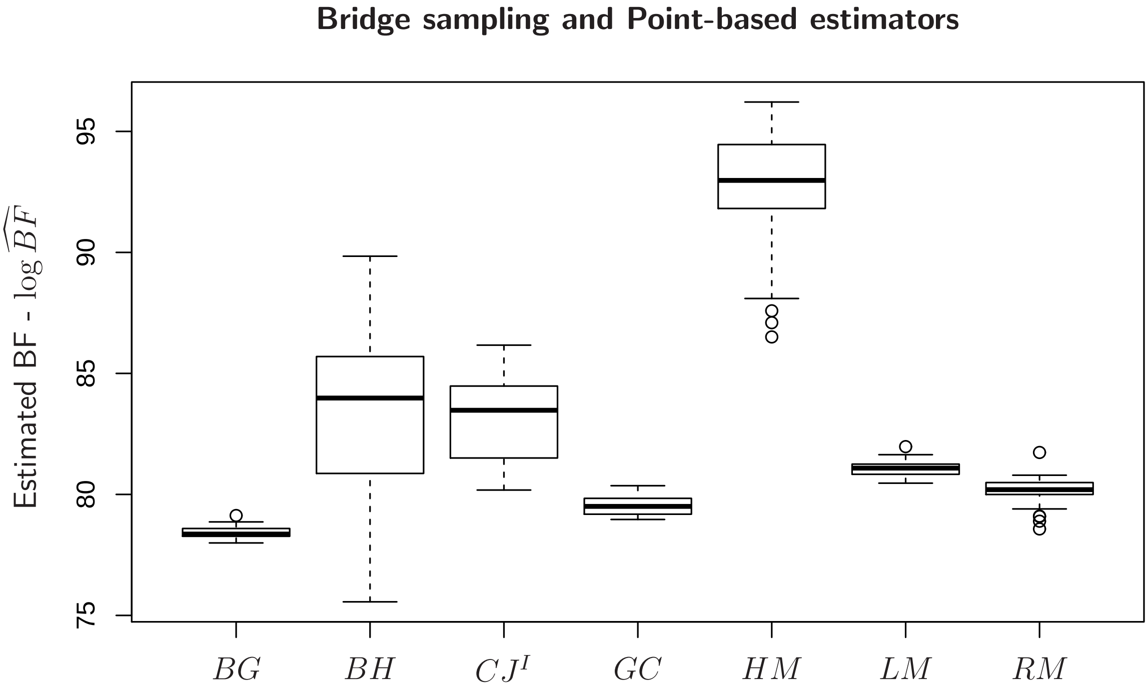


Figure 6.13: Dataset C: Bayes factor, all estimators ($p = 7$, $N = 800$, $k_{true} = 3$ vs $k = 1$).

Bridge sampling estimators: harmonic mean (HM), reciprocal mean (RM), bridge harmonic (BH) and geometric (BG). *Point-based* estimators: Gaussian copula (GC), independence Chib & Jeliazkov (CJ^I) and Laplace Metropolis (LM).

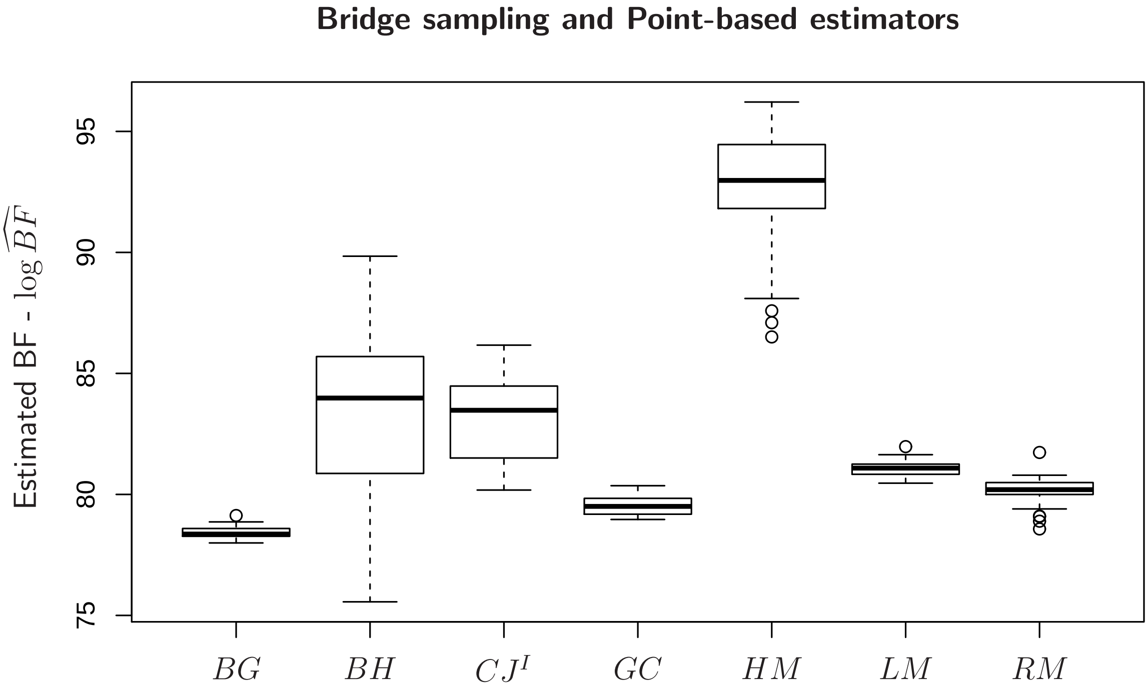


Figure 6.14: Dataset C: Bayes factor, all estimators ($p = 7$, $N = 800$, $k_{true} = 3$ vs $k = 2$).

