Generalizing the probability matrix decomposition model: an example of Bayesian model checking and model expansion*

Michel Meulders, Psychology Department, University of Leuven, Belgium
Andrew Gelman, Department of Statistics, Columbia University, New York, USA
Iven Van Mechelen, Psychology Department, University of Leuven, Belgium
Paul De Boeck, Psychology Department, University of Leuven, Belgium

January 8, 1998

Abstract

Probability matrix decomposition (PMD) models can be used to explain observed associations between two sets of elements. More specifically, observed associations are modeled as a deterministic function of B latent Bernoulli variables that are realized for each element. To estimate the parameters of this model, a sample of the posterior distribution is computed with a data augmentation algorithm. The obtained posterior sample can also be used to assess the fit of the model with the technique of posterior predictive checks. In this paper a PMD model is applied to data on psychiatric diagnosis. In checking the model for this analysis, we focus on the appropriateness of the prior distribution for a set of latent parameters. Based on the posterior distribution for the values of the parameters corresponding to the observed data, we conclude that a relatively flat prior distribution is inappropriate. In order to solve this problem, a mixture prior density with two beta distributed components is used to expand the model in a meaningful way.

Keywords: discrete data, mixture model, model checking, model expansion, prior distribution, posterior predictive check, psychometrics

1 Introduction

A crucial aspect of statistical analysis is checking the fit of models to observed data and, when appropriate, altering and generalizing the models to fit better. Gelman and Meng (1996) suggest a paradigm for model checking and improvement based on posterior predictive checks and continuous model expansion. In this paper, we use these tools to generalize a model in psychometrics. The conclusion of our analysis is *not* a flawless model, but rather one that is improved, in important aspects, compared to what came before. A key theme that runs through this example is that when a model is interpreted Bayesianly, it can be compared to data in a variety of ways. Model checking can be particularly effective for scientific understanding if it is used, not merely as an evaluation

*To appear in *Assumptions, Robustness, and Estimation Methods in Multivariate Modeling*, ed. J. Hox. This work was supported in part by Fellowship F/96/9 and Grant OT/96/10 of the Research Council of Katholieke Universiteit Leuven and Young Investigator Award DMS-9457824 and grant SBR-9708424 of the U.S. National Science Foundation.

device ("Our data pass a series of tests, thus we do not reject the model") but rather as a creative tool for uncovering aspects of reality that are imperfectly captured by the existing model.

For the rest of this paper, we focus on a particular model and application. Probability matrix decomposition (PMD) models have been introduced by Maris, De Boeck, and Van Mechelen (1996). In general, these models may be used to explain observed binary associations between two sets of elements, usually denoted by objects and attributes. In order to explain observed associations it is assumed that B latent Bernoulli variables are realized at both the object and the attribute side. Each latent variable indicates whether an object or an attribute has one of B latent features. Furthermore it is assumed that observed binary associations are a deterministic function of the realized latent variables.

With the PMD model, the distribution of both latent and observed variables, or rather the distribution of the augmented data, has a mathematically tractable form. As a result the EM algorithm (Dempster, Laird, and Rubin, 1977) may be used to compute maximum likelihood or posterior mode estimates of the model (Maris, De Boeck, and Van Mechelen, 1996). As an alternative, one could also turn to a fully Bayesian framework in order to solve the estimation problem for the PMD model. As a matter of fact one could use a data augmentation algorithm (Tanner and Wong, 1987) to compute a sample of the entire observed posterior distribution. As with the EM algorithm, data augmentation is most conveniently computed using properties of mathematical tractability of the augmented posterior distribution. Apart from solving the estimation problem, simulation of draws from the posterior distribution has three important advantages (see Gelman et al., 1995): (1) the computation of posterior intervals of any estimands of interest is straightforward, (2) the posterior sample can be used to investigate problems of local maxima and to trace trade-off relations between parameters in the model and (3) the posterior sample can be used to check the fit of the model using posterior predictive checks. Meulders et al. (1997) implemented a data augmentation algorithm for the PMD model and discuss each of the above-mentioned advantages.

In this paper we first briefly reconsider the model and the estimation of its parameters with a data augmentation algorithm. Second, we summarize how to use the sample from the posterior distribution in order to assess the fit of the model with the technique of posterior predictive checks. Third, we use the PMD model to analyze data on psychiatric diagnosis. In the phase of model checking we focus on the appropriateness of the prior distribution for a set of latent parameters. From the posterior distribution of the ensemble of these parameters, we conclude that a relatively flat prior density is not appropriate for these data. Consequently, in order to expand the model in a meaningful way, a mixture of two beta distributed components is used as the prior distribution.

