Statistica Sinica **13**(2003), 423-442

A BAYESIAN APPROACH TO THE SELECTION AND TESTING OF MIXTURE MODELS

Johannes Berkhof, Iven van Mechelen and Andrew Gelman

Catholic University Leuven, Free University Medical Center and Columbia University

Abstract: An important aspect of mixture modeling is the selection of the number of mixture components. In this paper, we discuss the Bayes factor as a selection tool. The discussion will focus on two aspects: computation of the Bayes factor and prior sensitivity. For the computation, we propose a variant of Chib's estimator that accounts for the non-identifiability of the mixture components. To reduce the prior sensitivity of the Bayes factor, we propose to extend the model with a hyperprior. We further discuss the use of posterior predictive checks for examining the fit of the model. The ideas are illustrated by means of a psychiatric diagnosis example.

Key words and phrases: Bayes factor, non-identifiability, hyperprior, latent class model, posterior predictive check, prior sensitivity, psychiatric diagnosis.

1. Introduction

The specification of a mixture model involves the selection of the number of components. This selection procedure can be performed in one of several ways. A possible strategy is to perform goodness-of-fit tests based on a likelihood ratio or Pearson chi-square statistic, and to extend the model until a reasonable fit is obtained. Another strategy is to compare alternative models by means of summary numbers including various information criteria. A summary number which may be singled out for its clarity is the Bayes factor (e.g., Berger and Sellke (1987); Kass and Raftery (1995)). A reason for computing the Bayes factor rather than performing a goodness-of-fit test is that the Bayes factor is based on weighing the alternative models by the posterior evidence in favor of each of them. Such evidence is not measured by the p-value of a goodness-of-fit test. A small p-value represents some evidence against a null hypothesis (Casella and Berger (1987); Berger and Sellke, (1987)), but a large p-value does not represent evidence in favor of the null. A second reason for computing the Bayes factor is that it can be used when comparing nonnested models. This makes the Bayes factor especially suitable for use in constrained mixture models where alternative models are nonnested (Clogg and Goodman (1984)).

Despite these advantages, model selection procedures based on a summary number like the Bayes factor may be criticized for not addressing the issue of model fit. Hence, we run the risk of selecting from a set of badly fitting alternatives. Goodness-of-fit tests, on the other hand, tell us whether a selected model is consistent with the observed data and, if none of the alternatives fits, stimulates us to search for new, better fitting models. Bayes factors and goodness-of-fit tests may therefore be applied simultaneously albeit for different purposes, model selection and examination of model fit.

In the present paper we discuss the Bayes factor as a model selection criterion used in combination with goodness-of-fit tests. The proposed model specification strategy is to formulate new models if the existing models have a bad fit, and to compare the new models with the initially selected model by means of the Bayes factor. The simultaneous use of goodness-of-fit tests and the Bayes factor helps us to arrive at a model which not only has a high posterior probability of being correct but also has a reasonable fit.

In the discussion of the Bayes factor, we address its computation and the choice of the prior. For the computation, we present a variant of Chib's estimator (1995). Chib's method is based on the Gibbs output and is especially suitable for mixture models. Yet, as noted by Neal (1999), we must modify Chib's estimator slightly. The posterior of a mixture model with Q components has $Q!$ modes due to permutability of the component labels, of which usually only a few are covered by the simulated posterior output. The reason for this is that the Gibbs sampler method mixes well within one of the modes but does not always mix well between the modes. Neal (1999) suggested inducing mixing between the modes by extending the Gibbs cycle with a relabeling transition for the mixture components. However, he also noted that this yields an estimator that is expensive to compute when the number of components is large. In this paper, we present a variant of Chib's estimator that accounts for the non-identifiability of the modes and is computationally less expensive than the modification proposed by Neal (1999).

The remainder of the paper is divided into five sections. After formalizing the mixture model (Section 2), the computation of the Bayes factor is discussed and the estimation methods are compared in a simulation study (Section 3). The choice of the prior is discussed in Section 4. In order to have a prior that is not contradicted by the given data set, we present a hierarchical Bayes procedure for estimating the hyperparameters of the prior distribution. Some goodness-of-fit tests are presented in Section 5. The replicated data to construct a reference distribution for the test quantities under the null are obtained from the posterior predictive distribution (Rubin (1984), Gelman, Meng and Stern (1996), Meng

(1994)). Section 6 contains some concluding remarks. The points made in the paper are illustrated by fitting a mixture model to psychiatric judgement data (Van Mechelen and de Boeck (1989)).

2. Mixture Model

For sake of generality, we use a multivariate setting in which the scores of N units on J variables are arranged in an $N \times J$ matrix $\boldsymbol{X}$. The i-th row of $\boldsymbol{X}$ is denoted by $X_i = (x_{i1}, \ldots, x_{iJ})$. A mixture model for this type of data assumes a latent partition of the units into Q classes, each class being characterized by some component density. The class membership of unit i is represented by the unobservable variable $Z_i = (z_{i1}, \ldots, z_{iQ})$ the q-th element of which is 1 if unit i is member of class q, and is 0 otherwise. Because the membership of unit i is unknown, the likelihood of X_i is a mixture of the component densities. In particular, if we denote the vector of mixing probabilities by $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_Q)^t$ and the other model parameters by $\boldsymbol{\pi}$, the likelihood of X_i is $p(X_i|\boldsymbol{\pi}, \boldsymbol{\lambda}) = \sum_{q=1}^{Q} \lambda_q p(X_i|\boldsymbol{\pi}, z_{iq} = 1)$. The posterior distribution of the model parameters $\boldsymbol{\theta} = (\boldsymbol{\lambda}', \boldsymbol{\pi}')'$ can be computed using chained data augmentation (Diebolt and Robert (1994); Rubin and Stern (1994)), alternately drawing from $p(\boldsymbol{\theta}|\boldsymbol{Z}, \boldsymbol{X})$ and $p(\boldsymbol{Z}|\boldsymbol{\theta}, \boldsymbol{X})$.

3. Model Selection

3.1. Definition of the Bayes factor

Suppose we have models $\mathcal{M}_1$ and $\mathcal{M}_2$. The Bayes factor is formally defined as the ratio of the posterior odds to the prior odds: $BF_{12} = \frac{[p(\mathcal{M}_1|\boldsymbol{X})/p(\mathcal{M}_2|\boldsymbol{X})]}{[p(\mathcal{M}_1)/p(\mathcal{M}_2)]}$. We see that if the prior model probabilities are equal, then BF_{12} is larger than 1 if $\mathcal{M}_1$ has a higher posterior probability. For computational purposes, it is more convenient to write the Bayes factor as the ratio of marginal likelihoods: $BF_{12} = p(\boldsymbol{X}|\mathcal{M}_1)/p(\boldsymbol{X}|\mathcal{M}_2)$. In the following, the computation of the marginal likelihood will be discussed in detail.

3.2. Computation of the marginal likelihood

The marginal likelihood of model $\mathcal{M}$ can be expressed as

$$p(\boldsymbol{X}|\mathcal{M}) = \int p(\boldsymbol{X}|\boldsymbol{\theta}, \mathcal{M}) p(\boldsymbol{\theta}|\mathcal{M}) d\boldsymbol{\theta}. \tag{1}$$

We have to approximate the integral in (1). Common approximation methods like importance sampling are effective when the posterior is unimodal, as noted

by DiCiccio, Kass, Raftery and Wasserman (1997). However, the posterior distribution generally has (at least) $Q!$ modes because the value of the posterior density function is invariant to a permutation of the class labels.

