A Bayesian approach to the selection and testing of latent class models

Johannes Berkhof and Iven van Mechelen

University of Leuven

Andrew Gelman

Columbia University

November 23, 2000

Please send all correspondence concerning this paper to Johannes Berkhof, Department of Psychology, Tiensestraat 102, B-3000 Leuven, Belgium (e-mail: hans.berkhof@psy.kuleuven.ac.be). The research reported in this paper 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.
A Bayesian approach to the selection and testing of latent class models

Abstract

An important part of a latent class analysis concerns the selection of the number of latent classes. In this paper, we discuss the Bayes factor as a selection tool. The discussion will focus on two aspects: (i) the computation of the Bayes factor and (ii) prior sensitivity. To deal with prior sensitivity, we propose to extend the model with a prior for the hyperparameters. 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: Bayes factor, Hyperprior, Latent class model, Posterior predictive check, Prior sensitivity, Psychiatric diagnosis
1 Introduction

Latent class analysis (Lazarsfeld, 1950; Goodman, 1974; Haberman, 1979) is a widely known method for analyzing dependencies among the observations in a frequency table. The latent class model accounts for the dependencies by assuming a mixture of multivariate independent Bernoulli distributions, each component of the mixture corresponding to a single class of subjects.

As for all mixture models, the specification of a latent class model involves the selection of the number of mixture components. This selection procedure can be performed in 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 Bayesian 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 favour 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 favour of the null. A second reason for computing the Bayes factor is that it can used when comparing nonnested models. This makes the Bayes factor especially suitable for use in constrained latent class 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 criticised 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 (possibly outside the latent class framework). Bayes factors and goodness-of-fit tests may therefore be applied simultaneously albeit for different
purposes, which are model selection and examination of model fit.

In the present paper we will discuss the Bayes factor as a model selection criterion in latent class analysis 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 the highest posterior probability (within a specified class of models) but also has a reasonable fit.

The remainder of this paper is divided in five sections. After formalizing the latent class model (Section 2), the definition and computation of the Bayes factor by means of Chib's method (1995) is discussed (Section 3). The latter method is based on the Gibbs output and is especially suitable for latent class models. Yet, we must slightly modify Chib's estimator of the Bayes factor because the posterior has permutable modes of which usually only a few are covered by the posterior output. A numerical correction for the Chib estimator will therefore be presented. The prior sensitivity of the Bayes factor is also discussed. In Section 4, a hierarchical Bayes procedure is presented for estimating the hyperparameters of the prior distribution. Some goodness-of-fit test quantities 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; Meng, 1994; Gelman et al., 1996). The final Section 6 contains some concluding remarks. The points made in the paper are illustrated by fitting a latent class model to psychiatric judgement data (Van Mechelen and de Boeck, 1989).

2 Latent class model for binary data

The 0-1 scores of $N$ subjects on $J$ items can be arranged in an $N \times J$ matrix $X$ of which the $i$-th row is denoted by $X_i = (x_{i1}, \ldots, x_{iJ})$. We assume that the dependencies in the $N$ patterns are captured by $Q$ latent classes, each subject being member of a single class. The membership of subject $i$ is
represented by the unobservable variable \( Z_i = (z_{i1}, \ldots, z_{iQ}) \) which \( q \)-th element equals 1 if subject \( i \) is member of class \( q \), and 0 otherwise.

Conditional on the value of \( Z_i \), the probability of scoring 1 on item \( j \) is independent of the probability of the score on item \( k \) (\( j \neq k \)). Hence, if subject \( i \) is a member of class \( q \), the likelihood of the response pattern of subject \( i \) can be written as

\[
p(X_i | \pi, z_{iq} = 1) = \prod_{j=1}^{J} \pi_{jq}^{x_{ij}} (1 - \pi_{jq})^{1-x_{ij}},
\]

where \( \pi_{jq} \) is the probability of scoring 1 on item \( j \) given \( z_{iq} = 1 \) and \( \pi \) is defined as \( \pi = \{ \pi_{jq}; j = 1, \ldots, J; q = 1, \ldots, Q \} \). Because the membership of subject \( i \) is unknown, the likelihood of \( X_i \) is a mixture probability. If we denote the vector of mixing probabilities by \( \lambda = (\lambda_1, \ldots, \lambda_Q)^t \), the likelihood of subject \( i \) equals

\[
p(X_i | \pi, \lambda) = \sum_{q=1}^{Q} \lambda_q p(X_i | \pi, z_{iq} = 1).
\]