2 The PMD model

In general, PMD models are used to explain replicated observed binary associations between two sets of elements, usually denoted by objects and attributes. The two sets may for instance contain patients and symptoms, situations and responses, countries and items, etc. Replications could be defined as opinions of several psychiatrists concerning whether a patient has a symptom, the responses of different persons in a situation, the answers of different inhabitants of a country to an item, etc. Formally, the observed variable Y_i^{oa} equals 1 if object o ($o = 1, \ldots, O$) has attribute a ($a = 1, \ldots, A$) at the ith replication ($i = 1, \ldots, I_{oa}$), and 0 otherwise.

The following two aspects of the model are needed in order to explain observed associations between objects and attributes.

1. It is assumed that B latent Bernoulli variables are realized at both the object and the attribute side. More specifically, for a triple (o, a, i), the latent variables $S_{ai}^{ob} \sim$ Bernoulli(ρ_{ob}) and $P_{oi}^{ab} \sim$ Bernoulli(τ_{ab}) ($b = 1, \ldots, B$) are realized. These variables equal 1 if object o and attribute a have feature b, and 0 otherwise.

2. It is assumed that the observed variables are a deterministic function of the corresponding latent variables. That is, $Y_i^{oa} = C(S_{ai}^{o1}, \ldots, S_{ai}^{oB}, P_{oi}^{a1}, \ldots, P_{oi}^{aB})$. Maris, De Boeck, and Van Mechelen (1996) describe several functions, which they call condensation rules. In this paper we use one of these rules, namely a disjunctive communality rule which is defined as follows:

$$Y_i^{oa} = 1 \Leftrightarrow \exists b : S_{ai}^{ob} = 1 \wedge P_{oi}^{ab} = 1 \ (b = 1, \ldots, B) \tag{1}$$

Maris, De Boeck, and Van Mechelen (1996) show that for rule (1) the probability that Y_i^{oa} equals 1 is given by:

$$\Pr(Y_i^{oa} = 1 | \rho_o, \tau_a) = 1 - \prod_b (1 - \rho_{ob}\tau_{ab}). \tag{2}$$

Now let f^{oa1} and f^{oa0} respectively denote the number of observed 1- and 0-responses with respect to the pair (o, a). Furthermore let $\mathbf{Y}$ and $\mathbf{Z}$ comprise all the observed and latent variables and let θ be the vector of all the parameters in the model. The observed posterior distribution may then be expressed as follows:

$$p(\theta|\mathbf{Y}) \propto p(\theta)p(\mathbf{Y}|\theta) = \prod_o \prod_a \Pr(Y_i^{oa} = 1|\theta)^{f^{oa1}} \Pr(Y_i^{oa} = 0|\theta)^{f^{oa0}}. \tag{3}$$

In the above expression one still has to make a choice with respect to the prior distribution $p(\theta)$. One possibility is to take $p(\theta) \propto 1$ so that the posterior density is proportional to the likelihood. However, Maris, De Boeck, and Van Mechelen (1996) show that this choice leads to computational

difficulties, because it does not guarantee the existence of maximum likelihood estimates within the interior of the parameter space. A useful alternative prior distribution, which guarantees the existence of posterior mode estimates within the boundaries of the parameter space, is $p(\theta) \sim \text{Beta}(\theta|2,2)$.

The observed posterior distribution for the PMD model is complex. However, the joint posterior distribution of observed and latent data—that is, the augmented posterior distribution $p(\theta|\mathbf{Y},\mathbf{Z})$—has a tractable form. Therefore one may easily use the EM algorithm to locate the posterior mode. Maris, De Boeck, and Van Mechelen (1996) provide details with respect to the implementation of the EM algorithm for the PMD model. An alternative approach, which also exploits the tractability of the augmented posterior density, is to use a data augmentation algorithm to compute a sample of the observed posterior distribution. Meulders et al. (1997) discuss the implementation of this algorithm for the PMD model in detail. In this paper we only briefly summarize the general scheme of this approach.