A simulation-based method that works better for multimodal posterior densities has been proposed by Chib (1995). Chib's estimator is based on the identity

$$p(\boldsymbol{X}) = \frac{p(\boldsymbol{X}|\boldsymbol{\theta}^*)p(\boldsymbol{\theta}^*)}{p(\boldsymbol{\theta}^*|\boldsymbol{X})}, \tag{2}$$

which holds for any $\boldsymbol{\theta}^*$. Conditioning on $\mathcal{M}$ is omitted in (2) to retain short expressions. The prior probability $p(\boldsymbol{\theta}^*)$ and the likelihood value $p(\boldsymbol{X}|\boldsymbol{\theta}^*)$ can be directly computed. The posterior probability $p(\boldsymbol{\theta}^*|\boldsymbol{X})$ can be estimated from the Gibbs output by

$$\hat{p}_{\mathrm{I}}(\boldsymbol{\theta}^*|\boldsymbol{X}) = \frac{1}{T}\sum_{t=1}^{T} p(\boldsymbol{\theta}^*|\boldsymbol{X}, \boldsymbol{Z}_{(t)}), \tag{3}$$

where $\boldsymbol{Z}_{(t)}$ is the t-th draw from $p(\boldsymbol{Z}|\boldsymbol{X})$ (Gelfand and Smith (1990)) and $\boldsymbol{\theta}^*$ is a chosen parameter vector that lies in one of the $Q!$ modal regions. For $\boldsymbol{\theta}^*$, we may choose one of the $Q!$ posterior modes. A rough approximation of the mode will usually suffice. Substituting (3) in (2) yields $\hat{p}_{\mathrm{I}}(\boldsymbol{X}) = p(\boldsymbol{X}|\boldsymbol{\theta}^*)p(\boldsymbol{\theta}^*)/\hat{p}_{\mathrm{I}}(\boldsymbol{\theta}^*|\boldsymbol{X})$. In order for the estimator $\hat{p}_{\mathrm{I}}(\boldsymbol{X})$ to be correct, the Markov chain $\{(\boldsymbol{\theta}_{(t)}, \boldsymbol{Z}_{(t)}); t = 1, \ldots, T\}$ has to explore all $Q!$ modal regions that exist because of the nonidentifiability of the mixture component labels. Neal (1999) noted that the probability of switching from one modal region to one of the other $Q! - 1$ modal regions may be very small in which case it is likely that some of the modes remain unexplored. An obvious way to handle this mixing problem is to include constraints on the model parameters. However, as noted by Celeux, Hurn and Robert (2000), these constraints have an influence on the performance of the sampler and may jeopardize the posterior inference. Neal (1999) suggested extending the Gibbs sampling scheme with relabeling transitions. He added, however, that this modification works satisfactorily only if the number of mixture components is small. Otherwise, the Markov chain will stay most of the time in the neighborhood of one of the relabelings of $\boldsymbol{\theta}^*$ rather than near $\boldsymbol{\theta}^*$ itself (if the number of components is 5, we already have $5! = 120$ modes) and will therefore be an inefficient estimator of $p(\boldsymbol{\theta}^*|\boldsymbol{X})$.

To illustrate the implementation of Neal's relabeling transitions, suppose that we have a model with two components ($Q = 2$). We start with a Gibbs chain $\{\boldsymbol{Z}_{(t)}; t = 1, \ldots, T\}$ in which labeling transitions have not been included. After having sampled this chain, we switch $z_{i1(t)}$ and $z_{i2(t)}$ for $i = 1, \ldots, N$ and $t = 1, \ldots, T$, yielding chain $\{\boldsymbol{Z}'_{(t)}; t = 1, \ldots, T\}$. Neal's modification of

Chib's estimator can then be defined as the average of $\{p(\boldsymbol{\theta}^*|\boldsymbol{X}, \boldsymbol{Z}_{(t)});\ t \text{ odd}\}$ and $\{p(\boldsymbol{\theta}^*|\boldsymbol{X}, \boldsymbol{Z}'_{(t)});\ t \text{ even}\}$. This estimator will be denoted as $\hat{p}_{\mathrm{II}}(\boldsymbol{\theta}^*|\boldsymbol{X})$. We can also use different transition schemes as long as fifty percent of the values are based on draws from $\{\boldsymbol{Z}_{(t)};\ t = 1, \ldots, T\}$ and fifty percent are based on draws from $\{\boldsymbol{Z}'_{(t)};\ t = 1, \ldots, T\}$.

Instead of using subchains, we also obtain a correct estimator of the posterior probability $p(\boldsymbol{\theta}^*|\boldsymbol{X})$ if we use all draws from $\{\boldsymbol{Z}_{(t)}; t = 1, \ldots, T\}$ and $\{\boldsymbol{Z}'_{(t)}; t = 1, \ldots, T\}$. We formulate a generalization of the latter estimator for a model with Q components ($Q \geq 2$). If the number of components is Q, we can perform $Q!-1$ different reorderings of the original chain. Each of these reorderings is obtained by performing a different permutation of the mixture component indices of $z_{iq(t)}$; the sth reordering ($s = 2, \ldots, Q!$) will be denoted as $\{\boldsymbol{Z}_{s(t)};\ t = 1, \ldots, T\}$. The original Gibbs chain corresponds to $s = 1$. The estimator of $p(\boldsymbol{\theta}^*|\boldsymbol{X})$ now becomes

$$\hat{p}_{\mathrm{III}}(\boldsymbol{\theta}^*|\boldsymbol{X}) = \frac{1}{Q!T} \sum_{s=1}^{Q!} \sum_{t=1}^{T} p(\boldsymbol{\theta}^*|\boldsymbol{X}, \boldsymbol{Z}_{s(t)}).$$

The above estimator is computationally intensive for large Q. In the following, we present a simulation-consistent estimator based on a stratification principle (Cochran (1977, p.87)). This estimator is based on the original Gibbs output $\{\boldsymbol{Z}_{1(t)};\ t = 1, \ldots, T\}$ and a smaller number of systematic draws from output where the mixture components indices have been permuted. The reason for distinguishing the non-permuted and permuted output is that the values of $p(\boldsymbol{\theta}^*|\boldsymbol{Z}_{s(t)})$ for the non-permuted output tend to be more variable than the values based on the permuted output (which are generally small as $\boldsymbol{\theta}^*$ is computed from the non-permuted output). We estimate $p(\boldsymbol{\theta}^*|\boldsymbol{X})$ by

$$\hat{p}_{\mathrm{V}}(\boldsymbol{\theta}^*|\boldsymbol{X}) = \frac{1}{Q!}\hat{p}_{\mathrm{I}}(\boldsymbol{\theta}^*|\boldsymbol{X}) + \frac{Q!-1}{Q!}\hat{p}_{\mathrm{IV}}(\boldsymbol{\theta}^*|\boldsymbol{X}),$$

where $\hat{p}_{\mathrm{I}}(\boldsymbol{\theta}^*|\boldsymbol{X})$ is based on the non-permuted output and $\hat{p}_{\mathrm{IV}}(\boldsymbol{\theta}^*|\boldsymbol{X})$ is based on the permuted output. Estimator $\hat{p}_{\mathrm{I}}(\boldsymbol{\theta}^*|\boldsymbol{X})$ is defined in (3) and

$$\hat{p}_{\mathrm{IV}}(\boldsymbol{\theta}^*|\boldsymbol{X}) = \frac{1}{(Q!-1)T_2} \sum_{s=2}^{Q!} \sum_{t=1}^{T_2} p(\boldsymbol{\theta}^*|\boldsymbol{X}, \boldsymbol{Z}_{s(tT/T_2)}).$$

Substituting $\hat{p}_{\mathrm{V}}(\boldsymbol{\theta}^*|\boldsymbol{X})$ into (2) yields an estimator for the marginal likelihood which will be denoted by $\hat{p}_{\mathrm{V}}(\boldsymbol{X})$.

3.3. Simulation example: mixture of two normals

To illustrate the need for modifying Chib's original estimator in case of bad mixing, we conducted a simulation study in which the original estimator

$\hat{p}_{\mathrm{I}}(\boldsymbol{\theta}^*|\boldsymbol{X})$ is compared to Neal's estimator $\hat{p}_{\mathrm{II}}(\boldsymbol{\theta}^*|\boldsymbol{X})$ and to the stratified sampling estimators. We postulated a mixture model with two normal components: $x_i \sim \lambda N(\mu_1, 1) + (1-\lambda)N(\mu_2, 1)$, $i = 1, \ldots, N$. For λ, we took a standard uniform prior and for μ_1 and μ_2, we took normal priors with mean zero and variance 100.