Bridge sampling estimators: harmonic mean (HM), reciprocal mean (RM), bridge harmonic (BH) and geometric (BG). *Point-based* estimators: Gaussian copula (GC), independence Chib & Jeliazkov (CJ^I) and Laplace Metropolis (LM).

Table 6.8: Marginal likelihood estimates for $LSAT\ data$ ($N = 1000$, $p = 5$ & $k = 2$)

Estimator	$\log \widehat{f(\mathbf{Y})}_{k=1}$	$\log \widehat{f(\mathbf{Y})}_{k=2}$	$\log \widehat{BF}_{12}$
HM	-2476.5	-2476.4	0.1
BH	-2492.7	-2491.9	0.8
RM	-2494.4	-2495.9	-1.5
BG	-2494.4	-2496.3	-1.8
LM	-2494.9	-2496.8	-1.9
CJ^I	-2495.1	-2496.1	-1.0
GC	-2495.2	-2498.3	-3.2

Bridge sampling estimators: harmonic mean (HM), reciprocal mean (RM), bridge harmonic (BH) and bridge geometric (BG). *Point-based* estimators: Gaussian copula (GC), independence Chib & Jeliazkov (CJ^I) and Laplace Metropolis (LM).

6.4.2.2 Workplace Industrial Relations Survey (WIRS)

With regard to the the WIRS data (see Section 4.5.4 for details), all estimators yielded ”decisive evidence” against the one-factor model, for the 5-item and 6-item scales (Table 6.9).

6.5 Discussion and concluding remarks

The examples presented in Section 6.3 to 6.4.2 indicate that there are repeatable patterns with regard to the estimated BML values and the associated MCE based on the BSE and PBE methods. In particular, the bridge geometric estimator is associated with the smallest error almost in all examples while the estimated BML value was well comparable with the ones derived by the majority of the methods. On the contrary, the harmonic mean estimator overestimated the likelihood in all examples. However, the largest MCE was present in the case of the bridge harmonic estimator. In fact the MCE associated with the BH was at least two times as large as the errors of the other estimators. However, unlike the HM, the estimated BML value was well comparable with the rest of the estimators. Finally, the RM estimator yielded BLM values close to the ones derived by BG, with well comparable errors. It occurs that the existence of the importance (reference) function $g(\boldsymbol{\vartheta})$ stabilises the BML estimators, even for high dimensional models

Table 6.9: Marginal likelihood estimates for *WIRS data* ($N = 1005$, $p = 5(6)$ & $k = 2$)

	Five items			Six items		
Estimator	$\log \widehat{f(\mathbf{Y})}_{k=1}$	$\log \widehat{f(\mathbf{Y})}_{k=2}$	$\log \widehat{BF}_{21}$	$\log \widehat{f(\mathbf{Y})}_{k=1}$	$\log \widehat{f(\mathbf{Y})}_{k=2}$	$\log \widehat{BF}_{21}$
HM	-2772.9	-2762.2	10.7	-3431.1	-3357.8	73.3
BH	-2789.6	-2773.4	16.2	-3453.2	-3385.6	67.6
RM	-2785.6	-2782.5	3.1	-3454.2	-3387.6	66.6
BG	-2785.9	-2783.6	2.3	-3454.3	-3388.2	66.1
LM	-2786.7	-2783.0	3.7	-3456.3	-3387.1	69.2
CJ^I	-2786.8	-2782.6	3.8	-3456.2	-3387.3	68.9
GC	-2786.8	-2784.3	2.5	-3456.6	-3388.2	68.5

Bridge sampling estimators: harmonic mean (HM), reciprocal mean (RM), bridge harmonic (BH) and bridge geometric (BG). *Point-based* estimators: Gaussian copula (GC), independence Chib & Jeliazkov (CJ^I) and Laplace Metropolis (LM).

(note that this was also true in the case of the importance posteriors). Finally, the four bridge sampling estimators are not computationally demanding (less than 20 seconds per 1000 additional iterations for the 3-factor model).

Regarding the point based methods, the three estimators yielded similar BML values in most cases, with close MCEs. It should be noted that the GC method yielded the smallest estimated BML, in all examples. Moreover, the error of the CJ^I estimator was increased (as compared to the LM and GC) in the case of the three factor model. This result is related to the fact that the CJ^I uses nested MG runs that, in the case of the three factor model, account for more than 2800 parameters. Even though the MCE can be reduced by increasing the number of iterations R, the CJ^I estimator is computationally demanding (ten minutes per 1000 additional iterations for the 3-factor model). However, the method is favored by the fact that the posterior ordinate is directly obtained by the Metropolis kernel, while no additional assumptions are imposed during the marginal likelihood estimation. On the contrary, the LM and the GC estimators are quick approximation techniques (less than a minute per per 1000 additional iterations for the 3-factor model) but they impose distributional restrictions for the posterior, such as normality or symmetry.

