Research Article

Johannes Reichl*

Estimating marginal likelihoods from the posterior draws through a geometric identity

<https://doi.org/10.1515/mcma-2020-2068>

Received February 25, 2020; accepted June 24, 2020

Abstract: This article develops a new estimator of the marginal likelihood that requires only a sample of the posterior distribution as the input from the analyst. This sample may come from any sampling scheme, such as Gibbs sampling or Metropolis–Hastings sampling. The presented approach can be implemented generically in almost any application of Bayesian modeling and significantly decreases the computational burdens associated with marginal likelihood estimation compared to existing techniques. The functionality of this method is demonstrated in the context of probit and logit regressions, on two mixtures of normals models, and also on a high-dimensional random intercept probit. Simulation results show that the simple approach presented here achieves excellent stability in low-dimensional models, and also clearly outperforms existing methods when the number of coefficients in the model increases.

Keywords: Bayesian statistics, model evidence, integrated likelihood, model selection, estimation of normalizing constant

MSC 2010: 62F15, 65C60, 65R20

1 Introduction

Bayesian model selection relies on the posterior probabilities of the H candidate models $\mathcal{M}_1, \dots, \mathcal{M}_H$ conditional on the data [3, 19]. Two major strategies exist for estimating the posterior probabilities $p(\mathcal{M}_h|\mathbf{y})$ for each of the $h = 1, \dots, H$ candidate models: (a) the model indicators are considered as part of the parameter space and jointly drawn during MCMC sampling together with all other unobserved quantities [17]. The drawback of this strategy is the necessity of including all candidate models in the estimation simultaneously. If the analyst then wishes to consider an additional model that was not a part of the initial candidate set, another simultaneous estimation needs to be carried out. This exacts a significant cost on the analyst in terms of computing time and effort. Strategy (b) estimates the marginal likelihood for each model which allows for easy calculation of the posterior probabilities independent from the estimation of the other candidate models [19, 27]. Despite this appealing characteristic, calculating the marginal likelihood is a non-trivial integration problem, and as such it is still associated with significant effort on the part of the analyst and potential imprecision in the case of high-dimensional or multi-level models. Comparative studies of existing estimation techniques for the marginal likelihood only provide clear evidence of precision for candidate models of lesser dimensions, while Bayesian analysis frequently requires more complex models [2, 12, 14]. Regardless of which of these two methodological approaches is used to identify the model with the highest posterior probability, reporting the marginal likelihood of this model supports the comparability and reproducibility of the modelling results and thus is a desirable practice.

This article presents a technique for estimating the marginal likelihood, which alleviates both of the drawbacks that hinder existing approaches. The method presented herein requires only a sample of the pos-

*Corresponding author: Johannes Reichl, Energy Institute at the Johannes Kepler University Linz, Altenberger Straße 69, 4040 Linz, Austria, e-mail: reichl@energieinstitut-linz.at. <https://orcid.org/0000-0003-1054-9454>

terior distribution as an input and is thus implementable as a generic function allowing for its use in a variety of applications. Additionally, the approach shows significantly less sensitivity to an increase in the number of model coefficients compared to existing approaches.

We start by defining the marginal likelihood of model $\mathcal{M}_h$ as

$$p(\mathcal{M}_h|\mathbf{y}) = \int_{\boldsymbol{\theta}_h} p(\mathbf{y}|\boldsymbol{\theta}_h) p(\boldsymbol{\theta}_h) d\boldsymbol{\theta}_h, \quad (1.1)$$

where $\boldsymbol{\theta}_h$ is a $1 \times K$ vector containing the K coefficients of model $\mathcal{M}_h$. For the derivation of the estimator in Section 2, we define $\boldsymbol{\theta}_h$ in such a way that it holds only non-linear dependent elements. For example, for a 3-dimensional covariance matrix parameter $\boldsymbol{\theta}_h$ contains only six values, and for a 5-dimensional vector of probabilities summing up to 1, $\boldsymbol{\theta}_h$ holds only four values. $p(\mathbf{y}|\boldsymbol{\theta}_h)$ refers to the likelihood of model $\mathcal{M}_h$, and $p(\boldsymbol{\theta}_h)$ is the prior distribution with domain $\boldsymbol{\Theta}_h$. Only for a very limited class of models does an analytical solution of (1.1) exist, while in the majority of models (1.1) presents a challenging integration problem. Suppressing the model index h henceforth, and considering the marginal likelihood of a model $\mathcal{M}$ is the normalizing constant of its posterior distribution $p(\boldsymbol{\theta}|\mathbf{y})$, we can rewrite Bayes theorem as

$$\frac{1}{p(\mathcal{M}|\mathbf{y})} = \frac{p(\boldsymbol{\theta}|\mathbf{y})}{p(\mathbf{y}|\boldsymbol{\theta}) p(\boldsymbol{\theta})}. \quad (1.2)$$

Let $\mathbf{A}$ be a bounded subset of prior domain $\boldsymbol{\Theta}$, then integrating both sides in (1.2) over $\mathbf{A}$ gives

$$\frac{1}{p(\mathcal{M}|\mathbf{y})} \int_{\mathbf{A}} \mathbf{1} d\boldsymbol{\theta} = \int_{\mathbf{A}} \frac{1}{p^*(\boldsymbol{\theta}|\mathbf{y})} p(\boldsymbol{\theta}|\mathbf{y}) d\boldsymbol{\theta},$$

where $p^*(\boldsymbol{\theta}|\mathbf{y})$ is used as an abbreviation for the non-normalized posterior $p(\mathbf{y}|\boldsymbol{\theta}) p(\boldsymbol{\theta})$ henceforth. A representation of the marginal likelihood is then found by:

$$p(\mathcal{M}|\mathbf{y}) = \frac{\int_{\mathbf{A}} \mathbf{1} d\boldsymbol{\theta}}{\int_{\mathbf{A}} \frac{1}{p^*(\boldsymbol{\theta}|\mathbf{y})} p(\boldsymbol{\theta}|\mathbf{y}) d\boldsymbol{\theta}}. \quad (1.3)$$

We refer to the integral in the denominator in (1.3) as the non-normalized posterior integral over $\mathbf{A}$ and abbreviate it by $\kappa_{\mathbf{A}}$. Integrating 1 over a K -dimensional bounded set has the geometric interpretation of a generalized volume, or hypervolume, and we will refer to the integral in the nominator in (1.3) as the volume of $\mathbf{A}$.

From 1.3 we see that the original integration problem in (1.1) can be split up into two separate problems, namely the volume of $\mathbf{A}$, and the non-normalized posterior integral over $\mathbf{A}$. Splitting up the estimation of the marginal likelihood into these two sub-problems reduces computation effort significantly, and leads to very stable estimates. In particular, in the case of high-dimensional models. The simpler computation results from the relative ease of implementing techniques to estimate a geometric volume as opposed to techniques for estimating a high-dimensional integral. Additionally, the posterior distribution can be used as an importance density for estimating $\kappa_{\mathbf{A}}$, where all drawbacks of earlier approaches exploiting the posterior distribution as importance density are overcome by limiting the region of integration to $\mathbf{A}$, as defined subsequently.

The remainder of this article is organized as follows: Section 2 details the proposed estimation method. Section 3 applies the developed method to several types of models, these are probit and logit regression, mixture of normals, and the random intercept probit. For each of the applications, a comparison to existing methods is presented, such as Chib's method [5], importance sampling, and bridge sampling [21]. Section 4 summarizes and concludes.

2 The approach

In this section we present a new estimator for the marginal likelihood by separately estimating the volume of $\mathbf{A}$ and the corresponding non-normalized posterior integral as defined in Section 1. We also present a method

for choosing $\mathbf{A}$ in such a way that the quotient of these estimators yields a stable estimate of the marginal likelihood.

First, we turn to the non-normalized posterior integral. A common technique for numerical integration is importance sampling, and since the posterior distribution is part of the numerator in the non-normalized posterior integral this suggests the posterior as the importance density. Since draws from the posterior distribution are usually available as a natural output of Bayesian analysis, the importance sampling estimator of the posterior integral $\hat{\kappa}_{\mathbf{A}}$ is available almost ad hoc once $\mathbf{A}$ has been defined:

$$\hat{\kappa}_{\mathbf{A}} = \frac{1}{L} \sum_{l=1}^L f(\boldsymbol{\theta}^{(l)}) \quad \text{with} \quad f(\boldsymbol{\theta}^{(l)}) = \begin{cases} \frac{1}{p^*(\boldsymbol{\theta}^{(l)}|\mathbf{y})} & \text{if } \boldsymbol{\theta}^{(l)} \in \mathbf{A}, \\ 0 & \text{else,} \end{cases} \quad (2.1)$$

where $\boldsymbol{\theta}^{(l)}$ with $l = 1, \dots, L$ refers to the posterior draws after the burn in. Earlier attempts for estimating the marginal likelihood by importance sampling have likewise exploited the posterior distribution as importance density, among them most prominently is the harmonic mean estimator $\hat{p}_{\text{HM}}$ [22]. In this approach, after some reordering, one integrates both sides of Bayes theorem over the whole prior domain:

$$\frac{1}{p(\mathcal{M}|\mathbf{y})} \int_{\boldsymbol{\theta}} p(\boldsymbol{\theta}) d\boldsymbol{\theta} = \int_{\boldsymbol{\theta}} \frac{1}{p(\mathbf{y}|\boldsymbol{\theta})} p(\boldsymbol{\theta}|\mathbf{y}) d\boldsymbol{\theta},$$

and since integrating the prior over its own domain yields 1, the harmonic mean estimator is given by

$$\hat{p}_{\text{HM}} = \left(\frac{1}{L} \sum_{l=1}^L \frac{1}{p(\mathbf{y}|\boldsymbol{\theta}^{(l)})} \right)^{-1}. \quad (2.2)$$

It is well known that $\hat{p}_{\text{HM}}$ suffers from pseudo-bias due to the use of an unrestricted prior domain as the region of integration [20], and from significant instability [12]. To be valid, the importance density must have its support defined across the entire region of integration. The posterior distribution however, generally does not have such a wide support in practice, and thus is an improper choice as importance density in this respect. The instability stems from the dominating effect that only a small number of draws from the MCMC sampler has on the sum in equation (2.2). Some of the L posterior draws usually result in low values of $p(\mathbf{y}|\boldsymbol{\theta}^{(l)})$, and since the likelihood enters the estimator reciprocal, these small values dominate $\hat{p}_{\text{HM}}$.