We simulated data sets of size 20 where the first 6 draws of each data set were taken from $N(-\delta, 1)$ and the next 14 draws were from $N(\delta, 1)$. The design factor δ was set equal to 0, 1 and 2, and for each value of λ, 100 data sets were simulated. For each data set, we computed $\hat{p}_{\mathrm{I}}(x)$, $\hat{p}_{\mathrm{II}}(x)$, $\hat{p}_{\mathrm{III}}(x)$, and $\hat{p}_{\mathrm{V}}(x)$ with $T_2/T = 0.5$, and $\hat{p}_{\mathrm{V}}(x)$ with $T_2/T = 0.1$. Regarding posterior simulation, we took a burn-in period of 10,000 draws and stored the subsequent 10,000 draws. The parameter $\boldsymbol{\theta}^*$ was calculated as $\operatorname{argmax}_{(t)}\{p(\boldsymbol{X}|\boldsymbol{\theta}_{(t)})p(\boldsymbol{\theta}_{(t)})\}$.

To examine the effect of the chain length on the performance of the estimators, we set T equal to $500, 1000, \ldots, 10,000$. The performance of the estimators was measured by the the mean squared deviation MSD $= \sum_{r=1}^{100}(\log \hat{p}_{(r)}(x) - \log p(x))^2/100$, where $\hat{p}_{(r)}$ is an estimate of the marginal likelihood $p(x)$ for the r-th data set. Since $p(x)$ is unknown, we estimated it by $\hat{p}_{\mathrm{III}}(x)$ using 100,000 new draws from the posterior. This number is sufficiently large to yield a very accurate approximation of the true marginal likelihood.

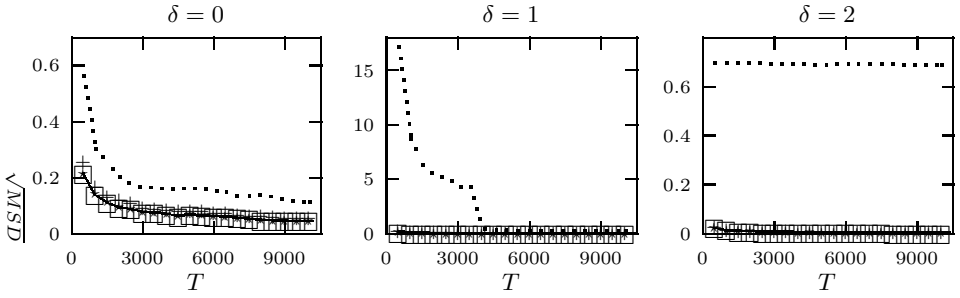

Figure 1. Square root of the mean squared deviation from the log marginal likelihood as a function of T. The estimators are $\log \hat{p}_{\mathrm{I}}(x)(\cdots)$, $\log \hat{p}_{\mathrm{II}}(x)(\square)$, $\log \hat{p}_{\mathrm{III}}(x)$ (—) , $\log \hat{p}_{\mathrm{V}}(x)$ with $T_2/T = 0.5(\star)$, and $\log \hat{p}_{\mathrm{V}}(x)$ with $T_2/T = 0.1(+)$.

If δ equals 0, all observations are drawn from the same distribution and we have a perfect mixing situation. We see from Figure 1 that $\hat{p}_{\mathrm{I}}(x)$ is the least efficient estimator. The estimators $\hat{p}_{\mathrm{II}}(x)$, $\hat{p}_{\mathrm{III}}(x)$, and $\hat{p}_{\mathrm{V}}(x)$ with $T_2/T = 0.5$ perform similarly and slightly better than $\hat{p}_{\mathrm{V}}(x)$ with $T_2/T = 0.1$. A possible reason for the relatively weak performance of $\hat{p}_{\mathrm{I}}(x)$ is that, although we have a perfect mixing situation, the permutable modes do not have to be covered

equally well by the posterior output. If δ equals 1, the two modes do not mix perfectly and the original estimator $\hat{p}_{\mathrm{I}}(x)$ is much less efficient than the other estimators when T is small. If δ equals 2, only one of the two permutable modes is explored and $\hat{p}_{\mathrm{I}}(x)$ and the other estimators do not tend to the same value if T tends to 10,000. To summarize, modifying $\hat{p}_{\mathrm{I}}(x)$ may improve the efficiency of the estimator also when the modes mix well.

3.4. The choice of T_2

To gain insight into the efficiency of $\hat{p}_{\mathrm{V}}(\boldsymbol{\theta}^*|\boldsymbol{X})$ as a function of T_2, let us consider the situation in which the modal regions mix well. Then the values of probabilities $p(\boldsymbol{\theta}^*|\boldsymbol{X}, \boldsymbol{Z}_{s(t)})$ tend to be of similar magnitude in the permuted and non-permuted output (and consequently, an estimator based on either the permuted or non-permuted output works fine). If we further assume that the probabilities $\{p(\boldsymbol{\theta}^*|\boldsymbol{X}, \boldsymbol{Z}_{s(t)}); s = 1, \ldots, Q!; t = 1, \ldots, T\}$ are independent, then the simulation standard errors of $\hat{p}_{\mathrm{I}}(\boldsymbol{\theta}^*|\boldsymbol{X})$ and $\hat{p}_{\mathrm{IV}}(\boldsymbol{\theta}^*|\boldsymbol{X})$ tend to be similar if an equal number of draws are taken from the permuted and non-permuted output (that is, $T = (Q! - 1)T_2$). If the number of draws are not equal ($T \neq (Q! - 1)T_2$), we have, approximately,

$$\frac{\text{s.e.}^2[\hat{p}_{\mathrm{I}}(\boldsymbol{\theta}^*|\boldsymbol{X})]}{\text{s.e.}^2[\hat{p}_{\mathrm{IV}}(\boldsymbol{\theta}^*|\boldsymbol{X})]} = \frac{(Q! - 1)T_2}{T}. \tag{4}$$

Under the independence assumption, the simulation standard error of $\hat{p}_{\mathrm{V}}(\boldsymbol{\theta}^*|\boldsymbol{X})$ is approximately

$$\text{s.e.}^2[\hat{p}_{\mathrm{V}}(\boldsymbol{\theta}^*|\boldsymbol{X})] \approx \frac{1}{Q!^2}\text{s.e.}^2[\hat{p}_{\mathrm{I}}(\boldsymbol{\theta}^*|\boldsymbol{X})] + \frac{(Q! - 1)^2}{Q!^2}\text{s.e.}^2[\hat{p}_{\mathrm{IV}}(\boldsymbol{\theta}^*|\boldsymbol{X})]. \tag{5}$$

If we substitute (4) into (5), we obtain

$$\text{s.e.}^2[\hat{p}_{\mathrm{V}}(\boldsymbol{\theta}^*|\boldsymbol{X})] \approx \frac{1 + (Q! - 1)T/T_2}{Q!^2}\text{s.e.}^2[\hat{p}_{\mathrm{I}}(\boldsymbol{\theta}^*|\boldsymbol{X})]. \tag{6}$$

From (6), it follows that if $T_2 \approx T/(Q! + 1)$, then the estimator $\hat{p}_{\mathrm{V}}(\boldsymbol{\theta}^*|\boldsymbol{X})$ has the same simulation standard error as Chib's estimator $\hat{p}_{\mathrm{I}}(\boldsymbol{\theta}^*|\boldsymbol{X})$.

If the modes do not mix well, we have two strata since the values of probabilities $p(\boldsymbol{\theta}^*|\boldsymbol{X}, \boldsymbol{Z}_{s(t)})$ tend to be smaller in the permuted output than in the non-permuted output. Consequently, the values in the permuted output tend to be less variable than the values in the non-permuted output. Under the independence assumption, the most efficient stratification estimator then has a value of T_2/T smaller than $1/(Q! + 1)$ (Cochran (1977, p.98)). For practical modeling, the value of $1/(Q! + 1)$ can be used as an upper bound for the ratio T_2/T.