In conclusion, the bridge geometric estimator is favored in terms of stability of the estimated values, MCE and computational time required. It is recommended as a generally applicable method, provided that an efficient importance function can be constructed. For multi-modal posteriors the importance posteriors (PSE) can be also considered. The LM method is also highly efficient and quick and is appropriate if the posterior is not expected to be highly skewed. The CJ^I method on the other hand, even though it is time consuming, is recommended as a benchmark method since it is directly connected to the Metropolis sampler (for instance it reveals MCMC issues such as the well known label switching). These results are well comparable with previously reported ones referring to other type of models, such as Han and Carlin (2000), Bos (2002), Lopes and West (2004), Ardia et al. (2009), Marin and Robert (2009), Fiorentini et al. (2012) and Friel and Wyse (2012) among others.

Chapter 7

Discussion and future research

‘‘Science never solves a problem without creating ten more "

George Bernard Shaw[*]

[*] George Bernard Shaw (1856-1950): the only person who has been awarded both a Nobel Prize in Literature (1925) and an Oscar (1938).

The wide use of the LVMs in a variety of scientific fields nowadays, prompts for the development of quicker and more efficient methods for latent variable modelling. Applied research requires models that are easy to use but at the same time able to account for increasingly complicated theoretical constructs. The LVMs ought to provide solutions in fields where the arbitrariness of the problem at hand needs to be confronted with robust mathematical and statistical tools. By nature, the LVMs are strongly related to the ideas of Bayesian statistics. However, due to the dimensionality of models and the corresponding computational burden, the LVMs are most often addressed in a semi rather than a fully Bayesian perspective. Beyond the advances in computing, the basic features and properties of the LVMs can be implemented in order to improve model estimation and evaluation. This thesis was motivated within this context and hopefully the findings presented here can be used to facilitate the research on the field. Shaw's quote however reminds us that the problems that need to be solved outnumber our solutions.

One of the most important features in Bayesian statistics is the prior information, which determines to a great degree the imposed model. Up to this day, the standard normal distribution is typically selected for the prior of the latent variables. Within the fully Bayesian approach this choice can be challenged in a straightforward manner in order to apply the latent variables methodology to different types of theoretical problems. With regard to the model parameters, on the other hand, different choices have been proposed in the literature. In the example presented here, a prior originally initiated within the GLMs methodology, was used for the case of binary data (see Section 2.2). This prior embodies a number of desired properties and can be expanded to other types of data. Moreover, the priors of the latent and item parameters, need to address also the issue of the identifiability in the case of the LVMs. The identifiability problem along with the so-called model switching problem is an open issue, quite rigorous to address, especially in the continuous case (latent trait models).

At this thesis the local independence assumption of the LVMs was used to reduce the error of the MCMC estimators (Chapter 3) and to simplify the CJ estimator in particular (Chapter 4). In a similar manner, the specific characteristics of the LVMs can be used to modify and improve the sampling algorithms employed to draw samples from the posterior. For instance, the Metropolis-within-Gibbs algorithm presented in this thesis can be modified based on the local independence assumption in order to reduce the computational time required, following for instance the ideas presented in the computation of the CJ^I estimator.

In the last chapter of this thesis, the close relationship between thermodynamics and Bayesian statistics was illustrated. Within the path sampling or the stepping stone sampling methodologies it becomes apparent that the laws of thermodynamics find a direct application in several aspects of statistics. Here, the Boltzmann-Gibbs distribution pertaining to different Hamiltonians was implemented to derive tempered transitions along paths, linking the distributions of interest at the endpoints. Existing marginal likelihood and Bayes factor estimators were reviewed along with their stepping-stone sampling analogues. New estimators were presented and the use of compound paths was introduced. The unified framework in thermodynamic integration offers new highways for research and further investigation. Here we discussed only some of the possible future research directions (Chapter 5). The thermodynamic integration can be proven a valuable tool in statistics, not only in order to facilitate new methods for model evaluation, but also in order to provide new intuition concepts such as the marginal likelihood, the prior and the posterior distributions and their divergencies, the Bayes factor and others.

Thank you for reading this thesis,

Silia Vitoratou

Bibliography

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

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

Algina, J. (1980). A note on identification in the oblique and orthogonal factor analysis models. *Psychometrika*, 45:393–396.

Ali, S. M. and Silvey, S. D. (1966). A general class of coefficients of divergence of one distribution from another. *Journal of the Royal Statistical Society, Series B*, 28(1):131–142.

Andersen, E. (1980). Comparing latent distributions. *Psychometrika*, 45(1):121–134.

Anderson, T. W. and Rubin, H. (1956). Statistical inference in factor analysis. In Newman, J., editor, *Proceedings of the third Berkley Symposium on Mathematical Statistics and Probability*, pages 111–150. University of California Press.

Ardia, D., Hoogerheide, L., and van Dijk, H. K. (2009). To Bridge, to Warp or to Wrap? A comparative study of Monte Carlo methods for efficient evaluation of marginal likelihoods. Tinbergen Institute Discussion Papers 09-017/4, Tinbergen Institute.