Given starting values $\theta^{(0)}$, the $(m+1)$st iteration of the data augmentation algorithm consists of the following two steps:

1. Imputation step: generate latent data $\mathbf{Z}^{(m+1)}$ from the conditional predictive distribution, $p(\mathbf{Z}|\theta^{(m)},\mathbf{Y})$.

2. Posterior step: draw a simulation $\theta^{(m+1)}$ of the parameter vector from the augmented posterior distribution, $p(\theta|\mathbf{Y},\mathbf{Z}^{(m+1)})$.

Tanner and Wong (1987) show that the subsequent values $\theta^{(1)},\theta^{(2)},\ldots$ form a Markov chain that, under some regularity conditions, converges to the posterior distribution. In practice, an important aspect is to monitor the convergence of the chain in order to determine the required number of iterations. Gelman and Rubin (1992) recommend to simulate multiple chains from different starting points and to judge approximate convergence when the statistic $\sqrt{\hat{R}}$, which measures the ratio of between- plus within-chain variation to within-chain variation, becomes close to 1 for each scalar estimand of interest. In this approach one should discard some initial iterations to exclude the influence of the starting points.

3 Model checking

Once a sample of the observed posterior distribution is available, assessment of the fit of the model is straightforward in a Bayesian framework using posterior predictive checks (Rubin, 1984, Gelman, Meng, and Stern, 1996). With this technique, model checking is basically a matter of comparing

observed data $\mathbf{Y}$ with replicated data $\mathbf{Y}^{\text{rep}}$ that could have been observed under the model if the experiment of today were replicated with the same value of θ.

To compare observed and replicated data, one usually defines a test quantity $T(\mathbf{Y})$ that is a function of the data only. Rubin (1984) defines the posterior predictive p-value as the probability that $T(\mathbf{Y}^{\text{rep}})$ exceeds or equals $T(\mathbf{Y})$. An extreme p-value indicates that $T(\mathbf{Y})$ is unlikely to have occurred under the model, which means that the model aspect measured by $T(\cdot)$ is questionable. In order to estimate the posterior predictive p-value, one has to carry out the following steps for each draw $\theta^{(l)}$ $(l = 1, \ldots, L)$ of the posterior distribution:

1. Generate $\mathbf{Y}^{\text{rep},l}$ from $p(\mathbf{Y}|\theta^{(l)})$.
2. Compute $T(\mathbf{Y}^{\text{rep},l})$.

This procedure would typically be repeated with several different test statistics $T(\cdot)$ in order to check the fit of different aspects of the model.
One can then estimate the posterior predictive p-value as the proportion of simulated values $T(\mathbf{Y}^{\text{rep},l})$ that exceed or equal $T(\mathbf{Y})$. In terms of a corresponding graphical representation, one may situate $T(\mathbf{Y})$ in the simulated reference distribution $T(\mathbf{Y}^{\text{rep}})$.

Gelman, Meng, and Stern (1996) also consider the use of test quantities $T(\mathbf{Y}, \theta)$ that are a function of both data and parameters, which they label realized discrepancy measures. In this case the posterior predictive p-value is defined as the probability that the realized discrepancy measure based on the replicated data $T(\mathbf{Y}^{\text{rep}}, \theta)$ exceeds or equals the realized discrepancy measure based on the observed data $T(\mathbf{Y}, \theta)$. For a realized discrepancy measure, the posterior predictive p-value may be estimated by the following procedure. For each draw $\theta^{(l)}$ $(l = 1, \ldots, L)$:

1. Generate $\mathbf{Y}^{\text{rep},l}$ from $p(\mathbf{Y}|\theta^{(l)}))$.
2. Compute $T(\mathbf{Y}^{\text{rep},l}, \theta^{(l)})$.
3. Compute $T(\mathbf{Y}, \theta^{(l)})$.

Subsequently, the posterior predictive p-value may be estimated as the proportion of simulated values $T(\mathbf{Y}^{\text{rep},l}, \theta^{(l)})$ that exceed or equal $T(\mathbf{Y}, \theta^{(l)})$. The corresponding graphical representation is a scatterplot of pairs $(T(\mathbf{Y}^{\text{rep},l}, \theta^{(l)}), T(\mathbf{Y}, \theta^{(l)}))$ $(l = 1, \ldots, L)$.