However, shortcomings of the harmonic mean estimator guide the way to a definition of $\mathbf{A}$ allowing a stable estimation of the marginal likelihood from identity (1.3). Firstly, to ensure the posterior distribution is a proper choice for the importance density, as required in our approach, the region of integration $\mathbf{A}$ must have full support of the posterior distribution. Secondly, to avoid instability, the region of integration $\mathbf{A}$ may only contain points for which the sum in (2.1) is stable independently of any specific run of the MCMC sampler. To address both of these requirements and at the same time allow for a simple estimation of the volume and the non-normalized posterior integral we define $\mathbf{A}$ as the intersection of two sets $\mathbf{A}_1$ and $\mathbf{A}_2$.

Set $\mathbf{A}_1$ is defined by a threshold value ρ and only those points $\boldsymbol{\theta}$ lie in $\mathbf{A}_1$ whose non-normalized posterior $p^*(\boldsymbol{\theta}|\mathbf{y})$ exceeds this threshold, such that

$$\boldsymbol{\theta} \in \mathbf{A}_1 \quad \text{if } p^*(\boldsymbol{\theta}|\mathbf{y}) > \rho.$$

Considering the series $p^{*,(1,\dots,L)} = p^*(\boldsymbol{\theta}^{(1)}|\mathbf{y}), \dots, p^*(\boldsymbol{\theta}^{(L)}|\mathbf{y})$, a natural way of determining the threshold ρ is to ensure that the lowest values of $p^{*,(1,\dots,L)}$ are not destabilizing the sum in (2.1) by setting ρ as a quantile of $p^{*,(1,\dots,L)}$. We define ρ as the median of $p^{*,(1,\dots,L)}$ because of its robustness even in models of high dimension, and because the posterior distribution has sufficient support to serve as importance density wherever the non-normalized posterior exceeds its median. Thereby, an almost perfectly stable estimation of (2.1) is ensured by excluding all values of $p^{*,(1,\dots,L)}$ from the estimation of $\hat{\kappa}_{\mathbf{A}}$ that might destabilize the sum in (2.1), and a graphical demonstration thereof is given subsequently after the derivation of the final set $\mathbf{A}$.

However, usage of $\mathbf{A}_1$ as $\mathbf{A}$ from (1.3) does not allow easy estimation of its volume and would constitute an integration problem with comparable burdens as the original formulation of the marginal likelihood in (1.1). To facilitate easy estimation of the volume of $\mathbf{A}$ a second set $\mathbf{A}_2$ is defined, in such a way that the volume of the intersection of $\mathbf{A}_1$ and $\mathbf{A}_2$ can be estimated by means of statistical standard techniques only.

As long as $\mathbf{A}_1$ and $\mathbf{A}_2$ have the same dimension K and an intersection $\mathbf{A}$, the volume of this intersection $V_{\mathbf{A}}$ can be written as $V_{\mathbf{A}} = \pi V_{\mathbf{A}_2}$, where π refers to the ratio of points lying in $\mathbf{A}_2$ that also lie in $\mathbf{A}_1$, and $V_{\mathbf{A}_2}$ is the volume of $\mathbf{A}_2$. Hence, for an easy estimation of $V_{\mathbf{A}}$ we define $\mathbf{A}_2$ in such a way that its volume $V_{\mathbf{A}_2}$ can be calculated analytically and that drawing uniformly from within it is efficient and feasible. Then π can simply be estimated by drawing K -dimensional vectors $\boldsymbol{\theta}^{(r)}$ uniformly from $\mathbf{A}_2$ such that

$$\hat{\pi} = \frac{1}{R} \sum_{r=1}^R I(\boldsymbol{\theta}^{(r)} \in \mathbf{A}_1),$$

where $I(\cdot)$ refers to the indicator function, and R is the number of random draws from within $\mathbf{A}_2$.

Notice that all of the above is valid for unimodal as well as for multimodal distributions. However, henceforth we derive the algorithm for estimating the marginal likelihood for unimodal distributions, while a minor change which will allow for the estimation of $p(\mathcal{M}|\mathbf{y})$ for multimodal distributions is presented in Section 3.2 and applied to a mixture of normal distributions.

Even though other definitions of $\mathbf{A}_2$ are possible, we choose a K -dimensional ellipsoid as $\mathbf{A}_2$. This choice shows outstanding efficiency of the resulting estimator, and to dispel doubts that the choice of an ellipsoid might limit the applicability of the approach, applications in Section 3 focus on posterior distributions with non-elliptical contours. The set of points $\boldsymbol{\theta}$ lying in $\mathbf{A}_2$ is thus defined by

$$\boldsymbol{\theta} \in \mathbf{A}_2 \quad \text{if } (\boldsymbol{\theta} - \boldsymbol{\theta}^*)\mathbf{C}(\boldsymbol{\theta} - \boldsymbol{\theta}^*)' < 1, \quad (2.3)$$

where $\mathbf{C}$ is a positive definite matrix of dimension $K \times K$ with its eigenvectors defining the principal axes of the ellipsoid. $\boldsymbol{\theta}^*$ is a point with support of the posterior distribution, and we define $\boldsymbol{\theta}^*$ as its posterior mode to ensure substantial overlap between $\mathbf{A}_2$ and $\mathbf{A}_1$.

The last step in defining $\mathbf{A}_2$ is thus choosing $\mathbf{C}$. Consider matrix $\mathbf{T} = (\boldsymbol{\theta}^{(1)'}, \dots, \boldsymbol{\theta}^{(L)'})'$, and its covariance matrix $\mathbf{D} = \text{cov}(\mathbf{T})$, we define

$$\mathbf{C} = (\alpha \mathbf{D})^{-1},$$

where α is a scalar with domain $\mathbb{R}^+$, and is employed as a tuning parameter in the presented approach. In all applications in Section 3, α was set in such a way that the resulting intersection of $\mathbf{A}_1$ and $\mathbf{A}_2$ contains about 49 % of the L posterior draws of $\boldsymbol{\theta}^{(l)}$. We include in the Appendix a trivial binary search method for finding the value of α . The choice of setting α to contain 49 % of the posterior draws can be understood intuitively: since ρ is defined as the median of $p^*, (1, \dots, L)$, set $\mathbf{A}_1$ contains exactly (for an even L) 50 % of the posterior draws, and since $\mathbf{A}$ is the intersection of $\mathbf{A}_1$ and $\mathbf{A}_2$, it cannot contain more than these 50 %. While the majority of points $\boldsymbol{\theta}^{(l)}$ for which $p^*(\boldsymbol{\theta}^{(l)}|\mathbf{y}) > \rho$ have comparably low distance to the posterior mode $\boldsymbol{\theta}^*$, in high-dimensional models with very non-elliptical contours some points far away from the mode can still fulfill $p^*(\boldsymbol{\theta}^{(l)}|\mathbf{y}) > \rho$. If α is chosen to include all 50 % of the posterior draws fulfilling this condition, the resulting ellipsoid from (2.3) might have much mass in its outer layers where only few points exceed the threshold ρ , with potentially negative effects on the efficiency of estimating $\hat{\pi}$. Setting α to include the 49 % of draws with lowest distance to the mode increases the efficiency of estimator $\hat{\pi}$, while no evidence for a negative impact on the stability of $\hat{\kappa}_{\mathbf{A}}$ is found.

Figure 1 compares the values of the non-normalized posterior $p^*(\boldsymbol{\theta}^{(l)}|\mathbf{y})$ for all $l = 1, \dots, L$ draws of the posterior distribution to those values of $p^*(\boldsymbol{\theta}^{(l)}|\mathbf{y})$ for which $\boldsymbol{\theta}^{(l)} \in \mathbf{A}$, for a low and a high-dimensional model. The almost completely flat lower boundary of the non-normalized posterior values under $\mathbf{A}$ allow for an extraordinarily stable estimation of the posterior integral in (2.1) independent of the dimension of the respective model.

Algorithm I for the estimation of the marginal likelihood is thus given by:

- (1) Run an MCMC sampler to obtain L posterior draws $\boldsymbol{\theta}^{(l)}$ after the burn in, calculate the series of non-normalized posterior density values $p^*, (1, \dots, L)$, and set ρ to its median, $\boldsymbol{\theta}^*$ to the posterior mode, and $\mathbf{D}$ to the covariance matrix of the posterior draws.
- (2) Define α in such a way that $0.49 \times L$ draws are in $\mathbf{A}$, where $\boldsymbol{\theta}^{(l)}$ is an element of $\mathbf{A}$ if $p^*(\boldsymbol{\theta}^{(l)}|\mathbf{y}) > \rho$ and $(\boldsymbol{\theta}^{(l)} - \boldsymbol{\theta}^*)(\alpha \mathbf{D})^{-1}(\boldsymbol{\theta}^{(l)} - \boldsymbol{\theta}^*)' < 1$.
- (3) Draw R points $\boldsymbol{\theta}^{(r)}$ from $\mathbf{A}_2$, count the number $\bar{r}$ of draws for which $p^*(\boldsymbol{\theta}^{(r)}|\mathbf{y}) > \rho$, and set $\hat{\pi} = \frac{\bar{r}}{R}$.