3.5. Example: latent class modeling

We consider data, collected by Van Mechelen and De Boeck (1989), that consist of 0-1 judgements made by an experienced psychiatrist about the presence of 23 psychiatric symptoms on 30 patients ($N = 30, J = 23$). A 0 was scored if the symptom was absent, and 1 if present (see Table 1). We assume that the dependencies in the symptom patterns are captured by Q mixture components, each patient a member of a single patient category. We postulate a latent class model (Lazarsfeld and Henry (1968); Goodman (1974)), meaning that the conditional likelihood of the q-th component or class, $p(X_i|\boldsymbol{\pi}, z_{iq} = 1)$ (see Section 2), is the product of 23 independent Bernoulli distributions, one for each symptom. We choose a $Dirichlet(1, \ldots, 1)$ prior distribution for the mixture probabilities $\boldsymbol{\lambda}$ and independent $Beta(\alpha, \alpha)$ prior distributions for the component dependent symptom probabilities in $\boldsymbol{\pi}$: $\{\pi_{j|q}; j = 1, \ldots, J; q = 1, \ldots, Q\}$. We set α equal to 0.5, 1 and 2.

Table 1. Dichotomous judgements (x present, . absent) about the occurrence of 23 symptoms in 30 patients. Symptoms and patients have been arranged in increasing order of symptom occurrence.

Symptom label	Patients
disorientation	.....x..
obsession/compulsion	...x.....
memory impairment	.....x.....x..
lack of emotion	.....x.....x
antisocial impulses or acts	....x.....xx.....
speech disorganization	.....x...x.x
overt anger	.....x.....x.....x....
grandiosity	.....x..x....x..x.....
drug abuse	x...x.....x.....x...
alcohol abuse	.....x.....xx.....x.x..
retardation	.....x..xx....x.x
belligerence/negativism	.....x.....xxx.....x...
somatic concerns	..x.....xx.....x.....x.xx
suspicion/ideas of persecution	.....xxx.....xx...xx
hallucinations/delusions	.....xxx.....xx...xx
agitation/excitement	.....x.....x.xx.....xxx..x.
suicide	.x....xxx..xx...xx...xx.....xx
anxiety	.xxx..xxxxxx...x.x..xxx.xxx..x
social isolation	x.....xx.xxxx..xx.xxxx.x.xxxx
inappropriate affect or behaviour	...xx.x..x....xxxx.xxxxxxxxxx
depression	xxx...xxxxxxxx..xx.xxxx..xx.xx
leisure time impairment	..xxxxxxxxxxxx.xxxxxxxxxxx.x
daily routine impairment	..xxxxxxxxxxxx.xxxxxxxxxxxx

We estimated models with one to five latent classes. Regarding posterior simulation, we simulated ten chains with independent starting values and a burn-in period of 10,000 draws per chain, and we stored the subsequent 100,000 observations. This number was sufficient to achieve convergence in the sense that $\sqrt{\hat{R}}$ was smaller than 1.1 for all estimands of interest (Gelman and Rubin (1992)). For $\boldsymbol{\theta}^*$, we chose $\text{argmax}_{(t)}\{p(\boldsymbol{X}|\boldsymbol{\theta}_{(t)})p(\boldsymbol{\theta}_{(t)})\}$ using only draws of the first chain. For the other nine chains, the mixture components indices were relabeled. For each chain separately, the relabeling was chosen to minimize the quadratic distance between the first draw $\boldsymbol{\theta}_{(1)}$ and $\boldsymbol{\theta}^*$ so that $\boldsymbol{\theta}_{(1)}$ and $\boldsymbol{\theta}^*$ come from the same modal region. These ten chains were then treated as non-permuted output of the Gibbs sampling procedure to distinguish them from the permuted output that is needed as well when computing $\hat{p}_{\text{V}}(\boldsymbol{\theta}^*|\boldsymbol{X})$. To study the robustness of the marginal likelihood estimator with respect to T_2, we set T_2 equal to 0, 10^3, 10^4 and 10^5.

Figure 2 presents the values of the logarithm of the estimated marginal likelihood $\hat{p}_{\text{V}}(\boldsymbol{X})$ and the standard error of $\log \hat{p}_{\text{V}}(\boldsymbol{X})$ based on the variability among the ten independent single-chain estimates of $\log p(\boldsymbol{\theta}^*|\boldsymbol{X})$. The patterns in Figure 2 are fairly constant across the panels where $T_2 = 10^3$, 10^4, or 10^5. By contrast, the patterns for the panel where $T_2 = 0$ are a bit elevated if the number of components is larger than two. To understand this result, we inspected the Gibbs output and noticed that the components do not mix at all for the two component model whereas they do mix (although not perfectly) for models with more components. The estimate $\hat{p}_{\text{V}}(\boldsymbol{X})$ with $T_2 = 0$ differs from the unmodified estimator $\hat{p}_{\text{I}}(\boldsymbol{X})$ by a factor $Q!$. Since $\hat{p}_{\text{I}}(\boldsymbol{X})$ overestimates the marginal likelihood by a factor $Q!$ if the modal regions do not mix at all (Neal (1999)), the use of $\hat{p}_{\text{V}}(\boldsymbol{X})$ is then appropriate even after setting T_2 equal to 0.

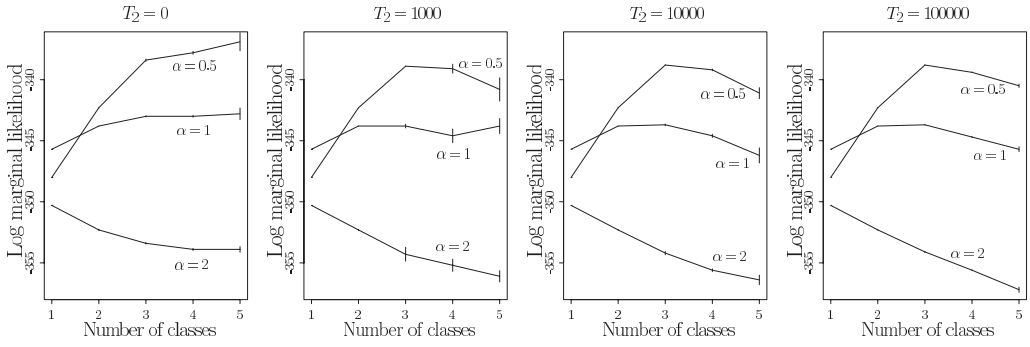

Figure 2. Log marginal likelihoods (with error bars showing standard errors from simulations) as a function of α and the number of classes, computed using $T = 10^6$ and four choices of T_2. The prior distributions are $\boldsymbol{\lambda} \sim Dirichlet(1, \ldots, 1)$; $\pi_{j|q} \sim Beta(\alpha, \alpha)$.

The standard errors are larger when $T_2 = 10^3$ than when $T_2 = 0$. To check whether this is reasonable, we compared T_2 to $T/(Q!+1)$. In Subsection 3.4, we showed that the value of $T/(Q!+1)$ is an approximation to the value of T_2 at which $\hat{p}_{\mathrm{V}}(\boldsymbol{X})$ is at least as efficient as $\hat{p}_{\mathrm{I}}(\boldsymbol{X})$. The standard error obtained when $T_2 = 0$ is equal to the one of $\log \hat{p}_{\mathrm{I}}(\boldsymbol{X})$, because then $\log \hat{p}_{\mathrm{V}}(\boldsymbol{X})$ and $\log \hat{p}_{\mathrm{I}}(\boldsymbol{X})$ differ only by a constant term. For the three to five-class model, the values of $T/(Q!+1)$ are 1.4×10^5, 4.0×10^4 and 8.3×10^3. Since these values are larger than 10^3, we expected the standard errors for $T_2 = 10^3$ to be larger than for $T_2 = 0$ as is indeed the case. In an analogous way, we compared the standard errors obtained when $T_2 = 10^4$ or 10^5 with the standard errors obtained when $T_2 = 0$. We found that the values of $T/(Q!+1)$ are not contradicted by Figure 2.