With the PMD model, assessment of the fit may focus on various aspects of the model. A first important aspect concerns the overall goodness of fit of the model, which can be measured using a Pearson χ^2 discrepancy measure (Gelman, Meng and Stern, 1996). For the PMD model this

discrepancy measure may be expressed as follows:

$$X^2(\mathbf{Y},\theta) = \sum_o \sum_a \left(\frac{(f^{oa1} - \mathrm{E}(f^{oa1}|\theta))^2}{\mathrm{E}(f^{oa1}|\theta)} + \frac{(f^{oa0} - \mathrm{E}(f^{oa0}|\theta))^2}{\mathrm{E}(f^{oa0}|\theta)} \right). \tag{4}$$

A second aspect concerns the relative fit of models with different numbers of features. Meulders et al. (1997) discuss the use of a likelihood ratio discrepancy measure in order to investigate this aspect of the PMD model. A similar measure was used by Rubin and Stern (1994) to determine the number of latent classes in a latent class analysis.

Besides general aspects, model checking may also focus on specific model assumptions, such as the independence of latent variables. Gelman et al. (1997) show that with the PMD model it is even meaningful to use test quantities that are functions of the latent data only. In the present paper model checking focuses on the appropriateness of the prior distribution.

4 Example

4.1 Data

In this paper we analyze data on psychiatric diagnosis collected by Van Mechelen and De Boeck (1990). In their study 15 psychiatrists were asked to judge descriptions of 30 patients with respect to 23 symptoms. Hence, the observed variables indicate whether a patient has a symptom in the opinion of several psychiatrists, whose judgments are considered to be replications. The PMD model may now be used to explain the observed variables through the following assumptions: (1) patients have each of B latent syndromes with a certain probability; (2) symptoms are characteristic to each of B latent syndromes with a certain probability; and (3) whether a patient has a symptom is a deterministic function of the syndromes associated with the patient, and of the syndromes for which the symptom is characteristic. Maris, De Boeck, and Van Mechelen (1996) analyzed these data with a PMD model involving a disjunctive communality rule, which means that a patient has a symptom if there is at least one syndrome for which it holds that the patient has that syndrome and that the symptom is characteristic of that syndrome. The authors used an EM algorithm to compute posterior mode estimates for models with from one to four syndromes. They subsequently argue that a model with three syndromes is preferable. In the following paragraphs we discuss a full Bayesian estimation and model checking for this model.

4.2 Bayesian estimation and model checking

4.2.1 Computation of posterior simulations

We simulate draws from the posterior distribution of the parameters of a disjunctive PMD model with independent $\mathrm{Beta}(\theta|2,2)$ prior distributions on the patient and symptom parameters θ_j. We

use a data augmentation algorithm to simulate four chains of 2500 iterations, of which the first 500 are discarded to remove the influence of the starting point. The convergence diagnostic $\sqrt{\hat{R}}$ is smaller than 1.1 for all parameters, so it may be concluded that further simulation would not improve much the precision of the simulated posterior distribution. We construct a sample of 2000 draws from this posterior distribution by gathering every 4th iteration of each chain (to save time in further computations, we do not keep every draw). Then, for each draw θ^l $(l = 1, \ldots, 2000)$, we draw a replicated dataset $\mathbf{Y}^{\text{rep},l}$ from the predictive distribution $p(\mathbf{Y}^{\text{rep}}|\theta)$.

4.2.2 Omnibus χ^2 test

The computed sample of the posterior predictive distribution may now be used to assess the fit of the model. We begin by evaluating the goodness of fit of the model with a Pearson χ^2 discrepancy measure. The resulting posterior predictive p-value equals 0.000, as $X^2(\mathbf{Y}, \theta^{(l)})$ exceeds $X^2(\mathbf{Y}^{\text{rep},l}, \theta^{(l)})$ for all the replicated datasets. Hence, the observed frequencies systematically deviate from the frequencies that could have been observed if the model were true and if the study were replicated with the same value of θ. Figure 1 displays a graphical representation of the pairs $(X^2(\mathbf{Y}^{\text{rep},l}, \theta^{(l)}), X^2(\mathbf{Y}, \theta^{(l)}))$ $(l = 1, \ldots, L)$. This picture is definitely more informative than merely reporting the posterior predictive p-value because the graph also shows the relative magnitudes of the realized and replicated discrepancies.