For \( Z_i \), we assume

\[
(Z_i | \pi, \lambda) \sim \text{Multinomial}(1, \lambda),
\]

and we further choose independent prior distributions for \( \lambda \) and \( \pi \):

\[
\lambda \sim \text{Dirichlet}(1, \ldots, 1); \quad \pi_{jq} \sim \text{Beta}(\alpha, \beta).
\]

The posterior distributions of \( \lambda \) and \( \pi \) can be computed using chained data augmentation (Diebolt and Robert, 1994; Rubin and Stern, 1994), alternately drawing from \( p(\pi, \lambda | Z, X) \) and \( p(Z | \pi, \lambda, X) \). These two densities are well-known: Conditional on \( X \) and \( Z \), the model parameters \( \lambda \) and \( \pi \) are distributed as

\[
\lambda | X, Z \sim \text{Dirichlet}(1 + \sum_{i=1}^{N} z_{i1}, \ldots, 1 + \sum_{i=1}^{N} z_{iQ})^t,
\]

\[
\pi_{jq} | X, Z \sim \text{Beta}(\alpha + \sum_{i=1}^{N} z_{iq} x_{ij}, \beta + \sum_{i=1}^{N} z_{iq} (1 - x_{ij})),
\]
and the distribution for the $i$-th row of $Z$ given $X$, $\lambda$, and $\pi$, is a multinomial distribution with parameters
\[
p(z_{iq} = 1|\pi, \lambda, X) = \frac{p(X_i|\pi, z_{iq} = 1)\lambda_q}{\sum_{q=1}^{Q} p(X_i|\pi, z_{iq} = 1)\lambda_q}, \quad q = 1, \ldots, Q.
\] (7)

3 Model selection

3.1 Definition of the Bayes factor

Suppose we have models $M_1$ and $M_2$. The Bayes factor is formally defined as the ratio of the posterior odds to the prior odds:
\[
BF_{12} = \frac{p(M_1|X)}{p(M_2|X)} / \frac{p(M_1)}{p(M_2)}
\] (8)

We see that if the prior model probabilities are equal, then $BF_{12}$ is larger than 1 if $M_1$ has a higher posterior probability. For computational purposes, it is more convenient to write the Bayes factor as the ratio of marginal likelihoods:
\[
BF_{12} = \frac{p(X|M_1)}{p(X|M_2)}.
\] (9)

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 $M$ can be expressed as
\[
p(X|M) = \int p(X|\pi, \lambda, M) p(\pi, \lambda|M) d\{\pi, \lambda\}.
\] (10)

The integral (10) can only be evaluated analytically for the one class model (i.e. the independence model). For latent class models with more than one class, we have to approximate the marginal likelihood. Common approximation methods like importance sampling are most effective when the posterior is unimodal as noted by DiCiccio et al. (1997). However, the posterior distribution 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

$$p(X) = \frac{p(X|\pi^*, \lambda^*) p(\pi^*, \lambda^*)}{p(\pi^*, \lambda^*|X)}, \quad (11)$$

which holds for any $$(\pi^*, \lambda^*)$$. Conditioning on $M$ is omitted in (11) to retain short expressions. The prior probability $p(\pi^*, \lambda^*)$ and the likelihood value $p(X|\pi^*, \lambda^*)$ can be directly computed. The posterior probability $p(\pi^*, \lambda^*|X)$ can be estimated from the Gibbs output as

$$\hat{p}_1(\pi^*, \lambda^*|X) = \frac{1}{T} \sum_{t=1}^{T} p(\pi^*, \lambda^*|X, Z(t)),$$ 

where $Z(t)$ is the $t$-th draw from $p(Z|X)$ (Gelfand and Smith, 1990). Substituting (12) in (11) yields

$$\hat{p}_1(X) = \frac{p(X|\pi^*, \lambda^*) p(\pi^*, \lambda^*)}{\hat{p}_1(\pi^*, \lambda^*|X)}.$$ 

(13)

For $$(\pi^*, \lambda^*)$$, we may choose one of the $Q!$ posterior modes. A rough approximation of the mode will usually suffice.