4. Choice of the Prior

4.1. Hierarchical prior

Figure 2 shows that the log marginal likelihood is rather sensitive to the choice of the prior distribution. If the models are compared at $\alpha = 2$, there is a preference for the one class model. At $\alpha = 1$, there is equal preference for the two- or three-class model, while at $\alpha = 0.5$ there is a preference for the three-class model.

The prior sensitivity of the Bayes factor is well known (Kass and Raftery (1995)) and demands a careful specification of the prior. A sensible approach is to use informative priors that reflect prior information on the model parameters. However, in practical modeling, prior information is not always available or is too vague to rely on. In that case, choosing a diffuse prior for $\boldsymbol{\theta}$ seems a quick way out but this does not work satisfactorily: If the prior is very diffuse, the value of $p(\boldsymbol{\theta}^*)$ is nearly zero, and by (2), the marginal likelihood is then nearly zero as well (Bartlett's or Lindley's paradox).

As outlined in the introduction, we propose a modeling strategy in which the selected model is checked for consistency with the observed data and, if none of the alternatives fits, we suggest a search for new, better-fitting models. Suppose now that it is not likely that the given data set is generated under the assumed prior distribution of the selected model. Then, a possible approach is to define a hierarchical extension of the mixture model in which the hyperparameters α and β are treated as unknown model parameters. This means that the α and β are essentially estimated from the data, which seems sensible unless one has a strong reason for favoring one of the specific choices of α and β. The underlying idea is that, by estimating the hyperparameters, we compare models for which the priors are at least not contradicted by the data. In the following, we illustrate the hierarchical Bayes approach for the psychiatric diagnosis example. We also

compare the prior and posterior for each model to check whether the hierarchical approach is worth the effort.

4.2. Example (continued): hierarchical modeling

In Section 3.5, we assumed either a symmetric $Beta(.5,.5)$, $Beta(1,1)$ or $Beta(2,2)$ prior for all component dependent symptom probabilities $\pi_{j|q}$. From now on, we relax this assumption and postulate the same, possibly asymmetric, $Beta(\alpha,\beta)$ prior for all symptom probabilities, with α and β being estimated rather than fixed. We follow Gelman, Carlin, Stern and Rubin (1995) and choose a diffuse hyperprior density for (α,β) that is uniform on $(\frac{\alpha}{\alpha+\beta}, \frac{1}{(\alpha+\beta)^{\frac{1}{2}}})$, in the range $\alpha/(\alpha+\beta)\epsilon(0,1)$ and $1/(\alpha+\beta)^{\frac{1}{2}}\epsilon(0,c)$, $c>0$. The expression $\alpha/(\alpha+\beta)$ is the mean of the $Beta(\alpha,\beta)$ distribution and $1/(\alpha+\beta)^{\frac{1}{2}}$ is a measure of dispersion (Gelman et al. (1995, p.131)).

Posterior simulation consists of subsequently drawing from $p(\boldsymbol{\theta}|\boldsymbol{X},\boldsymbol{Z},\alpha,\beta)$, $p(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta})$, and $p(\alpha,\beta|\boldsymbol{\theta})$. Because the latter distribution does not have a known form, we replace the last step of the Gibbs cycle by a Metropolis step. As a jumping distribution, we choose a uniform symmetric distribution around the current values of $\alpha/(\alpha+\beta)$ and $1/(\alpha+\beta)^{\frac{1}{2}}$.

We estimate the marginal likelihood from the Metropolis output. A general framework for estimating the marginal likelihood from the Metropolis-Hastings output is presented by Chib and Jeliazkov (2001). Here the estimator of the marginal likelihood is still based on identity (2) where, as before, the posterior probability $p(\boldsymbol{\theta}^*|\boldsymbol{X})$ is estimated by stratified sampling from the Gibbs chain and its reorderings. The estimators $\hat{p}_{\mathrm{I}}(\boldsymbol{\theta}^*|\boldsymbol{X})$ and $\hat{p}_{\mathrm{IV}}(\boldsymbol{\theta}^*|\boldsymbol{X})$ are defined as

$$\hat{p}_{\mathrm{I}}(\boldsymbol{\theta}^*|\boldsymbol{X}) = \frac{1}{T}\sum_{t=1}^{T} p(\boldsymbol{\theta}^*|\boldsymbol{X},\boldsymbol{Z}_{1(t)},\alpha_{(t)},\beta_{(t)}),$$

$$\hat{p}_{\mathrm{IV}}(\boldsymbol{\theta}^*|\boldsymbol{X}) = \frac{1}{(Q!-1)T_2}\sum_{s=2}^{Q!}\sum_{t=1}^{T_2} p(\boldsymbol{\theta}^*|\boldsymbol{X},\boldsymbol{Z}_{s(tT/T_2)},\alpha_{(tT/T_2)},\beta_{(tT/T_2)}).$$

The estimation of $p(\boldsymbol{X})$ involves approximating the prior probability $p(\boldsymbol{\theta}^*)$ which cannot be directly computed anymore. We write $p(\boldsymbol{\theta}^*)$ as

$$p(\boldsymbol{\theta}^*) = \int_0^\infty \int_{\max_{\{0,\frac{1}{c^2-\alpha}\}}}^\infty p(\boldsymbol{\theta}^*|\alpha,\beta)p(\alpha,\beta)d\beta d\alpha, \tag{7}$$

and approximate the double integral by means of Laplace bridge sampling (Meng and Wong (1996); diCiccio *et al.* (1997)) after having carried out the substitution $(u,v) = (\log(\alpha/\beta), \log(\alpha+\beta))$.

We simulated ten independent chains, each with a burn-in period of 10,000 draws. We set the upper bound c for the value of $1/(\alpha+\beta)^{\frac{1}{2}}$ equal to 10, yielding a sufficiently diffuse hyperprior. For the Metropolis step, we chose a symmetric uniform jumping distribution. The acceptance rate was larger than 0.2 for all models. We set T and T_2 equal to 10^6 and 10^5 and approximated the integral (7) using 50,000 draws from a normal approximation to the posterior density of (u, v).

The logarithms of the estimated marginal likelihood values were equal to -346.9, -340.8, -335.8, -335.7 and -335.8 for the models with one to five classes. The simulation standard error was never larger than 0.17. There is a clear preference for models with at least three classes. There is no preference for the three-, the four- or the five-class model, presumably because the number of patients is too small to be able to draw a distinction between them. The model selection results are different from the results in Section 3.5, in particular when α and β are fixed at 1 or at 2. For the latter values of the hyperparameters, the three-class model is not selected when compared to the two-class model. However, in the hierarchical model, the Bayes factor of the three-class model versus the two-class model is $\exp(-335.8)/\exp(-340.8) \approx 150$, which means that, under equal prior model probabilities, the posterior probability of the three-class model is 150 times higher than the posterior probability of the two-class model.

4.3. Example (continued): prior-posterior comparisons

In all non-hierarchical and hierarchical models under consideration, we have a single beta prior for all conditional symptom probabilities. To examine whether the priors are reasonable for the given data set, we constructed a histogram of a posterior draw of the conditional symptom probabilities for each of the models with at least two classes and plotted it together with the curve of a $Beta(\alpha, \beta)$ density (for a similar check of a psychometric model, see Meulders, Gelman, Van Mechelen and De Boeck (1998)). For the hierarchical model, we set the hyperparameters (α, β) equal to the posterior mode (α^*, β^*) (see Figure 3). We see that the curves and the histograms are similarly shaped for the hierarchical prior, which suggests that the $Beta(\alpha, \beta)$ prior is consistent with this aspect of the data. This is not the case for all models with a non-hierarchical prior. In particular, values near 0 are not well captured when assuming a symmetrical $Beta(0.5, 0.5)$, $Beta(1, 1)$ or $Beta(2, 2)$ prior. This supports our use of a hierarchical strategy.