4.2.3 Comparison of the number of zeroes in the dataset

A crude way to investigate the misfit of the model is a visual comparison of the observed frequencies (observed variables aggregated across psychiatrists) with a few replicated datasets. This comparison shows that the observed data matrix contains far more zero cells than a typical replicated data matrix. Actually, 225 out of $30 \times 23 = 690$ cells of the observed aggregated data matrix equal zero, whereas the simulated reference distribution of the number of zeroes has mean 88 and a 95% posterior interval of $[71, 104]$. In other words, for about 33% (225 out of 690) of all patient-symptom pairs it holds that all psychiatrists agree that the patient does not have the symptom, whereas this would occur significantly less frequently in the replicated data if the model were actually true.

4.2.4 Comparison of the distributions of the θ_j parameters

For a PMD model with a disjunctive communality rule, a zero in the replicated data matrix implies that for each syndrome either the patient has a low probability of having that syndrome or that the symptom has a low probability of being characteristic of that syndrome. Figure 2 displays histograms of one representative draw of the posterior distribution of respectively all patient parameters and all symptom parameters. It must be noticed that for patients and symptoms, one draw consists of

respectively $3 \times 30 = 90$ and $3 \times 23 = 69$ parameters. The histograms in Figure 2 show a peak near zero, but clearly this peak is not high enough to replicate the observed number of zeroes. One may wonder to what extent this result is caused by the $\text{Beta}(\theta|2,2)$ prior distribution. In other words, it is possible that this prior distribution pulls parameter estimates towards $\frac{1}{2}$. A more appropriate prior distribution could imply that a histogram of one draw of the posterior distribution resembles the prior. Figure 2 also indicates that we should consider different prior distributions for patient and symptom parameters. Indeed, relatively more symptom parameters are close to zero.

4.2.5 Using the results of the model check to suggest a direction for model expansion

In general, a mixture of two beta distributed components seems to be a good alternative for the prior distribution. This mixture may be expressed as follows:

$$p(\theta_j) = \lambda \text{Beta}(\theta_j|\alpha,\beta) + (1-\lambda)\text{Beta}(\theta_j|\alpha',\beta'). \tag{5}$$

The parameters of the mixture may be guessed from a histogram of many draws of the posterior distribution. For patient and symptom parameters, $(\lambda, \alpha, \beta, \alpha'\beta')$ are respectively chosen to equal (.50, 1, 1, 1, 6) and (.50, 1, 1, 1, 16). We decided to use fixed values for the shape parameters of the mixture components and for the probability that a parameter θ_j belongs to a specific component, rather than estimating these values from the data. The latter strategy would perhaps be preferable in principle but would require additional data augmentation steps and potential new modeling difficulties (for example, instability in the estimated hyperparameters) that we do not want to worry about here. Our main concern at this point is modeling the distribution of the θ_j parameters in a reasonable way. (In fact, the procedure of creating Figure 2 and using it to specify hyperparameters in the mixture model could be thought of as a crude, one-step form of data-augmentation.)

4.3 Model expansion

The implementation of a mixture prior distribution instead of a $\text{Beta}(\theta|2,2)$ distribution requires only a modification of the posterior step of the data augmentation algorithm. In the posterior step one has to draw θ from the augmented posterior distribution $p(\theta|\mathbf{Y},\mathbf{Z})$. And since the individual components of θ are independent conditional on $\mathbf{Z}$, Meulders et al. (1997) show that each component θ_j may be sampled from:

$$p(\theta_j|\mathbf{Y},\mathbf{Z}) \propto p(\theta_j)\theta_j^{t_{1j}}(1-\theta_j)^{t_{0j}}, \tag{6}$$

t_{1j} and t_{0j} being functions of the latent data that summarize the information about θ_j from the preceding imputation step. That is to say, for a patient parameter θ_j that indicates the probability that patient o has latent syndrome b, the values t_{1j} and t_{0j} equal the number of times that patient o has

syndrome b and does not have syndrome b, respectively, summing over symptoms and psychiatrists. In the same way, for a symptom parameter θ_j that indicates the probability that symptom a applies to syndrome b, t_{1j} and t_{0j} equal the number of times that symptom a does apply (respectively does not apply) to syndrome b, summing over patients and psychiatrists. It is now easy to see that $p(\theta_j|\mathbf{Y},\mathbf{Z}) \propto \text{Beta}(\theta_j|t_{1j}+2, t_{0j}+2)$ if $p(\theta_j) \sim \text{Beta}(\theta_j|2,2)$.