The probability of switching from one model region to one of the other $Q! - 1$ modal region 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), it is not always clear how to choose these constraints so that the sampler will mix efficiently. With regard to the computation of the Bayes factor, the lack of mixing can be handled by extending the Gibbs sequence $\{Z(t); t = 1, \ldots, T\}$ with $(Q! - 1)$ sequences of $T$ draws. Each of these sequences is obtained by performing a different permutation switch on the values of $\{Z(t); t = 1, \ldots, T\}$; the $s$-th sequence $(s = 1, \ldots, Q!)$ will be denoted as $\{Z_{s(t)}; t = 1, \ldots, T\}$. The nonpermuted Gibbs sequence corresponds to $s = 1$. The estimator of
\[ p(\pi^*, \lambda^*|X) \] now becomes

\[
\hat{p}_{\text{ext}}(\pi^*, \lambda^*|X) = \frac{1}{Q!T} \sum_{s=1}^{Q} \sum_{t=1}^{T} p(\pi^*, \lambda^*|X, Z_{s(t)}).
\] (14)

If the number of different posterior modes due to class label indeterminacy is large, it is computationally demanding to obtain the simulation estimator based on the extended Gibbs output (if the number of classes is 5, we already have 5! = 120 modes and hence 120 sequences). To account for this, a simulation-consistent estimator based on a stratification-type principle (Cochran, 1977, p. 87) is derived. This estimator is based on the nonpermuted Gibbs sequence \( Z_{1(t)}; t = 1, \ldots, T \) and a smaller number of systematic draws from the permuted sequences. The reason for splitting the sequences in two groups is that the values of \( p(\pi^*, \lambda^*|X, Z_s(t)) \) of the nonpermuted sequence tend to be more variable than the values based of the permuted sequence (which are generally small as \( \pi^* \) and \( \lambda^* \) are computed from the nonpermuted output). We estimate \( p(\pi^*, \lambda^*|X) \) as follows

\[
\hat{p}_{\text{nt}}(\pi^*, \lambda^*|X) = \frac{1}{Q!} \sum_{t=1}^{T} p(\pi^*, \lambda^*|X, Z_{1(t)}) + \frac{1}{Q!} \sum_{t=2}^{T_2} p(\pi^*, \lambda^*|X, Z_{s(t)})
\]

\[ = \frac{1}{Q!} \hat{p}_{\text{I}}(\pi^*, \lambda^*|X) + \frac{Q! - 1}{Q!} \hat{p}_{\text{II}}(\pi^*, \lambda^*|X). \] (15)

The \( T_2 \) elements of sequence \( \{u_1, \ldots, u_{T_2}\} \) are draws from \( \{1, \ldots, T\} \). Substituting \( \hat{p}_{\text{nt}}(\pi^*, \lambda^*|X) \) in (11) yields an estimator for the marginal likelihood which will be denoted by \( \hat{p}_{\text{nt}}(X) \).