Figure 1: Values of the non-normalized posterior $p^*(\theta^{(l)}|\mathbf{y})$ for all $l = 1, \dots, L$ draws of the MCMC sampler (black, in the background) overlaid by those values of $p^*(\theta^{(l)}|\mathbf{y})$ for which $\theta^{(l)} \in \mathbf{A}$ (red, in the front); left side: 4-dimensional model $\mathcal{M}_{16}$ from Section 3.1; right side: 142-dimensional model from Section 3.3. The flat lower boundary of the red points ensures stability of the sum in (2.1), i.e. stability of $\sum \frac{1}{p^*(\theta^{(l)}|\mathbf{y})}$.

- (4) Estimate the volume of $\mathbf{A}$ as $\hat{V}_{\mathbf{A}} = \hat{\pi} V_{\mathbf{A}_2}$, and obtain the estimator for the non-normalized posterior integral $\hat{\kappa}_{\mathbf{A}}$ from (2.1).
- (5) Calculate the final estimator of the marginal likelihood as

$$\hat{p}_{\mathbf{A}}(\mathcal{M}|\mathbf{y}) = \frac{\hat{V}_{\mathbf{A}}}{\hat{\kappa}_{\mathbf{A}}}.$$

An algorithm for efficiently drawing uniformly from within a hyperellipsoid is found in the Appendix which requires minimal computing time even for large values of R . Volume $V_{\mathbf{A}_2}$ is calculated rather than estimated, and the Appendix shows a recursive algorithm returning $\log(V_{\mathbf{A}_2})$ with minimal computing time even for very high K . Thus, the calculation of $\hat{\pi}$ and $V_{\mathbf{A}_2}$, and consequently $\hat{V}_{\mathbf{A}}$, is achieved with low computational effort and high precision.

3 Applications

In this section the developed approach is applied to several important classes of models. The estimator is compared to those existing methods that have consistently shown excellent performance within the respective class of models in earlier comparative studies. These are the approaches of Chib [5], along with importance sampling and bridge sampling [21]. Chib's method exploits the basic marginal likelihood equation

$$\hat{p}_{\text{CH}}(\mathbf{y}|\mathcal{M}) = \frac{p^*(\theta^*|\mathbf{y})}{\hat{p}(\theta^*|\mathbf{y})},$$

where θ^* is a high density point for which the posterior mode is used, and will be referred to throughout this section. Thus an estimate of the marginal likelihood requires an estimate of the posterior density at θ^* , and we follow the estimation technique for the posterior ordinate $\hat{p}(\theta^*|\mathbf{y})$ as outlined in [5] throughout this section. Chib's method does not require any input other than draws from the Gibbs sampler, however, for each variable block within the Gibbs sampler one needs to re-run a reduced sampler. In particular, for hierarchical and multi-level models, this incurs a certain degree of burden on the analyst. Furthermore, Chib's estimator as outlined above is applicable in a Gibbs setting only, while a generalization of this approach also allows estimation of the marginal likelihood based on output from the Metropolis–Hastings algorithm [6].

Importance sampling and bridge sampling both require as an input an importance density $q(\theta)$, which approximates the posterior distribution. The degree to which this approximation successfully mimics the posterior distribution has substantial impact on the efficiency of the estimators. For the importance sampling estimator $m = 1, \dots, M$ draws $\theta^{(m)}$ are obtained from $q(\theta)$, and an estimator is then given by averaging over

the non-normalized posterior values weighted by the importance density ordinates such that

$$\hat{p}_{\text{IS}}(\mathbf{y}|\mathcal{M}) = \frac{1}{M} \sum_{m=1}^M \frac{p^*(\boldsymbol{\theta}^m|\mathbf{y})}{q(\boldsymbol{\theta}^m)}.$$

If $q(\boldsymbol{\theta})$ does not have fatter tails than the posterior density, the importance density estimator can become unstable and resulting in high standard errors. The bridge sampling technique somewhat alleviates this problem by weakening the demands on the importance density with respect to its tail behavior [12]. The bridge sampling estimator of [21] is the limit of the recurrence

$$\hat{p}_{\text{BS}}^{(t)} = \frac{M_1^{-1} \sum_{m_1=1}^{M_1} \frac{p^*(\boldsymbol{\theta}^{(m_1)}|\mathbf{y})}{M_1 q(\boldsymbol{\theta}^{(m_1)}) + M_2 p^*(\boldsymbol{\theta}^{(m_1)}|\mathbf{y})/\hat{p}_{\text{BS}}^{(t-1)}}}{M_2^{-1} \sum_{m_2=1}^{M_2} \frac{q(\boldsymbol{\theta}^{(m_2)})}{M_2 q(\boldsymbol{\theta}^{(m_2)}) + M_2 p^*(\boldsymbol{\theta}^{(m_2)}|\mathbf{y})/\hat{p}_{\text{BS}}^{(t-1)}}},$$

where $\boldsymbol{\theta}^{(m_1)}$ refers to the m_1^{th} of M_1 draws of the importance density $q(\boldsymbol{\theta})$, while $\boldsymbol{\theta}^{(m_2)}$ refers to the m_2^{th} of M_2 draws of the posterior density. Throughout this section we use $M_1 = M_2 = M$.

The problem of finding an adequate importance density is not trivial, in particular for high-dimensional densities tuning of the importance density can become very time consuming. An approach to define the importance density $q(\boldsymbol{\theta})$ based on output from the Gibbs sampler is found in [8], where $\boldsymbol{\theta}$ is split into D blocks $(\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_D)$ and based on the conditional distributions of these parameter blocks a mixture importance density is constructed as a byproduct of the Gibbs sampler

$$q(\boldsymbol{\theta}) = \frac{1}{S} \sum_{s=1}^{l_S} \prod_{d=1}^D p^{(s)}(\boldsymbol{\theta}_d | \boldsymbol{\theta}_{j < d}^{(s)}, \boldsymbol{\theta}_{j > d}^{(s-1)}, \mathbf{y}), \quad (3.1)$$

where each mixture component $\prod_{d=1}^D p^{(s)}(\boldsymbol{\theta}_d | \boldsymbol{\theta}_{j < d}^{(s)}, \boldsymbol{\theta}_{j > d}^{(s-1)}, \mathbf{y})$ refers to the product of conditional densities from which $\boldsymbol{\theta}$ was drawn in the s^{th} iteration of the Gibbs sampler, and $l_1, \dots, l_S$ refer to a series of S Gibbs iterations chosen by the analyst, see [9, 11] for details. Then $q(\boldsymbol{\theta})$ converges to the posterior density when $S \rightarrow \infty$. Throughout this section an importance density as defined in (3.1) is used, unless stated otherwise.

Besides comparing point estimates, the expected relative mean-squared error (RE^2) gives evidence on the stability of the estimators, and respective estimation techniques for RE^2 are summarized in [12]. However, recent research investigated the accuracy of these estimation techniques [2] and calculated numerical standard errors of the outlined estimators for 500 independent runs of the MCMC sampler. The analysis shows that there is significant risk of overestimating or underestimating the standard errors under given techniques when these require as an input the specific lag at which the autocorrelation of the MCMC chain tapers off (which is the case for all estimators here except the importance sampling estimator $\hat{p}_{\text{IS}}$). However, since this paper aims to complete a pure comparison of the presented estimator to the existing ones, we abstract from this discussion and estimate RE^2 without relying on a measure of the autocorrelation by the bootstrap method, even though this increases computational effort significantly (see e.g. [18] for details about the bootstrap method). For each bootstrap replication of $\hat{p}_{\text{A}}$, we draw with replacement L out of the L posterior draws and carry out the estimation of $\hat{p}_{\text{A}}$ anew. The same applies for bootstrap replications of $\hat{p}_{\text{CH}}$ when the model is estimated by a one-block Gibbs sampler. For models requiring multi-block samplers only the high density point $\boldsymbol{\theta}^*$ and the conditional density of the first variable block can be calculated from the bootstrap sample, while for each of the remaining variable blocks the respectively reduced Gibbs sampler is required to be carried out anew to obtain the ordinate of the particular conditional density. Similarly, M out of the M importance draws are drawn to obtain a bootstrap replication of $\hat{p}_{\text{IS}}$, and these importance draws are combined with M out of the M selected posterior draws for $\hat{p}_{\text{BS}}$. Reported standard errors are calculated from 200 bootstrap replications each.

Defining the quantities L , M , S and R to yield comparable computational time with respect to the calculation of the different estimators, and thus allow for a fair comparison, would require tuning these quantities

in each application. However, we set these quantities in such way that $\hat{p}_A$ and $\hat{p}_{CH}$ show about the same computation time, while even under low values of S importance sampling and bridge sampling require significantly more time. In all subsequent applications we obtain $L = 100,000$ draws from the Gibbs sampler as input for $\hat{p}_A$ and $\hat{p}_{CH}$. Additionally, we set $R = 100,000$ to ensure precise estimation of $\hat{\pi}$ considering its standard error is $\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{R}}$. However, to ensure precise estimation even for high-dimensional distributions with very non-elliptical contours and potentially small values of $\hat{\pi}$, we define as a generic rule the lower bound of draws to be $\bar{r} = 1,000$. M and S are set in the respective application.