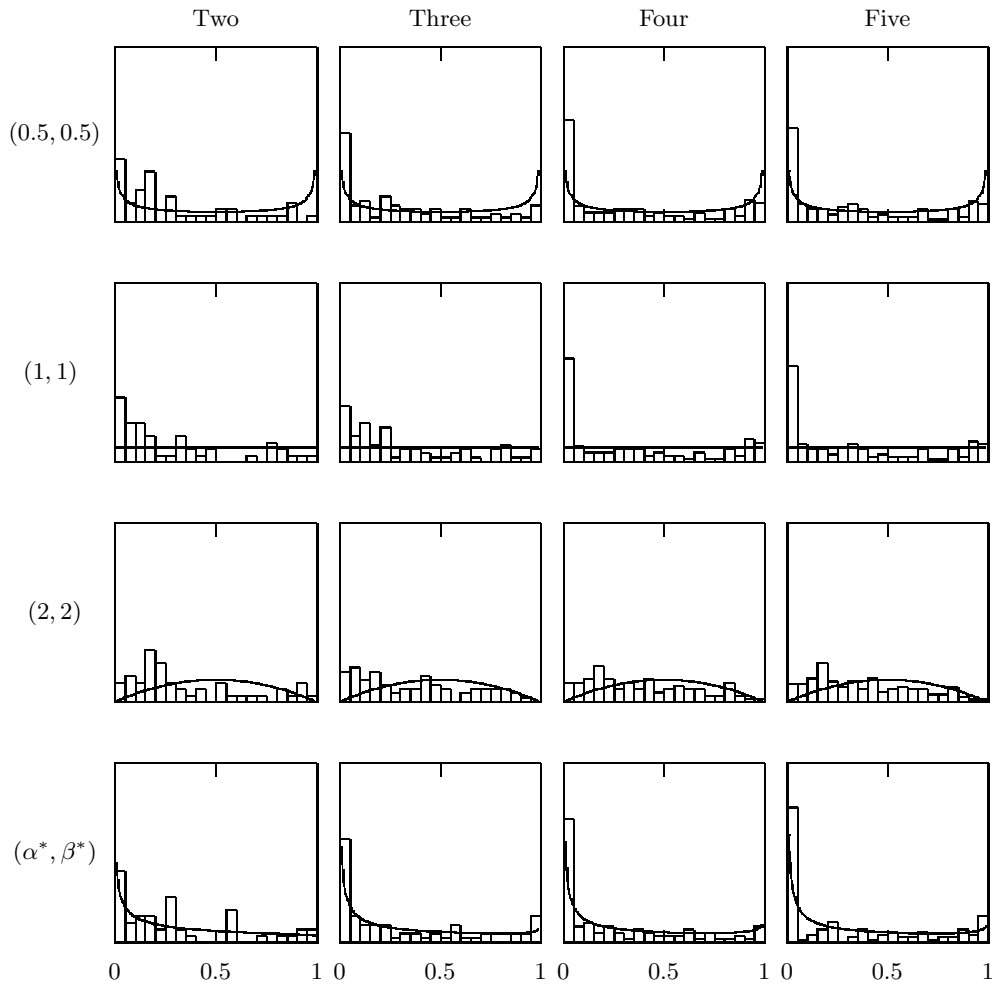

Figure 3. Histograms of a posterior draw of the conditional symptom probabilities. The first three rows correspond to models with $Beta(0.5, 0.5)$, $Beta(1, 1)$ and $Beta(2, 2)$ priors for the symptom probabilities. These prior densities are also drawn and rescaled in order to match with the histograms. The last row corresponds to models where a hyperprior is assumed for α and β; (α^*, β^*) is the posterior mode of (α, β).

5. Posterior Predictive Checking

5.1. Absolute and relative goodness-of-fit checks

As stated in the introduction, the Bayes factor does not solve the issue of model fit, that is, it does not reveal whether the selected model could have plausibly generated the observed data. Although, by adopting a hierarchical approach, we confined ourselves to priors that are not contradicted by the given data set,

the posterior may of course still be violated by the data. Since we would interpret the results from badly and well-fitting models differently, it makes sense to perform goodness-of-fit checks in addition to selecting a model by means of the Bayes factor. The goodness-of-fit model check is then used as a diagnostic tool which may help us improve the specification of the model. The replicates are drawn from the posterior predictive distribution, $p(\boldsymbol{X}^{\text{rep}}|\boldsymbol{X})$. From $p(\boldsymbol{X}^{\text{rep}}, \boldsymbol{\theta}, \boldsymbol{Z}|\boldsymbol{X}) \propto p(\boldsymbol{X}^{\text{rep}}|\boldsymbol{\theta}, \boldsymbol{Z})p(\boldsymbol{\theta}, \boldsymbol{Z}|\boldsymbol{X})$ (Rubin (1984); Gelman, Meng and Stern (1996); Meng (1994)), it follows that joint posterior draws $(\boldsymbol{X}^{\text{rep}}, \boldsymbol{\theta}, \boldsymbol{Z})$ are obtained by first sampling $(\boldsymbol{\theta}, \boldsymbol{Z})$ from $p(\boldsymbol{\theta}, \boldsymbol{Z}|\boldsymbol{X})$ and then $\boldsymbol{X}^{\text{rep}}$ from $p(\boldsymbol{X}^{\text{rep}}|\boldsymbol{\theta}, \boldsymbol{Z})$. The sampled vectors from the joint posterior distribution are denoted by $(\boldsymbol{X}_1^{\text{rep}}, \boldsymbol{\theta}_1, \boldsymbol{Z}_1), \ldots,$ $(\boldsymbol{X}_R^{\text{rep}}, \boldsymbol{\theta}_R, \boldsymbol{Z}_R)$.

Various goodness-of-fit quantities may be considered, including relative quantities such as a likelihood ratio test statistic in which the null model under consideration is compared to an alternative model (Rubin and Stern (1994)). In the following, we focus on absolute test quantities in which no alternative model is specified. Such test quantities are useful when checking for instance for outliers, residual dependencies, or distributional violations. The test quantities or discrepancy measures may also be functions of the model parameters and are denoted as $D(\boldsymbol{X}, \boldsymbol{\theta}, \boldsymbol{Z})$. A summary idea of the magnitude of the discrepancy can be obtained by comparing the posterior mean of $D(\boldsymbol{X}^{\text{rep}}, \boldsymbol{\theta}, \boldsymbol{Z})$ to the posterior mean of $D(\boldsymbol{X}, \boldsymbol{\theta}, \boldsymbol{Z})$. Outlyingness can also be summarized by the exceeding tail area probability, called posterior predictive p-value (Rubin (1984); Gelman, Meng and Stern (1996); Meng (1994)). It may be noted that posterior predictive checking using discrepancy measures tends to be conservative. Yet, modifications of the discrepancies to overcome this problem may be considered (Berkhof, Van Mechelen and Gelman (2002)).

5.2. Example (continued)

We illustrate posterior predictive checking in mixture models using the three-class model for the psychiatric diagnosis data as the null model. The posterior medians of the parameters of the model (computed after applying a $Q!$-means type of clustering analysis to the Gibbs output, see Celeux, Hurn, and Robert (2000)) are presented in Figure 4. Class 1 is associated with high probabilities on the symptoms agitation/excitement, suspicion/ideas of persecution, and hallucinations/delusions, being indicative of a psychosis syndrome. Class 2 is associated with depression, anxiety, and suicide and can be interpreted as an affective syndrome, while class 3 is associated primarily with alcohol abuse.