### 3.3 Choosing the ratio \( T_2/T \)

In this subsection, we study the relation between the ratio \( T_2/T \) and the efficiency of the estimators for the marginal likelihood. We start by making the following two additional assumptions: (1) \( p(\pi^*, \lambda^*|X, Z_{s(t)}) \) is independent of \( p(\pi^*, \lambda^*|X, Z_{r(u)}) \) if \( r \neq s \) or \( u \neq t \), and (2) \( p(\pi^*, \lambda^*|X, Z_{s(t)}) \) has a constant coefficient of variation. These assumptions do usually not hold in practice so that the presented results are only approximately correct. Under the two above assumptions, we have

\[
\text{s.e.}\{\hat{p}_{\text{II}}(\pi^*, \lambda^*|X)\} = \left( \frac{Q! \hat{p}_{\text{nt}}(\pi^*, \lambda^*|X)}{\hat{p}_{\text{I}}(\pi^*, \lambda^*|X)} - 1 \right) \sqrt\frac{T_1}{(Q! - 1)T_2} \text{s.e.}\{\hat{p}_{\text{I}}(\pi^*, \lambda^*|X)\}
\]
\[
\approx \left( \frac{1}{\rho} - 1 \right) \sqrt{\frac{T_1}{(Q! - 1)T_2}} \text{s.e.}\{\hat{p}_1(\pi^*, \lambda^*|X)\}, \tag{16}
\]

where
\[
\rho = \frac{\hat{p}_1(\pi^*, \lambda^*|X)}{Q! \hat{p}_{\text{ext}}(\pi^*, \lambda^*|X)}. 
\]

The term \(\rho\) expresses the degree of bad mixing and varies from 1/\(Q!\) in case all modal regions are equally well explored to 1 in case only one modal region is explored.

If all modal regions are equally well explored, there is no need for stratification and we may use the estimator \(\hat{p}_1(\pi^*, \lambda^*|X)\). By substituting (16) in (15), we find that in that case the estimator \(\hat{p}_{\text{st}}(\pi^*, \lambda^*|X)\) has a smaller standard error than \(\hat{p}_1(\pi^*, \lambda^*|X)\) if
\[
\frac{T_2}{T} > \frac{1}{Q! + 1}.
\]

If the modal regions are not equally well explored, the estimator \(\hat{p}_{\text{st}}(\pi^*, \lambda^*|X)\) is already more efficient than \(\hat{p}_1(\pi^*, \lambda^*|X)\) at a value smaller than 1/(\(Q! + 1\)). Therefore, a rule of thumb for choosing \(T_2/T\) is that by setting \(T_2/T\) equal to 1/(\(Q! + 1\)), the efficiency of the estimator does not drop when using \(\hat{p}_{\text{st}}(\pi^*, \lambda^*|X)\) instead of \(\hat{p}_1(\pi^*, \lambda^*|X)\). We can also derive the allocation between \(T_2\) and \(T\) that yields the most efficient estimator if the total sample size \((T + (Q! - 1)T_2)\) is fixed. By substituting (16) in (15), we obtain the optimal ratio
\[
\frac{T_2}{T} \approx \frac{1}{Q! - 1} \left( \frac{1}{\rho} - 1 \right)^{2/3}. \tag{17}
\]

### 3.4 Example: Psychiatric diagnosis data

We consider data collected by Van Mechelen and De Boeck (1989). The data 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. We assume that the dependencies in the symptom patterns are captured by \(Q\) latent classes, each patient being member of a single patient category. We chose a Dirichlet\((1, \ldots, 1)\) prior distribution for the class
<table>
<thead>
<tr>
<th>Symptom label</th>
<th>Patients</th>
</tr>
</thead>
<tbody>
<tr>
<td>disorientation</td>
<td>. . . . . . . . x . . . . . . . . . . . . . .</td>
</tr>
<tr>
<td>obsession/compulsion</td>
<td>..................................................</td>
</tr>
<tr>
<td>memory impairment</td>
<td>..................................................</td>
</tr>
<tr>
<td>lack of emotion</td>
<td>..................................................</td>
</tr>
<tr>
<td>antisocial impulses or acts</td>
<td>..................................................</td>
</tr>
<tr>
<td>speech disorganization</td>
<td>..................................................</td>
</tr>
<tr>
<td>overt anger</td>
<td>..................................................</td>
</tr>
<tr>
<td>grandiosity</td>
<td>..................................................</td>
</tr>
<tr>
<td>drug abuse</td>
<td>..................................................</td>
</tr>
<tr>
<td>alcohol abuse</td>
<td>..................................................</td>
</tr>
<tr>
<td>retardation</td>
<td>..................................................</td>
</tr>
<tr>
<td>belligerence/negativism</td>
<td>..................................................</td>
</tr>
<tr>
<td>somatic concerns</td>
<td>..................................................</td>
</tr>
<tr>
<td>suspicion/ideas of persecution</td>
<td>..................................................</td>
</tr>
<tr>
<td>hallucinations/delusions</td>
<td>..................................................</td>
</tr>
<tr>
<td>agitation/excitement</td>
<td>..................................................</td>
</tr>
<tr>
<td>suicide</td>
<td>..................................................</td>
</tr>
<tr>
<td>anxiety</td>
<td>..................................................</td>
</tr>
<tr>
<td>social isolation</td>
<td>..................................................</td>
</tr>
<tr>
<td>inappropriate affect or behaviour</td>
<td>..................................................</td>
</tr>
<tr>
<td>depression</td>
<td>..................................................</td>
</tr>
<tr>
<td>leisure time impairment</td>
<td>..................................................</td>
</tr>
<tr>
<td>daily routine impairment</td>
<td>..................................................</td>
</tr>
</tbody>
</table>