3.1 Probit and logit

The most exploited data with respect to comparing estimation techniques of the marginal likelihood are the prostatic nodal involvement data, see e.g. [5, 7, 14]. These data contain $N = 53$ patients with cancer of the prostate and the binary response indicates whether the cancer has spread to the surrounding lymph nodes. Potential explanatory variables are the patient's age at time of diagnosis (x_1), the level of serum acid phosphate (x_2), the binary result of an X-ray examination (x_3), the size of the tumor (x_4), with 0 for small and 1 for large, and the pathological grade of the tumor (x_5), with 0 if less serious and 1 if more severe. All of the above cited articles agree on the addition of an intercept for modeling the binary outcome.

Each of the aforementioned three instances in literature where these data are exploited for marginal likelihood estimation assume a different model specification and prior setting. We use a probit link [5, 7] as well as a logit link function [14], given by

$$\begin{aligned} \text{probit: } \Pr(y_i = 1|\boldsymbol{\beta}) &= \Phi(\mathbf{x}_i\boldsymbol{\beta}), \\ \text{logit: } \Pr(y_i = 1|\boldsymbol{\beta}) &= \frac{\exp(\mathbf{x}_i\boldsymbol{\beta})}{1 + \exp(\mathbf{x}_i\boldsymbol{\beta})}, \end{aligned}$$

where $\mathbf{x}_i$ is a row vector containing the explanatory variables of respondent i , the column vector $\boldsymbol{\beta}$ holds the corresponding coefficients, and $\Phi(\cdot)$ refers to the cumulative distribution function of a standard normal distribution. We use the prior setting in [7] for the priors in the probit model specification. Therefore, we use a normal prior with mean 0 and standard deviation 10 for the intercept, and normal priors with mean 0.75 and a standard deviation of 5 for the other explanatory variables. For achieving comparability between the probit and logit models, in the logit specification all parameters of the normal prior are scaled by 1.6 [14]. The probit models are estimated as outlined in [1], while the logit models are estimated by auxiliary mixture sampling as explained in [13] and [15].

Since the presented approach exploits the posterior distribution as importance density in a similar way as the harmonic mean estimator does, we also report $\hat{p}_{HM}$ in this section to demonstrate the substantial impact the proposed limitation of the region of integration to set $\mathbf{A}$ has. We set the number of draws of the importance density to $M = 20,000$ for $\hat{p}_{IS}$, and combine these draws with randomly selected ($M = 20,000$) posterior draws for the calculation of $\hat{p}_{BS}$. The mixture importance density is constructed from $S = 100$ components as outlined in (3.1).

Based on three independent runs of the MCMC sampler for each model, Table 1 displays the outcome of the different estimators of the marginal likelihood on the logarithmic scale. Firstly, for models of relatively low dimension, i.e. $\mathcal{M}_1 - \mathcal{M}_7$ and $\mathcal{M}_{10} - \mathcal{M}_{16}$, we see that $\hat{p}_{IS}$, $\hat{p}_{BS}$, and $\hat{p}_{CH}$ have the smallest standard errors, while the accuracy of $\hat{p}_A$ clearly allows for its application without increasing the risk of false model choice to a relevant degree. Secondly, we see that the advantage of existing methods vanishes when the dimension of the investigated model increases. In particular, for $\mathcal{M}_9$ and $\mathcal{M}_{18}$ we see that $\hat{p}_A$ outperforms the other estimators except the bridge sampling estimator $\hat{p}_{BS}$. Thus the estimator $\hat{p}_A$ is more robust to an increase in the complexity of the model than importance sampling and Chib's method in this example. This statement is also supported by consideration of the standard errors from the model of lowest dimension, $\mathcal{M}_1$ and the model of highest dimension, $\mathcal{M}_{18}$, which have increased by factor 5 or more for all other estimators, while the standard errors of $\hat{p}_A$ have not responded to an increase of the dimensionality within each group of models with the same link function.

Variables	Model	Probit link					Model	Logit link				
		$\ln(\hat{p}_{HM})$	$\ln(\hat{p}_{IS})$	$\ln(\hat{p}_{BS})$	$\ln(\hat{p}_{CH})$	$\ln(\hat{p}_A)$		$\ln(\hat{p}_{HM})$	$\ln(\hat{p}_{IS})$	$\ln(\hat{p}_{BS})$	$\ln(\hat{p}_{CH})$	$\ln(\hat{p}_A)$
$1, x_1$	$\mathcal{M}_1$	-37.229	-43.814	-43.815	-43.815	-43.815	$\mathcal{M}_{10}$	-36.824	-43.763	-43.761	-43.766	-43.761
		(0.316)	(0.001)	(0.000)	(0.001)	(0.004)		(0.071)	(0.001)	(0.001)	(0.002)	(0.006)
		-36.921	-43.815	-43.816	-43.815	-43.822		-37.437	-43.763	-43.761	-43.758	-43.743
		(0.091)	(0.001)	(0.000)	(0.001)	(0.004)		(0.228)	(0.001)	(0.001)	(0.002)	(0.008)
$1, \ln(x_2)$	$\mathcal{M}_2$	-36.945	-43.816	-43.816	43.814	-43.814	$\mathcal{M}_{11}$	-38.339	-43.761	-43.762	-43.761	-43.745
		(0.083)	(0.001)	(0.001)	(0.001)	(0.005)		(0.809)	(0.001)	(0.001)	(0.002)	(0.006)
		-34.931	-38.604	-38.604	-38.603	-38.607		-35.415	-38.420	-38.422	-38.424	-38.423
		(0.199)	(0.001)	(0.000)	(0.001)	(0.004)		(0.424)	(0.002)	(0.001)	(0.002)	(0.006)
$1, x_3$	$\mathcal{M}_3$	-34.743	-38.605	-38.605	-38.606	-38.606	$\mathcal{M}_{12}$	-35.113	-38.418	-38.421	-38.423	-38.424
		(0.065)	(0.001)	(0.001)	(0.001)	(0.003)		(0.163)	(0.002)	(0.001)	(0.002)	(0.006)
		-34.709	-38.604	-38.604	-38.601	-38.603		-35.180	-38.419	-38.422	-38.422	-38.419
		(0.051)	(0.001)	(0.001)	(0.002)	(0.003)		(0.308)	(0.002)	(0.001)	(0.002)	(0.007)
$1, x_4$	$\mathcal{M}_4$	-31.890	-35.974	-35.976	-35.976	-35.976	$\mathcal{M}_{13}$	-32.696	-35.837	-35.838	-35.836	-35.837
		(0.077)	(0.001)	(0.001)	(0.002)	(0.003)		(0.445)	(0.002)	(0.001)	(0.002)	(0.004)
		-31.679	-35.973	-35.975	-35.977	-35.974		-32.250	-35.837	-35.838	-35.848	-35.840
		(0.043)	(0.001)	(0.001)	(0.002)	(0.003)		(0.156)	(0.002)	(0.001)	(0.002)	(0.005)
$1, x_5$	$\mathcal{M}_5$	-32.071	-35.976	-35.977	-35.974	-35.977	$\mathcal{M}_{14}$	-32.124	-35.839	-35.837	-35.843	-35.824
		(0.123)	(0.001)	(0.001)	(0.002)	(0.003)		(0.143)	(0.001)	(0.001)	(0.002)	(0.004)
		-33.724	-37.875	-37.874	-37.873	-37.877		-33.913	-37.760	-37.758	-37.759	-37.759
		(0.080)	(0.001)	(0.001)	(0.002)	(0.004)		(0.123)	(0.001)	(0.001)	(0.002)	(0.004)
$1, \ln(x_2), x_4$	$\mathcal{M}_6$	-34.159	-37.874	-37.873	-37.875	-37.871	$\mathcal{M}_{15}$	-33.905	-37.760	-37.759	-37.770	-37.767
		(0.339)	(0.001)	(0.001)	(0.002)	(0.003)		(0.128)	(0.001)	(0.001)	(0.002)	(0.004)
		-33.593	-37.873	-37.874	-37.872	-37.873		-34.199	-37.757	-37.758	-37.761	-37.757
		(0.088)	(0.001)	(0.001)	(0.002)	(0.004)		(0.230)	(0.002)	(0.001)	(0.002)	(0.004)
$1, \ln(x_2), x_3, x_4$	$\mathcal{M}_7$	-35.330	-39.735	-39.736	-39.735	-39.736	$\mathcal{M}_{16}$	-35.588	-39.674	-39.673	-39.671	-39.673
		(0.061)	(0.001)	(0.001)	(0.001)	(0.004)		(0.095)	(0.001)	(0.001)	(0.002)	(0.004)
		-35.674	-39.736	-39.736	-39.738	-39.737		-36.230	-39.671	-39.674	-39.671	-39.684
		(0.154)	(0.001)	(0.001)	(0.001)	(0.004)		(0.352)	(0.002)	(0.001)	(0.002)	(0.004)
$1, \ln(x_2), x_3, x_4, x_5$	$\mathcal{M}_8$	-35.717	-39.735	-39.736	-39.735	-39.732	$\mathcal{M}_{17}$	-35.594	-39.673	-39.672	-39.673	-39.674
		(0.215)	(0.001)	(0.000)	(0.001)	(0.003)		(0.072)	(0.001)	(0.001)	(0.002)	(0.004)
		-32.237	-36.794	-36.795	-36.800	-36.797		-33.320	-36.589	-36.588	-36.586	-36.585
		(0.389)	(0.001)	(0.001)	(0.002)	(0.004)		(0.841)	(0.005)	(0.002)	(0.003)	(0.006)
$1, x_1, \ln(x_2), x_3, x_4, x_5$	$\mathcal{M}_9$	-34.155	-36.794	-36.796	-36.796	-36.797	$\mathcal{M}_{18}$	-32.115	-36.594	-36.588	-36.581	-36.588
		(1.075)	(0.002)	(0.001)	(0.002)	(0.004)		(0.193)	(0.003)	(0.002)	(0.003)	(0.006)
		-32.171	-36.797	-36.796	-36.797	-36.795		-32.474	-36.588	-36.586	-36.578	-36.574
		(0.324)	(0.002)	(0.001)	(0.002)	(0.004)		(0.288)	(0.002)	(0.002)	(0.003)	(0.009)
		-30.159	-35.201	-35.198	-35.201	-35.200		-31.681	-34.875	-34.881	-34.876	-34.879
		(0.734)	(0.004)	(0.003)	(0.003)	(0.004)		(1.118)	(0.005)	(0.003)	(0.005)	(0.004)
		-30.296	-35.202	-35.204	-35.210	-35.204		-29.535	-34.877	-34.880	-34.882	-34.873
		(0.558)	(0.004)	(0.002)	(0.004)	(0.004)		(0.139)	(0.007)	(0.002)	(0.005)	(0.007)
		-28.818	-35.197	-35.203	-35.194	-35.201		-28.885	-34.882	-34.877	-34.855	-34.883
		(0.137)	(0.004)	(0.002)	(0.003)	(0.004)		(0.101)	(0.004)	(0.003)	(0.005)	(0.006)
		-29.504	-36.493	-36.491	-36.501	-36.523		-29.504	-36.493	-36.491	-36.501	-36.523
		(0.147)	(0.009)	(0.002)	(0.004)	(0.004)		(0.094)	(0.015)	(0.003)	(0.006)	(0.006)
		-30.515	-36.889	-36.883	-36.885	-36.890		-29.452	-36.512	-36.493	-36.502	-36.468
		(0.614)	(0.004)	(0.002)	(0.004)	(0.004)		(0.131)	(0.007)	(0.003)	(0.006)	(0.005)
		-30.307	-36.888	-36.883	-36.876	-36.893		-29.731	-36.505	-36.490	-36.484	-36.452
		(0.638)	(0.005)	(0.003)	(0.004)	(0.004)		(0.099)	(0.006)	(0.003)	(0.007)	(0.008)
		-29.553	-41.211	-41.200	-41.199	-41.198		-30.963	-40.746	-40.746	-40.736	-40.730
		(0.244)	(0.007)	(0.003)	(0.005)	(0.005)		(0.630)	(0.014)	(0.004)	(0.008)	(0.006)
		-29.376	-41.200	-41.199	-41.193	-41.195		-30.138	-40.698	-40.727	-40.750	-40.728
		(0.165)	(0.007)	(0.003)	(0.005)	(0.004)		(0.208)	(0.035)	(0.005)	(0.012)	(0.007)
		-29.481	-41.198	-41.198	-41.198	-41.196		-30.014	-40.742	-40.730	-40.721	-40.697
		(0.182)	(0.007)	(0.003)	(0.005)	(0.005)		(0.218)	(0.022)	(0.005)	(0.008)	(0.007)