If the mixture of expression (5) is the prior distribution, one must sample each parameter θ_j from

$$p(\theta_j|\mathbf{Y},\mathbf{Z}) \propto [\lambda \text{Beta}(\theta_j|\alpha,\beta) + (1-\lambda)\text{Beta}(\theta_j|\alpha',\beta')]\ \theta_j^{t_{1j}}(1-\theta_j)^{t_{0j}}. \tag{7}$$

Using the definition of the beta distribution, this expression may be simplified as follows:

$$p(\theta_j|\mathbf{Y},\mathbf{Z}) \propto \lambda k_j \text{Beta}(\theta_j|t_{1j}+\alpha, t_{0j}+\beta) + (1-\lambda)k_j' \text{Beta}(\theta_j|t_{1j}+\alpha', t_{0j}+\beta'), \tag{8}$$

with

$$k_j = \frac{\Gamma(\alpha+\beta)\Gamma(t_{1j}+\alpha)\Gamma(t_{0j}+\beta)}{\Gamma(\alpha)\Gamma(\beta)\Gamma(t_{1j}+t_{0j}+\alpha+\beta)}, \tag{9}$$

and k_j' being defined similarly, but with α' and β' replacing α and β. To ensure that expression (8) is a probability density function, we notice that the normalizing constant equals:

$$\int p(\theta|\mathbf{Y},\mathbf{Z})d\theta = \lambda k_j + (1-\lambda)k_j', \tag{10}$$

as both beta distributions in expression (8) are probability density functions. Hence normalization of expression (8) yields

$$p(\theta_j|\mathbf{Y},\mathbf{Z}) = \pi_j \text{Beta}(\theta_j|t_{1j}+\alpha, t_{0j}+\beta) + (1-\pi_j)\text{Beta}(\theta_j|t_{1j}+\alpha', t_{0j}+\beta') \tag{11}$$

with

$$\pi_j = \frac{\lambda k_j}{\lambda k_j + (1-\lambda)k_j'}. \tag{12}$$

To draw θ_j from the probability density function in expression (11) we first compute the probability π_j that parameter θ_j belongs to the first component of the mixture and then sample the parameter from this component with probability π_j and from the second component with probability $1-\pi_j$. The θ_j's are independent in the posterior distribution.

The computation of π_j is not straightforward, because it involves the evaluation of $\Gamma(\cdot)$, which may have a large value depending on its argument. A possible solution is to use a logarithmic transformation in order to avoid computational overflow. More specifically one could perform the following steps to compute each π_j:

1. Compute

$$\begin{aligned} c_j = \log(\lambda k_j) \quad &= \quad \log(\lambda) + \log(\Gamma(\alpha+\beta)) + \log(\Gamma(t_{1j}+\alpha)) + \log(\Gamma(t_{0j}+\beta)) \\ &\quad - \log(\Gamma(\alpha)) - \log(\Gamma(\beta)) - \log(\Gamma(t_{1j}+t_{0j}+\alpha+\beta)). \end{aligned}$$

2. Similarly, compute $c'_j = log((1-\lambda)k'_j)$ by replacing $\lambda, \alpha, \beta, k_j$ by $1-\lambda, \alpha', \beta', k'_j$ in the above formula.

3. Compute $r_j = \exp((c_j - c'_j)/2)$.

4. Compute $\pi_j = \frac{r_j}{r_j + 1/r_j}$.

The parameters of the expanded model for the psychiatric diagnosis data are estimated with a data augmentation algorithm. Four chains of 8000 iterations are simulated and the first half of the iterations are discarded to remove the influence of the starting point. The convergence diagnostic $\sqrt{\hat{R}}$ is smaller than 1.1 for all the parameters, so it may be concluded that further simulation would not much improve the precision of the simulated posterior distribution. Subsequently, every 8th iteration of each chain is saved, and these are combined to construct a sample of 2000 draws from the posterior distribution.

The change in the parameters due to the use of the mixture model may be investigated via a comparison of the posterior means of the parameters under the original and under the expanded model. Figure 3 displays a scatterplot of the posterior means of patient and symptom parameters for the two models. In general, parameters that are already near zero under the original model, are closer to zero under the model with the mixture prior distribution. On the other hand, patient parameters with a rather high value ($\geq .70$) and symptom parameters with a moderate value ($\geq .40$ and $\leq .70$) slightly increase. Although the estimated parameters are quite similar under the two models, the differences are important, especially for the parameters near zero, which strongly affect the probability that certain symptoms are not attributed at all to some patients.