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.

weights vector \( \lambda \) and independent \( Beta(\alpha, \alpha) \) prior distributions for the class dependent symptom probabilities \( \pi_{j|q} \). We set \( \alpha \) equal to 0.5, 1, and 2 to examine the prior sensitivity of the Bayes factor.

We estimated models with one to five latent classes. Regarding posterior simulation, we simulated 10 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 \( \sqrt{R} \) was smaller than 1.1 for all model parameters (Gelman and Rubin, 1992). For computing \( \sqrt{R} \), the parameter labels need to be identified. This permutation problem can be handled by computing \( \sqrt{R} \) from the Gibbs output after having reordered the draws such that they all come from the same model region. The reordering can be done by applying a \( Q! \)-means type of
Figure 1: Log marginal likelihoods (with standard errors) as a function of $\alpha$ and the number of classes, computed using $T = 10^6$ and four choices of $T_2$. The model is $\lambda \sim Dirichlet(1, \ldots, 1)$; $\pi_{ij|q} \sim Beta(\alpha, \alpha)$. clustering analysis to the Gibbs output (see Celeux et al., 2000). The mode $(\theta^*, \lambda^*)$ which is needed for estimating the marginal likelihood was computed as $\text{argmax}_{(t)} \{ p(X|\pi_{(t)}, \lambda_{(t)})p(\pi_{(t)}, \lambda_{(t)}) \}$ over the draws of the first chain. The other chains were permuted such that each first draw and the mode $(\theta^*, \lambda^*)$ come from the same modal region. 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 1 presents the values of the logarithm of the estimated marginal likelihood $\hat{p}_\text{st}(X)$. The standard error of $\log \hat{p}(X)$ is based on the variability among the independent single-chain estimates of $\log p(\theta^*, \lambda^*|X)$. We can see from Figure 1 that $T_2$ has an effect on the precision of the estimate $\log \hat{p}_\text{st}(X)$ if the number of classes is at least equal to three. There is no effect for the two class model since in that situation there is no mixing between the permuted and non-permuted chain. We further see that setting $T_2$ equal to $10^3$ yields a relatively imprecise estimate of $p(X)$ which is even much larger than the standard error obtained if we set $T_2$ equal to zero. The latter can be explained as follows. In Subsection 3.3, we derived a lower bound bound of $1/(Q! + 1)$ being the value of $T_2/T$ at which $\hat{p}_\text{st}(X)$ is at least as efficient as $\hat{p}_I(X)$. The standard error of $\log \hat{p}_I(X)$ can be read from Figure 1 since it is equal to the standard error of $\log \hat{p}_\text{st}(X)$ if we set $T_2$ equal to
zero. For the 3 to 5-class model, the lower bound is much larger than $T_2/T = 10^{-3}$. Therefore, the standard error of $\log \hat{p}_{st}(X)$ is likely to be larger for $T_2 = 10^3$ than for $T_2 = 0$.