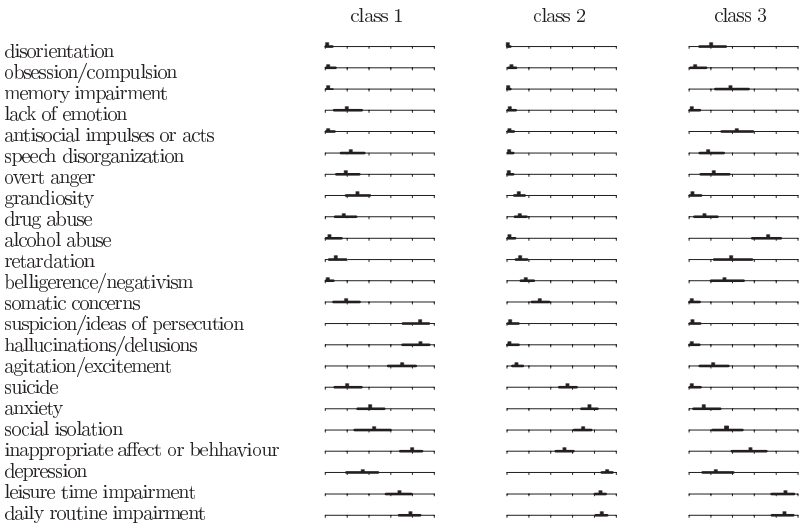

Figure 4. Posterior medians and 50% posterior intervals for the probability of each symptom being present, for each of the three classes. Each estimate and interval is overlaid on a $[0, 1]$ interval. Thus, for example, a patient in class 3 has an approximate 20% chance of having "disorientation," a nearly 0% chance of having "obsession/compulsion," and so forth.
The estimated marginal frequencies of each of the three latent classes in the population are (with 50% intervals): 23% (18%, 30%) for class 1, 58% (51%, 65%) for class 2, and 17% (12%, 23%) for class 3.

In a mixture model, each unit is assumed to be member of a single class. The posterior distribution of $(z_{i1}, \ldots, z_{iQ})$ expresses the uncertainty about the membership of unit i. The posterior mean of (z_{i1}, z_{i2}, z_{i3}) in the three-class model for the psychiatric diagnosis data contains a value larger than 0.9 for 21 out of 30 patients. This shows that most but not all patients are well classified in the three-class model.

In general, it is of interest to check whether the individual response patterns are well-fitted by the distributions of the corresponding mixture components. If this is not the case, the unit might belong to a component different from the ones included in the model. To examine this, we defined the patient-specific discrepancy measure $D_{1i}(\boldsymbol{X}, \boldsymbol{\theta}, \boldsymbol{Z}) = \sum_{q=1}^{Q} \sum_{j=1}^{J} |x_{ij} - \pi_{j|q}| I_{z_{iq}=1}$. In Figure 5a, we plot the posterior mean of $D_{1i}(\boldsymbol{X}^{\text{rep}}, \boldsymbol{\theta}, \boldsymbol{Z})$ against the posterior mean of $D_{1i}(\boldsymbol{X}, \boldsymbol{\theta}, \boldsymbol{Z})$. It appears from this figure that none of the patients is poorly fitted by the model, as the difference between the mean realized discrepancy value and the mean discrepancy value predicted under the model is never larger than 2.0 (the latter value pertaining to patient 30). The posterior predictive p-values can be read from plots $D_{1i}(\boldsymbol{X}^{\text{rep}}, \boldsymbol{\theta}, \boldsymbol{Z})$ against $D_{1i}(\boldsymbol{X}, \boldsymbol{\theta}, \boldsymbol{Z})$ (Gelman *et al.* (1995)).

In Figure 2, such a plot, based on 1000 draws from the posterior predictive distribution, is presented for patient 30. The posterior predictive p-value is the percentage of draws above the diagonal line and equals 0.06, which implies that the realized discrepancy of the data is somewhat higher than one might expect for replications under the model. For the other 29 subjects, the posterior predictive p-values are between 0.3 and 0.6. Although these model checking results do not indicate a serious model violation, it may still be interesting to inspect the data of patient 30 more closely. The posterior mean realized discrepancy turns out to be relatively large because this patient has symptoms that are typical for a psychosis syndrome but also symptoms that are typical for a depression. This gives some support for the existence of a separate class of patients which have both depressed and psychotic features.

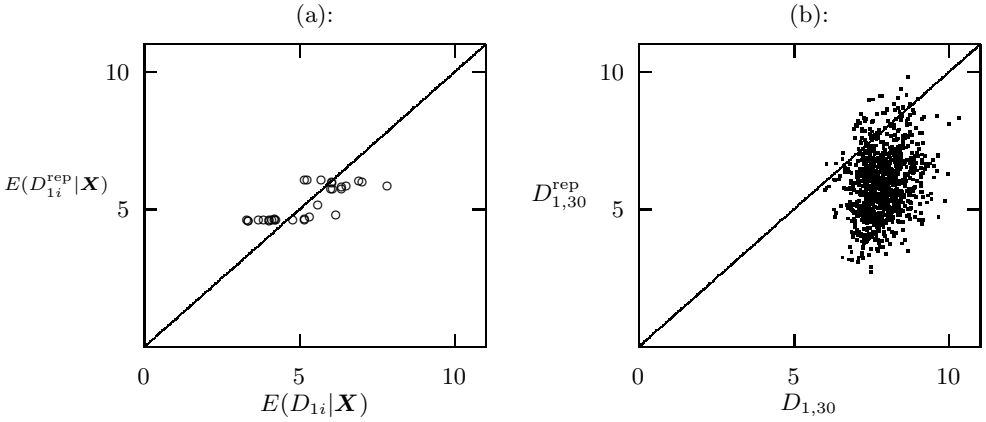

Figure 5. (a) Posterior predictive mean of $D_{1i}(\boldsymbol{X}^{\text{rep}}, \boldsymbol{\theta}, \boldsymbol{Z})$ against posterior mean of $D_{1i}(\boldsymbol{X}, \boldsymbol{\theta}, \boldsymbol{Z})$. (b) 1000 posterior simulations of $(D_{1,30}(\boldsymbol{X}^{\text{rep}}, \boldsymbol{\theta}, \boldsymbol{Z})$, $D_{1,30}(\boldsymbol{X}, \boldsymbol{\theta}, \boldsymbol{Z}))$. Panel a shows that the discrepancy for patient 30 (having the largest realized discrepancy value) is slightly higher than predicted under the model; however, the p-value of .06 (that is, 6% of the dots are above the diagonal line in panel b) indicates that the discrepancy could plausibly be explained by chance.

It is also interesting to check whether the postulated densities for the different mixture components are contradicted by the data. In the latent class model, we postulate a product of independent Bernoulli distributions for each mixture component. The independence assumption can be tested separately for each component by a posterior predictive check that uses only the information that is contained in pairs of item scores. Test quantities that are based on pairs of item scores have been proposed by Hoijtink (1998) and Reiser and Lin (1999). By means of a simulation study, Reiser and Lin (1999) showed that when the

number of subjects is small compared to the number of possible response patterns (i.e., when the data are sparse), a test quantity based on pairs of item scores has considerably more power than a test quantity based on the full response pattern. In order to formulate class-specific discrepancy measures, we define frequency n_{ab}^{jkq} which is the number of times that within class q, the values a and b are scored on symptoms j and k. Note that the value of n_{ab}^{jkq} can be computed only if the latent membership values are known. For class q, we use the discrepancy measure $D_{2q}(\boldsymbol{X}, \boldsymbol{\theta}, \boldsymbol{Z}) = \sum_{j=1}^{J-1} \sum_{k=j+1}^{J} p(n_{11}^{jkq}, n_{10}^{jkq}, n_{01}^{jkq}, n_{00}^{jkq} | \boldsymbol{\pi})$. The likelihood $p(n_{11}^{jkq}, n_{10}^{jkq}, n_{01}^{jkq}, n_{00}^{jkq} | \boldsymbol{\pi})$ is the density of the four response patterns for class q, n_{11}^{jkq}, n_{01}^{jkq}, n_{10}^{jkq}, and n_{00}^{jkq}, as implied by the overall latent class model. The posterior predictive p-values obtained for class 1, 2 and 3 are 0.43, 0.41 and 0.52, respectively. This shows that there is no evidence that the conditional independence assumption is violated for any of the three classes.