Table 1: Nodal involvement data; logarithm of different estimators of the marginal likelihood. $\hat{p}_{HM}$ refers to the harmonic mean estimator, importance sampling and bridge sampling are referenced by $\hat{p}_{IS}$ and $\hat{p}_{BS}$, and $\hat{p}_{CH}$ refers to Chib's method. The estimator proposed in this paper is referenced as $\hat{p}_A$. Relevant standard errors in parentheses, chosen model in bold face, results from three independent MCMC runs per model specification are reported.

3.2 Mixture of normals

A general formulation of a mixture distribution is given by

$$y_i \sim \sum_{s=1}^S \eta_s \cdot f(y_i | \theta_s).$$

Denoting the indicator holding the assignment of observation y_i to a mixture component by I_i , where this indicator takes labels $s = 1, \dots, S$, the model likelihood and the prior of the series $I^N = (I_1, \dots, I_N)$ are invariant to re-labelling. When the prior of the parameters θ_s is invariant too, then the posterior distribution of a mixture model is multimodal with respect to the $S!$ permutations of the indicator labels. This is a challenge for marginal likelihood estimation of mixture models [11].

Parameter estimation of mixture models through MCMC sampling frequently suffers from the un-identifiability of the indicator labels, which can result in undesired label switching. However, several techniques have been proposed to address the label switching problem [10, 26]. We first consider a posterior sample $\theta^{L,1} = (\theta^{(1)}, \dots, \theta^{(L)})$ in which all draws belong to one permutation of the indicator labels. Since the marginal likelihood is the normalizing constant of the whole unconstrained posterior distribution, not only of the subspace belonging to one permutation, calculating $\hat{p}_A$ from only $\theta^{L,1}$ would result in an incomplete estimate of the marginal likelihood. Thus we define a second sample of the posterior distribution $\theta^{L,S!} = (\theta^{(1)}, \dots, \theta^{(L)})$ in which the draws represent a balanced sample from all of the $S!$ permutations of the labels. In this section the random permutation sampler of [10] is applied to obtain $\theta^{L,S!}$, while other approaches exploring all modes of the posterior distribution could be used as well (see e.g. [4]).

A derivative of Algorithm I allowing the estimation of the marginal likelihood for multimodal mixture models is then found by first estimating ρ , θ^* , D , A , and thus $\hat{\pi}$ and $\hat{V}_A$, from $\theta^{L,1}$. Secondly based on the thereby defined A integral $\hat{\kappa}_A$ is subsequently estimated from $\theta^{L,S!}$. Thus $\hat{p}_A$ can be implemented for mixture models as generic as for unimodal distributions, however, two runs of the MCMC sampler are required. Nevertheless, since the construction of the importance density from (3.1) requires output from a Gibbs sampler run exploring all modes of the posterior as well, while usually a sample $\theta^{L,1}$ is additionally needed for parameter estimation, the requirement of two posterior samples does not mean additional burden compared to importance density based techniques in practice.

Here we demonstrate the function of the proposed estimator by applying it to the simulation example used in [10] and [16]. Two examples of a mixture of two normal distributions are defined. The first example is defined in such a way that the two states are easily separated by the Gibbs sampler and set $\eta_1 = \eta_2 = 0.5$, $\mu_1 = 1$, $\mu_2 = 2$, $\sigma_1^2 = 0.01$, and $\sigma_2^2 = 0.05$. The second example is defined with the same coefficients as the first example, but with $\sigma_1^2 = 0.1$ and $\sigma_2^2 = 0.15$, resulting in significant overlap of the distributions. We simulate 1,000 observations from each of the two mixture distributions, and prepare estimation by defining natural conjugate priors for μ_s and σ_s^2 . An inverse gamma distribution is selected as prior for σ_s^2 where the scale parameter and the shape parameter equal 3, imposing a vague prior considering the magnitudes of the true parameters. Then the prior of μ_s is a normal distribution with mean 0 and variance $100\sigma_s^2$. Vector (η_1, η_2) follows a priori a Dirichlet distribution with parameters $a_{\eta_1} = a_{\eta_2} = 5$. The Gibbs sampler as outlined in [23] is then applied to obtain the posterior sample $\theta^{L,1}$, and combined with the random permutation step to obtain $\theta^{L,S!}$.

The approach presented here is compared to importance sampling and bridge sampling where the importance density in (3.1) is constructed from $S = 100$ components, and M is again set to 20,000. As discussed in the literature (see e.g. [12]) Chib's method does not perform comparably well when applied to multimodal models as it does for unimodal models, and we omit reporting the outcomes of this estimator here. The left plot in Figure 2 displays a scatter plot of σ_1^2 and σ_2^2 from all $l = 1, \dots, L$ draws of the series $\theta^{L,S!}$ of the MCMC sampler (black, in the background) overlaid by those draws for which $\theta^{(l)} \in A$ (in the front). We see that the posterior distribution of σ_1^2 and σ_2^2 under the two permutations overlap to a significant extent and that set A is centered around one of the two permutations of the component labels.

Table 2 displays the results of the estimators from three independent runs of the Gibbs sampler. For the specification allowing a clear separation of the labels, importance sampling as well as bridge sampling show

σ_1^2, σ_2^2	$\ln(\hat{p}_{IS})$	$\ln(\hat{p}_{BS})$	$\ln(\hat{p}_A)$
0.01, 0.05	-376.606	-376.606	-376.600
	(0.001)	(0.001)	(0.007)
	-376.606	-376.606	-376.608
	(0.001)	(0.000)	(0.007)
	-376.605	-376.605	-376.593
0.1, 0.15	(0.001)	(0.001)	(0.007)
	-914.847	-914.839	-914.832
	(0.005)	(0.003)	(0.006)
	-914.846	-914.835	-914.838
	(0.008)	(0.002)	(0.006)
	-914.852	-914.838	-914.827
	(0.007)	(0.003)	(0.007)

Table 2: Simulated data of mixture of two normals; logarithm of different estimators for two specifications of the true parameters σ_1^2 and σ_2^2 . Importance sampling and bridge sampling are referenced by $\hat{p}_{IS}$ and $\hat{p}_{BS}$, and $\hat{p}_A$ refers to the estimator proposed in this paper. Relevant standard errors in parentheses, results from three independent MCMC runs per model are reported.