Figure 1 shows that the log marginal likelihood is rather sensitive to the prior distribution. If the models are compared at $\alpha = \beta = 2$, there is a preference for the one class model. At $\alpha = \beta = 1$, there is equal preference for the 2- or 3-class model while at $\alpha = \beta = 0.5$, there is a preference for the 3-class model. Because of this ambiguity, it seems advisable to include prior information about the model parameters. However, in practical latent class modelling prior information is not always available or too vague to rely on. In that case, a noninformative approach can be followed which consists of estimating the hyperparameters $\alpha$ and $\beta$ rather than setting them at a priori selected values.

4 A hierarchical extension of the model

4.1 Specification of the model

For the hyperparameters $\alpha$ and $\beta$, we follow Gelman et al. (1995) and choose a diffuse hyperprior density that is uniform on $\left(\frac{\alpha}{\alpha+\beta}, \frac{1}{\alpha+\beta}\right)$, in the range $\alpha/(\alpha + \beta) \in (0, 1)$ and $1/(\alpha + \beta) \in (0, c)$ $(c > 0)$. The expression $\alpha/(\alpha + \beta)$ is the mean of the $Beta(\alpha, \beta)$ distribution and $1/(\alpha + \beta)$ is a measure of dispersion (Gelman et al., 1995, p. 131).

Posterior simulation consists of subsequently drawing from $p(\pi, \lambda|X, Z, \alpha, \beta)$, $p(Z|X, \pi, \lambda)$, and $p(\alpha, \beta|\pi, \lambda)$. Because the latter distribution does not have a known form, the last step of the Gibbs cycle is replaced by a Metropolis step. As a jumping distribution, we may choose a uniform symmetric distribution around the current values of $\alpha/(\alpha + \beta)$ and $1/(\alpha + \beta)$.
Two classes

Three classes

Four classes

Five classes

Figure 2: Histograms of a posterior draw of the conditional symptom probabilities and $Beta(\alpha, \beta)$ density functions with $\alpha$ and $\beta$ set equal to their posterior modes. The beta densities are rescaled in order to match with the histograms.

4.2 Computation of the marginal likelihood

The estimator of the marginal likelihood is based on identity (11), where as before the posterior probability $p(\pi^*, \lambda^*|X)$ is estimated by stratified sampling from the non-permuted and permuted Gibbs output. The estimators $\hat{p}_I(\pi^*, \lambda^*|X)$ and $\hat{p}_{II}(\pi^*, \lambda^*|X)$ are defined as

$$
\hat{p}_I(\pi^*, \lambda^*|X) = \frac{1}{T} \sum_{t=1}^{T} p(\pi^*, \lambda^*|X, Z_{1(t)}, \alpha_{(t)}, \beta_{(t)})
$$

$$
\hat{p}_{II}(\pi^*, \lambda^*|X) = \frac{1}{(Q! - 1)T_2} \sum_{s=2}^{Q} \sum_{t=1}^{T_2} p(\pi^*, \lambda^*|X, Z_{s(u_t)}, \alpha_{(u_t)}, \beta_{(u_t)}) .
$$

The estimation of $p(X)$ also involves approximating the prior probability $p(\pi^*)$ which cannot be directly computed anymore. We write $p(\pi^*)$ as the following double integral:

$$
p(\pi^*) = \int_0^{\infty} \int_{\max(0, \frac{1}{2} - \alpha)}^{\infty} p(\pi^*|\alpha, \beta) p(\alpha, \beta) d\beta d\alpha .
$$

The double integral can be approximated by means of bridge sampling (Meng and Wong, 1996) or importance sampling (diCiccio et al., 1998). The precision of the estimator may be improved by replacing $(\alpha, \beta)$ by $(u, v) = (\log(\alpha/\beta), \log(\alpha + \beta))$, since the posterior distribution of $(u, v)$ tends to be less skewed than that of $(\alpha, \beta)$. 