6. Concluding Remarks

A distinction that is often drawn in statistical modeling is one between model selection and assessment of model fit. As to the former, we discussed the common Bayesian selection criterion, the Bayes factor, whereas for the latter we relied on posterior predictive checks. We illustrated with our example analysis that Bayes factors and posterior predictive checks can be meaningfully combined and supplement each other with complementary information although they stem from quite different research traditions.

When applying the Bayes factor, we had to deal with two statistical problems. The first problem is multimodality of the posterior, which typically occurs in mixture models. We showed how the Bayes factor can be computed from the Gibbs output by means of a modification of Chib's (1995) method that accounts for multimodality. Such a modification is required only when some modal regions are not visited by the Gibbs sampler. We applied the modification of Chib's method to a latent class model but it can be applied to mixtures of other types of distributions as well. It is also of interest to consider the same method when exploring the posterior using a tempered annealing scheme (Neal (1996); Celeux *et al.* (2000)). The second problem we had to deal with is prior sensitivity. In general, it is well known that the Bayes factor may be prior sensitive, even to changes in the prior distribution that essentially have no influence on the posterior distribution. To account for prior sensitivity, several variants of the Bayes factor have been proposed in the literature (for an overview, see Gelfand and Dey (1994)). However, these variants cannot be calibrated in terms of evidence in favor of either model. An alternative is to compute the Bayes factor for a set of different prior distributions (Kass and Raftery (1995)). This approach

forces us to define a class of "reasonable" models. For the latent class models considered when analyzing the psychiatric judgement data, reasonable prior distributions for the symptom probabilities could be symmetric beta densities with the hyperparameters set at values between 0 and 2. Yet the choice of the hyperparameters is rather arbitrary and a change in the hyperparameters may affect the Bayes factor considerably, as shown in the example. A hierarchical extension of the latent class model was shown to provide a neat way out; such an extension implies the choice of a prior that is not contradicted by the data, which may further yield acceptable prior-posterior comparisons. The application of such a hierarchical approach is not limited to the latent class case and may be considered for any type of mixture model. In general, it seems sensible to use priors that are not contradicted by the data.

Regarding posterior predictive checks, an unlimited number of test quantities could be considered. As for the psychiatric diagnosis example, we focused on how well the patients are classified to one of the syndrome classes and whether the conditional independence assumption of the model holds. Different checks could have been presented as well (see for instance, Hoijtink and Molenaar (1997)). Our purpose was not to give an overview of possible posterior predictive checks but to illustrate that posterior predictive checking is a usefool tool for examining the fit of a mixture model to the data.

Acknowledgements

The research reported in this paper was carried out while the first author was employed at the Catholic University Leuven and was supported by the Research Fund of the K.U. Leuven (grant OT/96/10) and the U.S. National Science Foundation. The authors are grateful to Jan Beirlant and Geert Verbeke for their helpful comments on a previous draft of this paper.

References

Berkhof, J., Van Mechelen, I. and Gelman, A. (2002). Posterior predictive checking using antisymmetric discrepancy functions. Manuscript.

Berger, J. O. and Sellke, T. (1987). Testing a point null hypothesis: the irreconcilability of P values and evidence (with discussion). *J. Amer. Statist. Assoc.* **82**, 112-122.

Casella, G. and Berger, R. L. (1987). Reconciling Bayesian and frequentist evidence in the one-sided testing problem (with discussion). *J. Amer. Statist. Assoc.* **82**, 106-111.

Celeux, G., Hurn, M. and Robert, C. P. (2000). Computational and inferential difficulties with mixture posterior distributions. *J. Amer. Statist. Assoc.* **95**, 957-970.

Chib, S. (1995). Marginal likelihood from the Gibbs output. *J. Amer. Statist. Assoc.* **90**, 1313-1321.

Chib, S. and Jeliazkov, I. (2001). Marginal likelihood from the Metropolis-Hastings output. *J. Amer. Statist. Assoc.* **96**, 270-281.

Clogg, C. C. and Goodman, L. A. (1984). Latent structure analysis of a set of multidimensional contingency tables. *J. Amer. Statist. Assoc.* **79**, 762-771.

Cochran, W. G. (1977). *Sampling Techniques*, 3rd edition. Wiley, New York.

DiCiccio, T. J., Kass, R. E., Raftery, A. and Wasserman, L. (1997). Computing Bayes factors by combining simulation and asymptotic approximations. *J. Amer. Statist. Assoc.* **92**, 903-915.

Diebolt, J. and Robert, C. P. (1994). Estimation of finite mixture distributions through Bayesian sampling. *J. Roy. Statist. Soc. B* **56**, 363-375.

Gelfand, A. E. and Dey, D. K. (1994). Bayesian model choice: asymptotics and exact calculations. *J. Roy. Statist. Soc. B* **56**, 501-514.

Gelfand, A. E. and Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *J. Amer. Statist. Assoc.* **85**, 398-409.

Gelman, A., Carlin, J. B., Stern, H. S. and Rubin, D. B. (1995). *Bayesian Data Analysis*. Chapman and Hall, London.

Gelman, A., Meng, X.-L. and Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies (with discussion). *Statist. Sinica* **6**, 733-807.

Gelman, A. and Rubin, D. B. (1992). Inferences from iterative simulation using multiple sequences (with discussion). *Statist. Science* **7**, 457-511.

Goodman, L. A. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika* **61**, 215-231.

Hoijtink, H. and Molenaar, I. W. (1997). A multidimensional item response model: constrained latent class analysis using the Gibbs sampler and posterior predictive checks. *Psychometrika* **62**, 171-189.

Hoijtink, H. (1998). Constrained latent class analysis using the Gibbs sampler and posterior predictive p-values: Applications to educational testing. *Statist. Sinica* **8**, 691-711.

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

Kass, R. E. and Raftery, A. E. (1995). Bayes factors. *J. Amer. Statist. Assoc.* **90**, 773-795.

Meulders, M., Gelman, A., Van Mechelen, I. and De Boeck, P. (1998). Generalizing the probability matrix decomposition model: an example of Bayesian model checking and model expansion. In *Assumptions, Robustness and Estimation Methods in Multivariate Modeling* (Edited by J. Hox).

Meng, X.-L. (1994). Posterior predictive p-values. *Ann. Statist.* **22**, 1142-1160.

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

Neal, R. M. (1996). Sampling from multimodal distributions using tempered transitions. *Statist. Comput.* **6**, 353-366.

Neal, R. M. (1999). Erroneous results in "Marginal likelihood from the Gibbs output". Manuscript. www.cs.utoronto.ca/∼radford/chib-letter.pdf.

Reiser, M. and Lin, Y. (1999). A goodness-of-fit test for the latent class model when expected frequencies are small. *Sociological Methodology* **29**, 81-111.

Rubin, D. B. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *Ann. Statist.* **12**, 1151-1172.

Rubin, D. B. and Stern, H. S. (1994). Testing in latent class models using a posterior predictive check distribution. In *Latent Variable Analysis: Applications for Developmental Research*, (Edited by A. Von Eye and C. C. Clogg), pp.420-438. Sage, Thousand Oaks, CA.

Van Mechelen, I. and De Boeck, P. (1989). Implicit taxonomy in psychiatric diagnosis: A case study. *J. Social and Clinical Psychology* **8**, 276-287.

Department of Clinical Epidemiology and Biostatistics, Free University Medical Center, PO Box 7057, 1007MB Amsterdam, the Netherlands.

E-mail: h.berkhof@vumc.nl

Department of Psychology, Catholic University Leuven, Tiensestraat 102, 3000, Leuven Belgium.

E-mail: iven.vanmechelen@psy.kuleuven.ac.be

623 Mathematics Bldg, Department of Statistics, Columbia University, New York, NY10027.

E-mail: gelman@stat.columbia.edu

(Received December 2000; accepted September 2002)