Figure 2: Scatter plot of two coefficients from all $l = 1, \dots, L$ draws of the MCMC sampler (black, in the background) overplotted by those draws of the coefficients for which $\theta^{(l)} \in A$ (in the front); left side: σ_1^2 and σ_2^2 (from series θ^{L, S_1}) from Section 3.2; right side: μ and σ^2 of the model with $K = 28$ from Section 3.3.

very low standard errors and outperform the estimator presented in this article. However, in absolute terms the standard errors of $\hat{p}_A$ are still very small and would usually have no practical impact on model choice. For the parameter specification with overlap of the distributions of σ_1^2 and σ_2^2 bridge sampling again has the smallest standard errors, while the magnitudes have increased noticeably. Additionally, we see that the standard errors of importance sampling and the method presented here are on the same scale.

We interpret the results from this simulation study as evidence that $\hat{p}_A$ provides unbiased and stable estimates when applied to multimodal models. Considering the significant challenges associated with marginal likelihood estimation of such mixture models, and the limited applicability of Chib's estimator to them, the presented approach can significantly contribute to making the task of model selection less cumbersome for this important class of models.

3.3 Random intercept probit

In this subsection the sensitivity of the discussed estimators of the marginal likelihood to an increase of the dimensionality of the candidate models is explored. This paper provides for the first time a comparison of the

discussed estimation techniques for a high-dimensional unit-level model, and discloses the shortcomings of existing approaches. As one instance of a comparative study exploring the existing techniques with respect to a unit-level model, Frühwirth-Schnatter and Wagner [14] fit a random intercept logit model, with up to $K = 25$ coefficients, to agricultural data and compare the estimates from Chib's method, as well as from importance and bridge sampling under the mixture importance density from (3.1). In this section we increase the number of model dimensions in five applications up to $K = 142$ to demonstrate the extraordinary stability of the presented estimator in comparison to the existing approaches.

The applications of this section exploit data on reading proficiency in US schools. Data stem from the "Program for International Student Assessment" (PISA), and here data from the year 2009 is used as provided in [25]. Reading proficiency is measured in terms of a continuous score, however, since limited dependent variable models have shown to be a bigger challenge with respect to marginal likelihood estimation, the score is censored in this example. Consequently, the observed dependent variable $\mathbf{y}$ is defined according to whether a student exceeded the median score of all students (1) or fell behind it (0). As explanatory variables we use a student's grade (x_1 ; a number between 9 and 11), gender (x_2 ; 1 for female, 0 for male), whether a student's family has lived in the USA for at least three generations (x_3 ; 1 for yes, 0 for no), the index of economic, social and cultural status (x_4 ; a continuous value centered at 0), and an indicator whether a student has stated they enjoy reading (x_5 ; a continuous value centered at 0). Skrongdal and Rabe-Hesketh [24] estimate an older version of the PISA data and add a random intercept for each school to capture overdispersion. In this application we use a probit link function, such that the model is given by

$$y_{it} = \begin{cases} 1 & \text{if } z_{it} > 0, \\ 0 & \text{else,} \end{cases} \quad z_{it} = \mathbf{x}_i \boldsymbol{\beta} + \gamma_t + \varepsilon_{it}, \quad \varepsilon_{it} \sim N(0, 1), \quad \gamma_t \sim N(\mu, \sigma^2), \quad (3.2)$$

where $t = 1, \dots, T$ refers to the school of student i with her or his characteristics being held by the row vector $\mathbf{x}_i$. We define a normal prior for $\boldsymbol{\beta}$ such that $\boldsymbol{\beta} \sim N(\mathbf{0}, 100 \times \mathbf{I})$, where $\mathbf{I}$ refers to the identity matrix with five columns. We specify a natural conjugate prior for μ and σ^2 . The variance of the random intercept is estimated to be 0.28 in [24], and we specify an inverted Gamma distribution as its prior with $\sigma^2 \sim G^{-1}(2, 0.5)$ having a median of around 1.2 and thus being vague enough for this application. The prior of μ is then a normal distribution with mean 0 and variance $100\sigma^2$. The Gibbs sampler is outlined in the Appendix.

This model is estimated for five different data sets. The first data set contains all 21 public schools in the PISA sample from the Northeast of the US, the second data set contains the 28 schools of the US region West, the third contains the 36 schools from the Midwest, and the fourth data set contains the 50 schools from the remaining region, South. For the fifth data set we combine these regions and estimate reading proficiency for all 135 public schools in the sample. Considering the model estimates an individual intercept for each school, the total number of coefficients K is the number of schools in the respective data set plus 7.

The estimation techniques explored in this paper are the method presented in this article, Chib's method, and importance and bridge sampling. As discussed earlier, the mixture importance distribution from equation (3.1) approaches the posterior distribution for $S \rightarrow \infty$, and $S = 100$ performed well for the low-dimensional models from the proceeding sections. However, in a hierarchical model each of the S components $\prod_{d=1}^D p^{(s)}(\boldsymbol{\theta}_d | \boldsymbol{\theta}_{j < d}^{(s)}, \boldsymbol{\theta}_{j > d}^{(s-1)}, \mathbf{y})$ of the mixture importance density becomes a very narrow and almost spike shaped distribution conditional on a draw of the latent variables $\mathbf{Z} = (z_{11}, \dots, z_{NT})'$. As a result the final importance density has hardly any mass in between its S spikes and clearly fails the requirement to have fatter tails than its target distribution, with negative impact on the stability of $\hat{p}_{IS}$.

The hedgehog-like shape of the importance density may also induce problems for bridge sampling. When constructing the importance density from (3.1) in an unsupervised manner, it is likely that some of the posterior draws $\boldsymbol{\theta}^{(1)}, \dots, \boldsymbol{\theta}^{(M_2)}$ intended to be used during bridge sampling have originally been drawn from one of the S components of the mixture importance density. If a draw $\boldsymbol{\theta}^{(m_2)}$ has been drawn from one of the S components of the importance density (i.e. from the center region of one of the spikes), then the magnitude of $q(\boldsymbol{\theta}^{(m_2)})$ is exceedingly high compared to an importance density constructed from components M_2 which none of the posterior draws have been drawn from. As a result, the sum in the denominator of the bridge sampling estimator $\sum_{m_2=1}^{M_2} \frac{q(\boldsymbol{\theta}^{(m_2)})}{M_2 q(\boldsymbol{\theta}^{(m_2)}) + M_2 p^*(\boldsymbol{\theta}^{(m_2)} | \mathbf{y}) / \hat{p}_{IS}^{(t-1)}}$ is dominated by these high values of $q(\boldsymbol{\theta}^{(m_2)})$ with neg-

ative effects on the stability of the estimator. To avoid this source of instability, we split the posterior sample and construct the importance density from the first half of the L posterior draws $\boldsymbol{\theta}^{(1)}, \dots, \boldsymbol{\theta}^{(L/2)}$ and select the M_2 posterior draws for usage during bridge sampling from the second half $\boldsymbol{\theta}^{(L/2)}, \dots, \boldsymbol{\theta}^{(L)}$.

Following the preceding discussion about possible drawbacks of the mixture importance density when applied to unit-level models, we compare it to another approach for the unsupervised construction of an importance density. Likewise as for (3.1) $\boldsymbol{\theta}$ is split into D blocks $(\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_D)$, and the marginal distribution of each block $q(\boldsymbol{\theta}_d)$ is estimated from the posterior draws. The product of the estimated marginal densities is then used as importance density (see e.g. [7]):

$$q(\boldsymbol{\theta}) = \prod_{d=1}^D \hat{q}(\boldsymbol{\theta}_d). \quad (3.3)$$

Here, the marginal distribution of μ is estimated by a normal distribution with its mean being the mean of the posterior draws $(\mu^{(1)}, \dots, \mu^{(L)})$, and its variance as the variance of the posterior draws. The marginal distribution of σ^2 is approximated by a log-normal distribution with its mean as the mean of the log-posterior draws $(\log(\sigma^{2,(1)}), \dots, \log(\sigma^{2,(L)}))$, and its variance as the variance of this series. Likewise, $q(\boldsymbol{\theta}_d)$ of $\boldsymbol{\beta}$ and of the random intercepts $\gamma_1, \dots, \gamma_T$ are approximated by normal distributions with their means and their variances estimated from the respective posterior draws.

As this construction of the importance density decreases its ability of drawing from high-density regions of the posterior distribution compared to the mixture importance density approach, we set $M = 1,000,000$. For the mixture importance density, we set $S = 1,000$ and $M = 50,000$. Considering these demanding settings for both importance densities, importance sampling and bridge sampling require long computational runtime, while calculation of $\hat{p}_A$ is much faster.

Table 3 displays the results of the comparative study. While the magnitudes of the estimates cannot be compared between the different values of K as these are related to different data sets, the standard errors allow conclusions about the sensitivity of the respective estimators to an increase of the number of coefficients of the underlying model. Firstly, we see that constructing the importance density as in (3.3) can be ruled out for this model and we skip reporting estimates beyond $K = 28$. Secondly, we see from the high standard errors that, even for this relatively low value of K , importance sampling applying the mixture importance density produces less reliable estimates than the remaining three techniques, and this reliability worsens with an increase in K . However, up to $K = 57$ all three remaining techniques provide estimates with relatively low standard errors, while $\hat{p}_A$ is by far the most stable for all values of K . Thirdly, in practice one would usually wish to analyze all public schools in the US in the same model, which is marked in Table 3 through $K = 142$. For this model we see that all existing techniques are not able to produce stable estimates, while in contrast $\hat{p}_A$ provides outstandingly stable estimates with low standard errors showing only very limited sensitivity to the increase in the number of model coefficients from $K = 28$ to $K = 142$.