Baignères, T., Sepehrdad, P., and Vaudenay, S. (2010). Distinguishing distributions using Chernoff information. In *Proceedings of the 4th international conference on Provable security*, ProvSec'10, pages 144–165, Berlin, Heidelberg. Springer-Verlag.

Barankin, E. and Maitra, E. (1963). Generalisation of the Fisher-Darmois-Koopman-Pitman theorem on sufficient statistics. *Sankhyā A*, 25:217–244.

Bartholomew, D. and Knott, M. (1999). *Latent variable models and factor analysis.* Kendall's Library of Statistics, 7. Wiley.

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.

Bartholomew, D. J., Steele, F., Moustaki, I., and Galbraith, J. (2008). *Analysis of multivariate social science data*. Chapman & Hall/CRC, 2 edition.

Bayarri, M. J. and Berger, J. O. (1997). Measures of surprise in Bayesian analysis. ISDS Discussion Paper 97-46, Duke University.

Bayarri, M. J. and Berger, J. O. (1999). Quantifying surprise in the data and model verification. In Bernardo, J. M., Berger, J. O., Dawid, A. P., and Smith, A. F. M., editors, *Bayesian Statistics 6*, pages 53–82. Oxford University Press.

Beguin, A. and Glas, C. (2001). MCMC estimation and some model-fit analysis of multidimensional IRT models. *Psychometrika*, 66(4):541–561.

Behrens, G., Friel, N., and Hurn, M. (2012). Tuning tempered transitions. *Statistics and Computing*, 22(1):65–78.

Bekker, P. (1986). A note on the identification of restricted factor loading matrices. *Psychometrika*, 51:607–611.

Berger, J. (1996). An overview of robust Bayesian analysis. *Journal of the American Statistical Association*, (91):1343–1370.

Berger, J., Bernardo, J., M., and Sun, D. (2009). The formal definition of reference priors. *The Annals of Statistics*, (37):905–938.

Berger, J. O. and Pericchi, L. (1996). The intrinsic Bayes factor for model selection and prediction. *Journal of the American Statistical Association*, (91):109–122.

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

Bhattacharyya, A. (1943). On a measure of divergence between two statistical populations defined by their probability distributions. *Bulletin of the Calcutta Mathematical Society*, 35:99–109.

Binder, K. (1986). Introduction: theory and techinical aspects of Monte Carlo simulations. In Binder, K., editor, *Monte Carlo methods in Statistical Physics*, Topics in current physics 7. Berlin: Springer.

Binet, A., Simon, T., and Kite, E. (1916). *The development of intelligence in children: (the Binet-Simon Scale).* Publications of the Training School at Vineland, Department of Research. Williams & Wilkins, New Jersey, USA.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In Lord, F. M. and Novick, M. R., editors, *Statistical theories of mental test scores*, pages 397–479. Addison-Weselley, Reading, MA.

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. (1997). A brief history of item theory response. *Educational Measurement: Issues and Practice*, 16(4):21–33.

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

Bolt, D. M. and Lall, V. F. (2003). Estimation of Compensatory and Noncompensatory Multidimensional Item Response models using Markov chain Monte Carlo. *Applied Psychological Measurement*, 27(6):395–414.

Bos, C. S. (2002). A comparison of marginal likelihood computation methods. Tinbergen Institute Discussion Papers 02-084/4, Tinbergen Institute.

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

Bregman, L. (1967). The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *{USSR} Computational Mathematics and Mathematical Physics*, 7(3):200 – 217.

Brooks, S. P. (1998). Markov chain Monte Carlo method and its application. *The Statistician*, 47(1):69–100.

Browne, M. W. (2001). An overview of analytic rotation in Exploratory Factor Analysis. *Multivariate Behavioral Research*, 36:111–150.

Calderhead, B. and Girolami, M. (2009). Estimating Bayes factors via thermodynamic integration and population MCMC. *Computational Statistics and Data Analysis*, 53(12):4028–4045.

Carlin, B. P. and Chib, S. (1995). Bayesian model choice via Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(3):473–484.

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

Casella, G. and George, E. I. (1992). Explaining the Gibbs Sampler. *The American Statistician*, 46(3):167–174.

Chen, M.-H., Ibrahim, J. G., and Shao, Q.-M. (2000). Power prior distributions for generalized linear models. *Journal of Statistical Planning and Inference*, 84:121–137.

Chernoff, H. (1952). A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. *The Annals of Mathematical Statistics*, 23(4).

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

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

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

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

Cover, T. M. and Thomas, J. A. (1991). *Elements of Information Theory*. Wiley-Interscience, New York, NY, USA.

Crooks, G. E. (1999). Entropy production fluctuation theorem and the nonequilibrium work relation for free energy differences. *Physical review E. Statistical physics, plasmas, fluids, and related interdisciplinary topics*, 60:2721–2726.

Crooks, G. E. and Sivak, D. A. (2011). Measures of trajectory ensemble disparity in nonequilibrium statistical dynamics. *Journal of Statistical Mechanics: Theory and Experiment*, 2011(6).

Csiszár, I. (1963). Eine informationstheoretische Ungleichung und ihre anwendung auf den Beweis der ergodizität von Markoffschen Ketten. *Publications of the Mathematical Institute of Hungarian Academy of Sciences*, 8:95–108.