13
4.3 Example (continued)

We simulated 10 independent sequences, 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. For the Metropolis step, we chose a symmetric uniform jumping distribution yielding an acceptance rate larger than 0.2 for all models. We set $T$ and $T_2$ equal to $10^6$ and $10^5$ and approximated $p(\pi^*)$ by a Laplace bridge sampling estimate (Meng and Wong, 1996; diCiccio et al., 1998). We based the latter estimate on 50,000 draws from the Gibbs output and 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 $-343.6$, $-337.5$, $-332.5$, $-332.4$, and $-332.5$ for the models with 1 to 5 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 either the 3-, the 4- or the 5-class model, presumably because the number of patients is too small to be able to draw a distinction between these models. The results are different from the results in Section 3.4, in particular when $\alpha$ and $\beta$ are fixed at 1 or at 2. For the latter values of the hyperparameters, the 3-class model is not preferred to the 2-class model. However, in the hierarchical model, the Bayes factor of the 3-class model versus the 2-class model is $\exp(-260.3)/\exp(-265.3) \approx 150$, which means that under equal prior model probabilities, the 3-class model has 150 times the posterior probability of the 2-class model.

By using a hierarchical model, we have not eliminated the prior sensitivity of the marginal likelihood but rather have shifted it to a higher level. The presented hierarchical extension is therefore only appropriate when the assumption of the same $Beta(\alpha, \beta)$ distribution for all conditional symptom probabilities is reasonable. To examine this, we constructed for each of the models with at least two classes a histogram of a posterior draw of the conditional symptom probabilities. The plotted curves are $Beta(\alpha, \beta)$ densities with $\alpha$ and $\beta$ set equal to the posterior modes of the hyperparameters. We see in Figure 2 that the curves and the histograms are similarly shaped which suggests that the $Beta(\alpha, \beta)$ model is consistent with this aspect of the data.
disorientation | class 1 | class 2 | class 3
---|---|---|---
obssession/compulsion | | | |
memory impairment | | | |
lack of emotion | | | |
antisocial impulses or acts | | | |
speech disorganization | | | |
overt anger | | | |
grandiosity | | | |
drug abuse | | | |
alcohol abuse | | | |
retardation | | | |
belligerence/negativism | | | |
somatic concerns | | | |
suspicion/ideas of persecution | | | |
hallucinations/delusions | | | |
agitation/excitement | | | |
suicide | | | |
anxiety | | | |
social isolation | | | |
inappropriate affect or behaviour | | | |
depression | | | |
leisure time impairment | | | |
daily routine impairment | | | |

Figure 3: Posterior medians and 50% intervals for the probability of each symptom being present, for each of the three classes. Each estimate and interval is overlain 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.

The posterior modes of the parameters of the 3-class model are presented in Figure 3. To interpret the classes, it is helpful to search for symptoms for which the mode of $\pi_{jq}$ is near 1 for one value of $q$ and near 0 for the other values of $q$. It follows that class 1 is associated with the symptoms agitation/excitement, suspicion/ideas of persecution, and hallucinations/delusions, being indicative for a psychosis syndrome. Class 2 is associated with depression, anxiety, and suicide and can be interpreted as an affective syndrome, and class 3 is primarily associated with alcohol abuse, and may therefore be considered an implicit addiction syndrome.
5 Posterior predictive checks

5.1 Goodness-of-fit tests

As stated in the introduction, the Bayes factor does not solve the issue of model fit, that is, it does not tell whether the selected model could have plausibly generated the observed data. Since we would interpret the results from a badly and well fitting model differently, it makes sense to perform goodness-of-fit tests in addition to selecting a model by means of the Bayes factor. The goodness-of-fit test is then used as a diagnostic model check which may help us improve the specification of the model. The replicates are drawn from the posterior predictive distribution, $p(X^{\text{rep}}|X)$. The sampled vectors from the posterior predictive distribution will be denoted by $X^{\text{rep}}_1, \ldots, X^{\text{rep}}_R$. The statistical significance can be summarized by the exceeding tail area probability called posterior predictive $p$-value (Rubin, 1984; Gelman et al., 1996; Meng, 1994). 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, however, we will focus on absolute test quantities. The 3-class model for the psychiatric diagnosis data is taken as the null model.