The outstanding stability of $\hat{p}_A$ is a result of directly addressing the magnitude of the non-normalized posterior as the criterion to qualify for $\mathbf{A}$. For importance density based methods the importance density has to approximate the posterior distribution in all K dimensions, while $\hat{p}_A$ only relies on the univariate aggregate $p^*(\boldsymbol{\theta}|\mathbf{y})$. This is a significant advantage with respect to the stability of the estimates. The right plot in Figure 2 compares all posterior draws of the coefficients μ and σ^2 of the data with $K = 28$, to those draws lying in $\mathbf{A}$, and demonstrates that the posterior distribution has very non-elliptical contours obviously not hampering the stability of $\hat{p}_A$.

4 Conclusion and discussion

This article develops a new approach to marginal likelihood estimation and compares it to existing estimation techniques. This comparison reveals two major advantages of the presented approach. Firstly, it requires only a sample of posterior draws as an input from the analyst, and these draws can come from any sampling scheme, such as Gibbs sampling or Metropolis–Hastings sampling. Furthermore, and in contrast to some widely used earlier approaches, only draws from one run of the MCMC sampler are required, except for

US region	K	$\ln(\hat{p}_{IS}^1)$	$\ln(\hat{p}_{BS}^1)$	$\ln(\hat{p}_{IS}^2)$	$\ln(\hat{p}_{BS}^2)$	$\ln(\hat{p}_{CH})$	$\ln(\hat{p}_A)$
Northeast	28	-396.244	-376.876	-363.203	-363.139	-363.086	-363.136
		(0.659)	(0.537)	(0.076)	(0.012)	(0.054)	(0.014)
		-382.218	-364.083	-363.028	-363.155	-363.183	-363.167
		(1.009)	(0.862)	(0.185)	(0.012)	(0.049)	(0.010)
West	35	-383.503	-364.308	-363.204	-363.169	-363.168	-363.145
		(0.641)	(0.540)	(0.176)	(0.012)	(0.051)	(0.009)
				-503.501	-503.159	-503.219	-503.199
				(0.219)	(0.024)	(0.090)	(0.017)
Midwest	43			-503.781	-503.180	-503.091	-503.174
				(0.262)	(0.024)	(0.097)	(0.012)
				-503.828	-503.125	-503.003	-503.189
				(0.150)	(0.023)	(0.100)	(0.010)
South	57			-667.515	-667.026	-667.117	-667.033
				(0.234)	(0.028)	(0.127)	(0.013)
				-667.364	-667.012	-667.014	-667.031
				(0.243)	(0.029)	(0.123)	(0.012)
all	142			-667.241	-666.962	-667.127	-667.036
				(0.471)	(0.027)	(0.095)	(0.016)
				-891.596	-892.433	-892.246	-892.427
				(0.779)	(0.054)	(0.158)	(0.012)
				-893.633	-892.432	-892.628	-892.432
				(0.536)	(0.058)	(0.255)	(0.013)
				-893.341	-892.568	-892.370	-892.421
				(0.475)	(0.053)	(0.184)	(0.016)
				-2366.396	-2352.958	-2350.232	-2351.877
				(0.561)	(1.001)	(0.447)	(0.017)
				-2365.387	-2354.952	-2351.547	-2351.920
				(0.637)	(1.815)	(1.264)	(0.016)
				-2364.521	-2350.454	-2352.491	-2351.913
				(0.689)	(0.539)	(1.338)	(0.021)

Table 3: PISA data; logarithm of different estimators and for five different data sets. Importance sampling and bridge sampling using an importance density constructed as product of estimated marginal densities as in (3.3) are referenced by $\hat{p}_{IS}^1$ and $\hat{p}_{BS}^1$, while $\hat{p}_{IS}^2$ and $\hat{p}_{BS}^2$ use a mixture importance density constructed as in (3.1), $\hat{p}_{CH}$ refers to Chib's method, and the estimator proposed in this paper is referenced as $\hat{p}_A$. Relevant standard errors in parentheses, results from three independent MCMC runs per data set are reported

multi-modal posterior distributions where a second run is required. Hence, the presented approach can be implemented generically for many important classes of models. This property substantially decreases the burden for the analyst associated with computing the marginal likelihood.

The second advantage of the presented approach is improved stability when the number of model coefficients increases, as illustrated in Section 3.3. Herein, the high level of stability, in contrast to the existing approaches, was shown up to a 142-dimensional random intercept probit. While this result can encourage analysts to report the marginal likelihood as part of Bayesian analyses more frequently, no general rule can be derived as to the number of dimensions for which the method will provide stable estimates. Therefore, assessments of the relationship between the number of dimensions, the type of the analyzed model, and the precision of the resulting estimates are suggested for future research.

Comparable uncertainty remains regarding the best choice of the tuning parameters for the presented approach, these are ρ and α . However, despite the very different model types on which the presented approach was tested in Section 3, all applications utilized the same values for these parameters. The high stability of the estimates in these applications suggests that the precision of the method may not be sensitive to the choice of tuning parameters. Nevertheless, future research should provide a better understanding how and to what extent an optimized choice of these parameters could further improve the precision of the method. As a final remark, and as mentioned in Section 2, we note that in principle the choice of the geometric object

for $\mathbf{A}_2$, which was chosen to be a K -dimensional ellipsoid in this article, is a tuning parameter. The availability of methods for drawing uniformly from within an ellipsoid, and for providing the log of its volume even when K is high, make it a convenient and efficient choice. Nevertheless, it cannot be ruled out that a different definition of $\mathbf{A}_2$ could further improve the properties of the presented approach.

A Subroutines for Algorithm I

A.1 Log of volume of K -dimensional ellipsoid

Algorithm 1 returns the log of the volume of a K -dimensional ellipsoid.

Algorithm 1. $\log(V_{\mathbf{A}_2})$ of a K -dimensional ellipsoid.

```

1: procedure GETVOLUME( $\alpha, \mathbf{D}$ )
   #Defines vectors required for calculation
2:    $\mathbf{v}^{(0, \dots, K)}$  ... empty vector with  $i = 0, \dots, K$  elements
3:    $\mathbf{e}^{(1, \dots, K)}$  ... vector holding the  $k = 1, \dots, K$  eigenvalues of  $\alpha \mathbf{D}$ 
   #Sets the starting values of the recurrence
4:    $\mathbf{v}^{(0)} = \log(1)$ ;  $\mathbf{v}^{(1)} = \log(2)$ 
   #Calculates the volume of  $i$ -dimensional spheres of unit radius
5:   for  $i = 2$  to  $K$  do
6:      $\mathbf{v}^{(i)} = \log(\frac{2\pi}{i-1}) + \mathbf{v}^{(i-2)}$ 
7:   end for
   #Scales the volume of  $K$ -dimensional unit sphere with respect to main axes
8:    $\log(V_{\mathbf{A}_2}) = \mathbf{v}^{(K)} + \sum_{k=1}^K \log(\sqrt{e^{(k)}})$ 
9: end procedure

```

A.2 Drawing uniformly from within K -dimensional ellipsoid

Algorithm 2 returns one point $\boldsymbol{\theta}^{(r)}$ uniformly drawn from within a K -dimensional ellipsoid.

Algorithm 2. Drawing uniformly from within a K -dimensional ellipsoid.

```

1: procedure GETELLDRAW( $\boldsymbol{\theta}^*, \alpha, \mathbf{D}$ )
   #Generate a random radius from the uniform distribution
2:    $\mathbf{rs} = \text{rUnif}(0, 1)$ 
   #Generate vector holding  $K$  draws from the standard normal distribution
3:    $\mathbf{pt}^{(1, \dots, K)} = \text{rNorm}(K; 0, 1)$ 
   #Get scaling for the point onto the surface of a unit hypersphere
4:    $\lambda_{\text{sur}} = \sum_{k=1}^K \mathbf{pt}^{(k)^2}$ 
5:    $\lambda_{\text{hyp}} = \mathbf{rs}^{1/K} / \sqrt{\lambda_{\text{sur}}}$ 
   #Defines vectors required for calculation
6:    $\mathbf{eVals}^{(1, \dots, K)}$  ... vector holding the  $k = 1, \dots, K$  eigenvalues of  $\alpha \mathbf{D}$ 
7:    $\mathbf{eVecs}^{(1, \dots, K)}$  ... matrix holding the  $k = 1, \dots, K$  eigenvectors of  $\alpha \mathbf{D}$  in columns
   #Scale point  $\mathbf{pt}^{(1, \dots, K)}$  to the ellipsoid using the eigenvalues and rotate with the eigenvectors and add centroid
8:    $\boldsymbol{\theta}^{(r)} = (\lambda_{\text{hyp}} \mathbf{pt}^{(1, \dots, K)}) \mathbf{eVals}^{(1, \dots, K)} \mathbf{eVecs}^{(1, \dots, K)^T} + \boldsymbol{\theta}^*$ 
9: end procedure

```

A.3 Binary search method for choosing α

Algorithm 3 returns a value α such that *about* $L' = 0.49 \times L$ posterior draws lie in $\mathbf{A}$.