Dale, A. (1999). *A history of the inverse probability: From Thomas Bayes to Karl Pearson*. Sources and Studies in the History of Mathematics and Physical Sciences. Springer-Verlag, London, 2nd edition.

Dean, N. and Raftery, A. (2010). Latent class analysis variable selection. *Annals of the Institute of Statistical Mathematics*, 62(1):11–35.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B*, 39:1–38.

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.

Dunn, J. (1973). A note on a sufficiency condition for uniqueness of a restricted factor matrix. *Psychometrika*, 38:141–143.

Dunson, D., Palomo, J., and Bollen, K. (2005). Bayesian structural equation modeling. Technical report, Statistical and Applied Mathematical Sciences Institute.

Elffers, H., Bethlehem, J., and Gill, R. (1978). Indeterminacy problems and the interpretation of factor analysis results. *Statistica Neerlandica*, 32:181–199.

Evans, M., Gilula, Z., and Guttman, I. (1988). *Latent Class Analysis of Two-way Contingency Tables by Bayesian Methods*. Technical report (University of Toronto. Dept. of Statistics). University of Toronto, Department of Statistics.

Everitt, B. S. (1984). *Introduction to Latent Variable Models*. Chapman and Hall, London,UK.

Fan, Y., Wu, R., Chen, M., Kuo, L., and Lewis, P. (2011). Choosing among partition models in Bayesian phylogenetics. *Molecular Biology and Evolution*, 28(2):523–532.

Fienberg, S. E. (2006). When did Bayesian inference become "Bayesian"? *Bayssian Analysis*, 1(1):1–41.

Fiorentini, G., Planas, C., and Rossi, A. (2012). The marginal likelihood of dynamic mixture models. *Computational Statistics and Data Analysis*, 56(9):2650–2662.

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.

Fox, J. and Glas, C. A. W. (2001). Bayesian estimation of a multilevel IRT model using Gibbs sampling. *Psychometrika*, 66(2):271–288.

Frenkel, D. (1986). Free-energy computation and first-order phase transition. In Ciccoti, G. and Hoover, W. G., editors, *Molecular-Dynamics simulation of Statistical - Mechanical systems*, pages 151–188. Amsterdam: North Holland.

Friel, N., Hurn, M., and Wyse, J. (2012). Improving power posterior estimation of statistical evidence. *To appear in Statistics and Computing.*

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.

Friel, N. and Wyse, J. (2012). Estimating the evidence - a review. *Statistica Neerlandica*, 66(3):288–308.

Frühwirth-Schnatter, S. and Lopes, H. (2010). Parsimonious Bayesian factor analysis when the number of factors is unknown. Technical report, Booth School of Business, University of Chicago.

Galton, F. (1888). Co-relations and their measurement, chiefly from anthropometric data. *Proceedings of The Royal Society of London*, 45:135–145.

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. (1994). Path sampling for computing normalizing constants: identities and theory. Technical Report 376, University of Chicago, Dept. Statistics.

Gelman, A. and Meng, X. (1998). Simulating normalizing constants: from importance sampling to bridge sampling to path sampling. *Statistical Science*, 13(2):163–185.

Gelman, A., Roberts, G. O., and Gilks, W. R. (1996). Efficient Metropolis jumping rules. In Bernardo, J., Berger, J., Dawid, A., and Smith, A., editors, *Bayesian statistics 5: Proceedings of the Fifth Valencia International Meeting*, pages 599–607. Oxford University Press, New York.

Gelman, A. and Shalizi, C. R. (2013). Philosophy and the practice of Bayesian statistics. *British Journal of Mathematical and Statistical Psychology*, 66(1):8–38.

Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6(6):721–741.

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

Geyer, C. J. (1991). Markov chain Monte Carlo maximum likelihood. In *Computing Science and Statistics: Proc. 23rd Symp. Interface*, pages 156–163.

Geyer, C. J. and Thompson, E. A. (1995). Annealing Markov chain Monte Carlo with applications to ancestral inference. *Journal of the American Statistical Association*, 90(431):909–920.

Ghosh, J. and Dunson, D. (2009). Default prior distributions and efficient posterior computation in Bayesian factor analysis. *Journal of Computational and Graphical Statistics*, 18(2):306–320.

Ghosh, J., Herring, A. H., and Siega-Riz, A. M. (2011). Bayesian variable selection for Latent Class Models. *Biometrics*, 67(3):917–925.

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.

Gilks, W., Richardson, S., and Spiegelhalter, D. (1996). Introducing Markov chan Monte Carlo. In W.R.Gilks, S. R. and Spiegelhalter, D., editors, *Markov Chain Monte Carlo in Practice*, pages 165–187. London: Chapman and Hall.

Glas, C. A. W. and Meijer, R. R. (2003). A Bayesian approach to Person Fit Analysis in Item Response Theory models. *Applied Psychological Measurement*, 27(3):217–233.

Goodman, L. (1978). *Analyzing qualitative categorical data: log-linear models and latent-structure analysis.* Edited by Magidson, J. Abt Books, Inc., Cambridge.

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

Green, P. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82:711–732.

Gustafson, P. (1996). Local sensitivity of inferences to *p*rior *m*arginals. *Journal of the American Statistical Association*, (91):774–781.