In a latent class model, each subject 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 subject $i$. The posterior mean of $(z_{i1}, z_{i2}, z_{i3})$ in the 3-class model for the psychiatric diagnosis data contains a value larger than 0.9 for 21 out of 30 patients. This shows that most of the patients are well classified by the 3-class model. For the other 9 patients, it is of interest to check whether their response patterns are well fitted by the distributions of the corresponding patient classes. If this is not the case, the patient might belong to a patient class different from the ones included in the model. To examine this, we defined the following discrepancy measure:

$$D_1(X, \pi, Z) = \sum_{j=1}^{J} |x_{ij} - \pi_{j|q}| I_{z_{i|q}}(1).$$
The posterior predictive p-value can be read from a plot of $D(X^{rep}, \theta)$ versus $D(X, \theta)$ (Gelman et al., 1995). In Figure 4, such a plot is presented for patient 21 based on 1000 draws from the posterior predictive distribution. The discrepancy for the observed data is estimated to be about 1.5 times as high as in the replications. 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 higher, but not extremely higher, than what might occur under replications under the model.

For the other 8 subjects, the posterior predictive p-values are between 0.3 and 0.6. The p-value for patient 21 is relatively small 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 depressive and psychotic features.

An important assumption of the latent class model is that within each class, the scores on the items are independent. This assumption can be tested for each class separately by a posterior predictive check with discrepancy measure

$$D_2(X, \pi, Z) = \sum_{j=1}^{J-1} \sum_{k>j} \frac{\sum_{i=1}^{I} z_{iq}!}{n_{00q}! \cdot n_{01q}! \cdot n_{10q}! \cdot n_{11q}!} \cdot p(x_j, x_k | z, \pi) ,$$

where $n_{11q} = \sum_i z_{iq} x_{ij} x_{ik}$, $n_{01q} = \sum_i z_{iq} (1 - x_{ij}) x_{ik}$, $n_{10q} = \sum_i z_{iq} x_{ij} (1 - x_{ik})$, and $n_{00q} =$
\[ \sum_i z_{iq}(1 - x_{ij})(1 - x_{ik}), \text{ and with } x_j \text{ being the } j\text{-th column of } X. \] The probability \( p(x_j, x_k | z, \pi) \) is premultiplied by a factor so that it becomes the multinomial density of the four response patterns \( n_{11q}, n_{01q}, n_{10q}, \) and \( n_{00q}. \) The discrepancy measure \( D_2(X, \pi, Z) \) 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 Reiser and Lin (1999) and Hoijtink (1998). 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. A problem with \( D_2(X, \pi, Z) \) is that it is assumed that class \( q \) is identified. As mentioned in Section 3.4, this problem can be dealt with by applying a \( Q! \)-means type of clustering analysis to the Gibbs output and permuting the values in the non-identification clusters (see Celeux et al., 2000). 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 learned from our latent class analysis that, although Bayes factors and posterior predictive checks stem from quite different research traditions, they can meaningfully combined and supplement each other with complementary information.

Regarding the Bayes factor for the latent class model, 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 confined ourselves to estimation of the marginal likelihood using output from the Gibbs sampler but the presented estimator can also be applied when the posterior is explored using a Metropolis scheme or a tempered annealing scheme (Neal, 1996; Celeux et al., 2000). Then we have only draws from the posterior distribution $p(\pi, \lambda | X)$ and we need to complement them with draws from the conditional predictive distribution $p(Z | \pi, \lambda, X)$. The marginal likelihood is again computed as the average over draws from the conditional predictive distribution. 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 favour 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 under consideration, 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; if the hyperparameter is set to 0.5, the 3-class model is selected, but the 1-class model is selected if the hyperparameter is set at 2. A hierarchical extension of the latent class model was shown to provide a neat way out.

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, 1998), but we confined ourselves to checks that capture aspects of the latent class model that are of key interest when analyzing psychiatric judgement data.

References