Algorithm 3. Binary search method for setting α .

```

1: procedure GETALPHA( $\theta^{(1)}, \dots, \theta^{(L)}, L, \theta^*, \mathbf{D}$ )
   # Set number  $L'$  of posterior draws that shall lie in  $\mathbf{A}$ , and set starting values of the algorithm
2:    $L' = 0.49 \times L$ ;
3:    $\alpha_{\text{low}} = 0.01$ ;  $\alpha_{\text{high}} = 100$ ;  $\alpha_{\text{max}} = \alpha_{\text{high}}$ ;  $\alpha_{\text{min}} = \alpha_{\text{low}}$ ;  $\delta_\alpha = \alpha_{\text{high}} - \alpha_{\text{low}}$ ;  $c = 0$ 
   # Start loop and define stop condition
4:   while  $\delta_\alpha > 0.1$  do
5:      $\delta_\alpha = \alpha_{\text{high}} - \alpha_{\text{low}}$ 
     # Count the number  $L'_{\text{low}}$  of posterior draws  $\theta^{(1)}, \dots, \theta^{(L)}$  for which
6:      $\theta^{(l)} \in \mathbf{A}$  if  $(p^*(\theta^{(1)}|\mathbf{y}) > \rho \ \& \ (\theta^{(l)} - \theta^*)(\alpha_{\text{low}}\mathbf{D})^{-1}(\theta^{(l)} - \theta^*)' < 1)$ ,
     # and count the number  $L'_{\text{high}}$  of posterior draws  $\theta^{(1)}, \dots, \theta^{(L)}$  for which
7:      $\theta^{(l)} \in \mathbf{A}$  if  $(p^*(\theta^{(1)}|\mathbf{y}) > \rho \ \& \ (\theta^{(l)} - \theta^*)(\alpha_{\text{high}}\mathbf{D})^{-1}(\theta^{(l)} - \theta^*)' < 1)$ .
     # Set new values of  $\alpha_{\text{low}}$  and  $\alpha_{\text{high}}$  based on the number of posterior draws lying in  $\mathbf{A}$  in the current iteration
8:     if  $L'_{\text{high}} \geq L'$  then
9:       if  $c = 0$  then
10:         $\alpha_{\text{max}} = \max(\alpha_{\text{high}}, \alpha_{\text{max}})$ 
11:       else
12:        if  $\alpha_{\text{high}} < \alpha_{\text{max}}$  then
13:           $\alpha_{\text{max}} = \alpha_{\text{high}}$ 
14:        end if
15:       end if
16:        $\alpha_{\text{high}} = (\alpha_{\text{low}} + \alpha_{\text{high}})/2$ 
17:        $c = 1$ 
18:     else
19:       if  $c = 1$  then
20:         $\alpha_{\text{high}} = \min(\alpha_{\text{high}} + 10, \alpha_{\text{max}})$ 
21:       else
22:         $\alpha_{\text{high}} = \alpha_{\text{high}} + 100$ 
23:       end if
24:     end if
25:     if  $L'_{\text{low}} < L'$  then
26:       if  $\alpha_{\text{low}} > \alpha_{\text{min}}$  then
27:         $\alpha_{\text{min}} = \alpha_{\text{low}}$ 
28:       end if
29:        $\alpha_{\text{low}} = (\alpha_{\text{low}} + \alpha_{\text{high}})/2$ 
30:     else
31:        $\alpha_{\text{low}} = \max((1 + \alpha_{\text{low}})/2, \alpha_{\text{min}})$ 
32:     end if
33:   end while
34:    $\alpha = \alpha_{\text{high}}$ 
35: end procedure

```

B Details of the Gibbs sampler from Section 3.3

Gibbs sampling from the model specified in (3.2) requires drawing from the following distributions:

- (1) Draw z_{it} from $p(z_{it}|\boldsymbol{\beta}, y_t)$ for all $i = 1, \dots, N$ and all $t = 1, \dots, T$.
- (2) Draw coefficients vector $\boldsymbol{\beta}$ from $p(\boldsymbol{\beta}|\mathbf{Z}, y_1, \dots, y_T)$.
- (3) Draw random intercepts γ_t from $p(\gamma_t|z_t, \boldsymbol{\beta}, \sigma^2, \mu)$ for all $t = 1, \dots, T$.
- (4) Draw the variance of the random intercepts σ^2 from $p(\sigma^2|y_1, \dots, y_T)$.
- (5) Draw the mean of the random intercepts μ from $p(\mu|y_1, \dots, y_T, \sigma^2)$.

Prior distributions and corresponding parameters are defined in Section 3.3, and drawing from distributions (1) to (5) from above is standard in Bayesian econometric literature, and we refer the interested reader to [23] for details.

Acknowledgment: This research was enriched through discussions and comments from Sylvia Frühwirth-Schnatter and Jed Cohen.

Funding: The research was partly done in course of the openENTRANCE project that has received funding under the European Union's Horizon 2020 research and innovation programme, grant agreement No. 835896.

References

- [1] J. H. Albert and S. Chib, Bayesian analysis of binary and polychotomous response data, *J. Amer. Statist. Assoc.* **88** (1993), no. 422, 669–679.
- [2] D. Ardia, N. Baştürk, L. Hoogerheide and H. K. van Dijk, A comparative study of Monte Carlo methods for efficient evaluation of marginal likelihood, *Comput. Statist. Data Anal.* **56** (2012), no. 11, 3398–3414.
- [3] J.-M. Bernardo and A. F. M. Smith, *Bayesian Theory*, Probab. Math. Stat., John Wiley & Sons, Chichester, 1994.
- [4] G. Celeux, M. Hurn and C. P. Robert, Computational and inferential difficulties with mixture posterior distributions, *J. Amer. Statist. Assoc.* **95** (2000), no. 451, 957–970.
- [5] S. Chib, Marginal likelihood from the Gibbs output, *J. Amer. Statist. Assoc.* **90** (1995), no. 432, 1313–1321.
- [6] S. Chib and I. Jeliazkov, Marginal likelihood from the Metropolis-Hastings output, *J. Amer. Statist. Assoc.* **96** (2001), no. 453, 270–281.
- [7] P. Congdon, *Applied Bayesian Modelling*, Probab. Math. Stat., John Wiley & Sons, Chichester, 2003.
- [8] S. Frühwirth-Schnatter, Bayesian model discrimination and Bayes factors for linear Gaussian state space models, *J. Roy. Statist. Soc. Ser. B* **57** (1995), no. 1, 237–246.
- [9] S. Frühwirth-Schnatter, Fully Bayesian analysis of switching Gaussian state space models, *Ann. Inst. Statist. Math.* **53** (2001), 31–49.
- [10] S. Frühwirth-Schnatter, Markov chain Monte Carlo estimation of classical and dynamic switching and mixture models, *J. Amer. Statist. Assoc.* **96** (2001), no. 453, 194–209.
- [11] S. Frühwirth-Schnatter, Estimating marginal likelihoods for mixture and Markov switching models using bridge sampling techniques, *Econom. J.* **7** (2004), no. 1, 143–167.
- [12] S. Frühwirth-Schnatter, *Finite Mixture and Markov Switching Models*, Springer Ser. Statist., Springer, New York, 2006.
- [13] S. Frühwirth-Schnatter and R. Frühwirth, Auxiliary mixture sampling with applications to logistic models, *Comput. Statist. Data Anal.* **51** (2007), no. 7, 3509–3528.
- [14] S. Frühwirth-Schnatter and H. Wagner, Marginal likelihoods for non-Gaussian models using auxiliary mixture sampling, *Comput. Statist. Data Anal.* **52** (2008), no. 10, 4608–4624.
- [15] A. Fussl, S. Frühwirth-Schnatter and R. Frühwirth, Efficient MCMC for binomial logit models, *ACM Trans. Model. Comput. Simul.* **23** (2013), no. 1, Article No. 3.
- [16] J. Geweke, Interpretation and inference in mixture models: simple MCMC works, *Comput. Statist. Data Anal.* **51** (2007), no. 7, 3529–3550.
- [17] P. J. Green, Reversible jump Markov chain Monte Carlo computation and Bayesian model determination, *Biometrika* **82** (1995), no. 4, 711–732.
- [18] T. Hastie, R. Tibshirani and J. Friedman, *The Elements of Statistical Learning. Data Mining, Inference, and Prediction*, 2nd ed., Springer Ser. Statist., Springer, New York, 2009.
- [19] R. E. Kass and A. E. Raftery, Bayes factors, *J. Amer. Statist. Assoc.* **90** (1995), no. 430, 773–795.

- [20] P. Lenk, Simulation pseudo-bias correction to the harmonic mean estimator of integrated likelihoods, *J. Comput. Graph. Statist.* **18** (2009), no. 4, 941–960.
- [21] X.-L. Meng and W. H. Wong, Simulating ratios of normalizing constants via a simple identity: A theoretical exploration, *Statist. Sinica* **6** (1996), no. 4, 831–860.
- [22] M. A. Newton and A. E. Raftery, Approximate Bayesian inference with the weighted likelihood bootstrap. With discussion and a reply by the authors, *J. Roy. Statist. Soc. Ser. B* **56** (1994), no. 1, 3–48.
- [23] P. E. Rossi, G. M. Allenby and R. McCulloch, *Bayesian Statistics and Marketing*, Probab. Math. Stat., John Wiley & Sons, Chichester, 2005.
- [24] A. Skrondal and S. Rabe-Hesketh, Prediction in multilevel generalized linear models, *J. Roy. Statist. Soc. Ser. A* **172** (2009), no. 3, 659–687.
- [25] T. A. B. Snijders and R. J. Bosker, *Multilevel Analysis. An Introduction to Basic and Advanced Multilevel Modeling*, 2nd ed., Sage Publications, Los Angeles, 2012.
- [26] M. Stephens, Dealing with label switching in mixture models, *J. R. Stat. Soc. Ser. B Stat. Methodol.* **62** (2000), no. 4, 795–809.
- [27] A. Zellner, *An Introduction to Bayesian Inference in Econometrics*, Probab. Math. Stat., John Wiley & Sons, New York, 1971.