Guttman, L. (1955). The determinancy of factor score matrices with implications fro five other basic problems of common factor theory. *British Journal of Statistical Psychology*, 8:65–81.

Hambleton, R., Swaminathan, H., and Rogers, H. (1991). *Fundamentals of Item Response Theory.* Measurement Methods for the Social Science. SAGE Publications, Newbury Park, CA, USA.

Han, C. and Carlin, B. (2000). MCMC Methods for computing Bayes Factors: a comparative review. *Journal of the American Statistical Association*, 96:1122–1132.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109.

Heinen, T. (1996). *Latent Class and Discrete Latent Trait Models: Similarities and Differences*. Advanced Quantitative Techniques in the Social Sciences. SAGE Publications Inc, Thousand Oaks.

Hellinger, E. (1909). Neue Begründung der Theorie quadratischer Formen von unendlichvielen Veränderlichen. *Journal f*
ddotur die reine und angewandte Mathematik, 136:210–271.

Hoijtink, H. (1998). Constrained Latent Class analysis using the Gibbs sampler and Posterior Predictive p-values: applications to educational testing. *Statistica Sinica*, 8:691–711.

Hoijtink, H. (2001). Confirmatory Latent Class analysis: Model selection using Bayes factors and (pseudo) likelihood ratio statistics. *Multivariate Behavioral Research*, 36(4):563–588.

Hoijtink, H. and Molenaar, I. (1997). A multidimensional item response model: Constrained Latent Class analysis using the Gibbs sampler and Posterior Predictive checks. *Psychometrika*, 62(2):171–189.

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.

Ibrahim, J. G. and Chen, M.-H. (2000). Power Prior Distributions for Regression Models. *Statistical Science*, 15(1):46–60.

Janssen, R., Tuerlinckx, F., Meulders, M., and De Boeck, P. (2000). A hierarchical IRT model for criterion-referenced measurement. *Journal of Educational and Behavioral Statistics*, 25:285–306.

Jarzynski, C. (1997). Nonequilibrium equality for Free Energy differences. *Physical Review Letters*, 78(14):2690–2693.

Jeffreys, H. (1946). An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences*, 186(1007):453–461.

Jeffreys, H. (1961). *Theory of probability*. Oxford *U*niversity *P*ress”, Oxford.

Joe, H. (1997). *Multivariate Models and Multivariate Dependence Concepts*. London: Chapman and Hall.

Johnson, D. and Sinanovic, S. (2000). Symmetrizing the Kullback-Leibler distance. Technical report, IEEE Transactions on Information Theory.

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.

Jöreskog, K. (1969). A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika*, 34:183–202.

Jöreskog, K. G. (1970). A general method for analysis of covariance structures. *Biometrika*, 57:239–251.

Julier, S. (2006). An empirical study into the use of Chernoff information for robust, distributed fusion of Gaussian mixture models. In *Information Fusion, 2006 9th International Conference on*, pages 1–8.

Kakizawa, Y., Shumway, R., and Taniguchi, N. (1998). Discrimination and clustering for multivariate time series. *Journal of the American Statistical Association*, 93(441):328–340.

Kang, T. and Cohen, A. S. (2007). IRT model selection methods for dichotomous items. *Applied Psychological Measurement*, 31(4):331–358.

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

Kass, R. E. and Wasserman, L. (1996). The selection of prior distributions by formal rules. *Journal of the American Statistical Association*, 91(435):1343–70.

Kim, J.-S. and Bolt, D. M. (2007). Estimating Item Response Theory models using Markov chain Monte Carlo methods. *Educational Measurement: Issues and Practice*, 26(4):38–51.

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

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.

Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *Annals of Mathematical Statistics*, 22:49–86.

Langeheine, R. and Rost, J. (1988). *Latent Trait and Latent Class Models*. Plenum Press, New York, USA.

Lartillot, N. and Philippe, H. (2006). Computing Bayes factors using Thermodynamic Integration. *Systematic Biology*, 55:195–207.

Lawley, D. N. and Maxwell, A. E. (1963). Factor Analysis as a statistical method. *Journal of the Royal Statistical Society. Series D (The Statistician)*, 12(3):209–229.

Lazarsfeld, P. F. and Henry, N. W. (1968). *Latent Structure Analysis*. Houghton Mifflin, Boston, USA.

Lee, S. (2007). *Structural Equation Modeling: A Bayesian Approach*. Wiley Series in Probability and Statistics. Wiley.

Lee, S.-Y. and Song, X.-Y. (2001). Hypothesis testing and model comparison in two-level Structural Equation Models. *Multivariate Behavioral Research*, 36(4):639–655.

Lee, S.-Y. and Song, X.-Y. (2003). Bayesian analysis of Structural Equation Models with dichotomous variables. *Statistics in Medicine*, 22:3073–3088.

Lee, Y. and Nelder, J. (2009). Likelihood inference for models with unobservables: another view. *Statistical Science*, 24(3):255–269.

Lefebvre, G., Steele, R., and Vandal, A. C. (2010). A path sampling identity for computing the Kullback-Leibler and J divergences. *Computational Statistics and Data Analysis*, 54(7):1719–1731.

Lewis, M. (1994). *Multilevel Modeling of Discrete Event History Data Using Markov Chain Monte Carlo Methods. Unpublished doctoral dissertation.* PhD thesis, University of Washington, Dept. of Statistics.

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.

Liese, F. and Vajda, I. (2006). On divergences and informations in statistics and information theory. *IEEE Transactions on Information Theory*, 52(10):4394–4412.

Lopes, F. L. (2003). Expected posterior priors in factor analysis. *Brazilian Journal of Probability and Statistics*, 17:91–105.

Lopes, H. F. and West, M. (2004). Bayesian model assessment in factor analysis. *Statistica Sinica*, 14:41–67.

Lord, F. (1952). *A theory of test scores (Psychometric Monograph No. 7)*. Retrieved from http://www.psychometrika.org/journal/online/MN07.pdf. Psychometric Corporation, Rischmond, VA.

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.

Lunn, D., S.-D. T. A. and Best, N. (2009). The BUGS project: Evolution, critique and future directions (with discussion). *Statistics in Medicine*, 28:3049–3082.

Marin, J.-M. and Robert, C. P. (2009). Importance sampling methods for Bayesian discrimination between embedded models.

Marinari, E. and Parisi, G. (1992). Simulated Tempering: A new Monte Carlo scheme. *EPL (Europhysics Letters)*, 19(6):451.

McCullagh, P. and Nelder, J. (1989). *Generalized Linear Models.* Monographs on statistics and applied probability. Chapman and Hall.

McCulloch, R. and Rossi, P. (1992). Bayes factors for nonlinear hypotheses and likelihood distributions. *Biometrika*, 79(4):663–676.

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.

Merhav, N. (2010). Statistical physics and Information Theory. In *Foundations and trends in communications and Information Theory*, volume 6, pages 1–212. Boston - Delft: Now Publishers.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equation of State Calculations by Fast Computing Machines. *The Journal of Chemical Physics*, 21(6):1087–1092.

Meyn, S. and Tweedie, R. L. (2009). *Markov Chains and Stochastic Stability.* Cambridge University Press, New York, NY, USA, 2nd edition.

Mislevy, R. J. (1986). Bayes modal estimation in item response models. *Psychometrika*, 51(2):177–195.

Moustaki, I. (1996). A latent trait and a latent class model for mixed observed variables. *British journal of mathematical and statistical psychology*, 49(2):313–334.

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

Mutheén, B. and Asparouhov, T. (2012). Bayesian structural equation modeling: A more flexible representation of substantive theory. *Psychological Methods*, 17(3):313–335.

Neal, R. (1996). Sampling from multimodal distributions using tempered transitions. *Statistics and Computing*, 6(4):353–366.

Neal, R. M. (1993). Probabilistic inference using Markov chain Monte Carlo methods. Technical Report CRG-TR-93-1, University of Toronto.

Neal, R. M. (2001). Annealed importance sampling. *Statistics and Computing*, 11(2):125–139.

Neuenschwander, B., Branson, M., and Spiegelhalter, D. (2009). A note on the power prior. *Statistics in Medicine*, 28:3562–3566.

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

Nielsen, F. (2011). Chernoff information of exponential families. *Computing Research Repository*, abs/1102.2684.

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

Ntzoufras, I. (2011). *Bayesian Modeling using WinBUGS*. Wiley Series in Computational Statistics. Wiley.

Ntzoufras, I., Dellaportas, P., and Forster, J. (2000). Bayesian variable and link determination for Generalised Linear Models. *Journal of Statistical Planning and Inference*, 111(1-2):165–180.

Nussbaum,M and Szkoła, A. (2009). The Chernoff lower bound for symmetric quantum hypothesis testing. *The Annals of Statistics*, 37(2):1040–1057.

Ogata, Y. (1989). A Monte Carlo method for high dimensional integration. *Numerische Mathematik*, 55:137–157.

Parzen, E. (1992). Time series, statistics, and information. New directions in time series analysis. Part I, Proc. Workshop, Minneapolis/MN (USA) 1990, IMA Volumes in Mathematics and Its Applications 45, 265-286.

Patz, R. J. and Junker, B. W. (1999a). Applications and extensions of MCMC in IRT: Multiple item types, missing data, and rated responses. *Journal of Educational and Behavioral Statistics*, 24(4):342–366.

Patz, R. J. and Junker, B. W. (1999b). A straightforward approach to Markov chain Monte Carlo methods for item response models. *Journal of Educational and Behavioral Statistics*, 24(2):146–178.

Perez, J. M. and Berger, J. O. (2002). Expected-posterior prior distributions for model selection. *Biometrika*, 89(3):491–512.

Polasek, W. (2000). Factor analysis and outliers: A Bayesian approach. Technical report, Institute of Statistics and Econometrics, University of Basel.

R Core Team (2013). *R a language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria.

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.

Raftery, A. and Banleld, J. (1991). Stopping the Gibbs sampler, the use of morphology, and other issues in spatial statistics. *Annals of the Institute of Statistical Mathematics*, 43(430):32–43.

Raftery, A., Newton, M., Satagopan, J., and Krivitsky, P. (2007). Estimating the integrated likelihood via posterior simulation using the harmonic mean identity. *Bayesian Statistics*, (8):1–45.

Raftery, A. E. (1996). Hypothesis testing and model selection. In Richardson, W. G. S. and Spiegelhalter, D., editors, *Markov Chain Monte Carlo in Practice*, pages 165–187. London: Chapman and Hall.

Raïffa, H. and Schlaifer, R. (1961). *Applied statistical decision theory.* Studies in managerial economics. Division of Research, Graduate School of Business Adminitration, Harvard University.

Rasch, G. (1960). *Probabilistic Models for Some Intelligence and Attainment Tests.* Paedagogiske Institut, Copenhagen.

Rauber, T., Braun, T., and Berns, K. (2008). Probabilistic distance measures of the Dirichlet and Beta distributions. *Pattern Recognition*, 41(2):637 – 645.

Rényi, A. (1961). On measures of entropy and information. In *Proceedings of the 4th Berkeley Symposium on Mathematics, Statistics and Probability*, pages 547–561.

Richardson, M. (1936). The relation between the difficulty and the differential validity of a test. *Psychometrika*, 1(2):33–49.

Rizopoulos, D. and Moustaki, I. (2008). Generalized latent variables models with non-linear effects. *British journal of mathematical and statistical psychology*, 61(2):415–438.

Rousseeuw, P. J. and van Zomeren, B. (1990). Unmasking multivariate outliers and leverage points (with discussion). *Journal of the American Statistical Association*, 85:633–651.

Rubin, D. and Thayer, D. (1982). EM algorithms for ML factor analysis. *Psychometrika*, 47(1):69–76.

Sahu, S. K. (2002). Bayesian estimation and model choice in item response models. *Journal of Statistical Computation and Simulation*, 72(3):217–232.

Sato, M. (1991). A study of an identification problem and substitute use of principal component analysis in factor analysis. *Hiroshima Mathematical Journal*, 22:607–611.

Scheines, R., Hoijtink, H., and Boomsma, A. (1999). Bayesian estimation and testing of structural equation models. *Psychometrika*, 24(1):37–52.

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.

Sheng, Y. (2008). A MATLAB package for Markov Chain Monte Carlo with a multi-unidimensional IRT model. *Journal of Statistical Software*, 28(10):1–20.

Sinharay, S. (2005). Assessing fit of unidimensional item response theory models using a bayesian approach. *Journal of Educational Measurement*, 42(4):375–394.

Sinharay, S. (2006). Bayesian item fit analysis for unidimensional item response theory models. *British Journal of Mathematical and Statistical Psychology*, 59(2):429–449.

Sinharay, S., Johnson, M. S., and Stern, H. S. (2006). Posterior predictive assessment of item response theory models. *Applied Psychological Measurement*, 30(4):298–321.

Sinharay, S. and Stern, H. (2002). On the Sensitivity of Bayes Factors to the prior distributions. *The American Statistician*, 56:196–201.

Skrondal, A. and Rabe-Hesketh, S. (2004). *Generalized Latent Variable Modeling Multilevel, Longitudinal, and Structural Equation Models*. Chapman & Hall, Boka Raton.

Song, P. X.-K. (2000). Multivariate Dispersion Models Generated from Gaussian Copula. *Scandinavian Journal of Statistics*, 27:305–320.

Spearman, C. (1904). General intelligence, objectively determined and measured. *American Journal of Psychology*, 15:201–293.

Spiegelhalter, D. J. and Smith, A. F. M. (1982). Bayes factors for linear and log-linear models with vague prior information. *Journal of the Royal Statistical Society. Series B (Methodological)*, 44:377–387.

Spiegelhalter, S. D., Best, N. G., Carlin, B. P., and Linde, A. V. D. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(4):583–639.

Steiger, J. (1979). Factor indeterminacy in the 1930's and the 1970's some interesting parallels. *Psychometrika*, 44:157–167.

Stigler, S. M. (1983). Who discovered Bayes's Theorem? *The American Statistician*, 37(4a):290–296.

Stone, C. A. (2000). Monte Carlo based null distribution for an alternative goodness-of-fit test statistic in IRT models. *Journal of Educational Measurement*, 37(1):58–75.

Sturtz, S., Ligges, U., and Gelman, A. (2005). R2WinBUGS: A Package for Running WinBUGS from R. *Journal of Statistical Software*, 12(3):1–16.

Tanner, M. A. and Wong, W. H. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*, 82(398):528–540.

Thurstone, L. (1931). Multiple factor analysis. *Psychological Review*, 38:406–427.

Thurstone, L. (1947). *Multiple-factor analysis: a development and expansion of the vectors of the mind*. The University of Chicago Press, Chicago, USA.

Tierney, L. and Kadane, J. (1986). Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association*, 81:82–86.

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

van Onna, M. (2002). Bayesian estimation and model selection in ordered latent class models for polytomous items. *Psychometrika*, 67(4):519–538.

Williams, J. (1978). A definition for the common-factor analysis model and the elimination of problems of factor score indeterminacy. *Psychometrika*, 43:293–306.

Wilson, E. (1928). Review of "The abilities of man, their nature and measurment" by C. Spearman. *Science*, 67:244–248.

Xie, W., Lewis, P., Fan, Y., Kuo, L., and Chen, M. (2011). Improving marginal likelihood estimation for Bayesian phylogenetic model selection. *Systematic Biology*, 60(2):150–160.