The obtained posterior sample may further be used to check the fit of the expanded model in various aspects such as the overall goodness of fit of the model or the number of zero cells in the data matrix. First, we evaluate the overall goodness of fit of the expanded model using a Pearson χ^2 discrepancy measure. The result shows that the expanded model still has to be rejected: again $X^2(\mathbf{Y}, \theta^{(l)})$ exceeds $X^2(\mathbf{Y}^{\text{rep},l}, \theta^{(l)})$ for each replicated dataset, which suggests that further model improvement is possible and desirable.

Second, we compute the distribution of the frequency of zero cells in the replicated aggregated data. The mean of this distribution equals 117, and the 95% posterior interval is $[99, 135]$. Therefore, the number of zero cells in the observed data (225) is still implausible for data that were generated under the model and with the same value of θ. Yet, the expanded model predicts substantially more zero cells than the original model (the posterior interval of the number of zeroes under the original model is $[71, 104]$), hence the model expansion appears to be a step in the right direction. Furthermore the prior distribution of the expanded model is more realistic as it better resembles a

histogram of one draw of the posterior distribution (see Figure 4).

In addition to improving fit, the expanded model has the interesting psychological interpretation that some patient-syndrome pairs and some symptom-syndrome pairs are highly unlikely (these pairs that correspond to the mixture component located near zero) whereas the other patient-syndrome pairs and other symptom-syndrome pairs are more likely, with probabilities that vary approximately uniformly between 0 and 1. Future modeling can build on this perspective, which links probabilistic and deterministic matrix decomposition models.

5 Conclusion

The above analysis illustrates that the Bayesian framework offers a powerful approach with respect to model checking. It allows one to focus on very general model aspects, such as omnibus goodness-of-fit measures like χ^2 discrepancies, or one may evaluate specific aspects such as the number of zero cells in the observed data matrix. The above analysis shows the flexibility of model expansion within a Bayesian framework. Furthermore, the model expansion that was illustrated appears to decrease the discrepancy between observed and replicated data. In particular, the number of cells with zero entries in the observed data matrix is better fit under the expanded model (although there is still room for much improvement).

The model check that was particularly useful in this example—comparing the ensembles of the estimated patient and symptom parameters θ_j to their assumed prior distributions—was only feasible because of the internal replication in the model. After all, it would be hard for any single θ_j parameter to be inconsistent with an assumed $\text{Beta}(\theta|2,2)$ prior distribution. A set of 90 or 69 such parameters, however, can easily be compared with the model, as in Figure 2. This illustrates the general point that model checks—like models themselves—are most powerful when they incorporate structure in the underlying problem being studied. In this case, we refer to the structure inherent in having 90 replications of the patient/syndrome interactions and 69 replications of the symptom/syndrome interactions. Model checks that did not take advantage of this replication (for example, by testing the fit of the model on each of the θ_j parameters individually) would not be able to reveal the global fitting problem displayed in Figure 2.

On the whole, one may stress the flexibility of model checking in a Bayesian framework. Computation of posterior predictive checks is straightforward for any quantity of interest once a sample of the posterior distribution is available. In general this quantity may be a function of the data only or a function of both data and parameters. An important consequence of this flexibility is that it leads to a better understanding of the various model aspects and thus to a better comprehension of the data at hand.

Other approaches to Bayesian model building are also possible (see, e.g., Raftery, 1996, for a paradigm based on discrete model averaging). The purpose of this paper is not to claim that the posterior predictive approach is best, but rather to illustrate its use in a particular example of interest, in which model checking and expansion have been important tools in moving us toward models that fit the data better and that make substantive sense.

References

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

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

Gelman, A., and Meng, X. L. (1996). Model checking and model improvement. In *Practical Markov Chain Monte Carlo*, ed. W. Gilks, S. Richardson, and D. Spiegelhalter, 189–201. London: Chapman and Hall.

Gelman, A., Meng, X. M., and Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica* **4,** 733–807.

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

Gelman, A., Van Mechelen, I., Heitjan, D., and Meulders, M. (1997). *Bayesian model checking for missing and latent data problems using posterior predictive checks.* Unpublished manuscript, Department of Statistics, Columbia University, U.S.A.

Maris, E., De Boeck, P., and Van Mechelen, I. (1996). Probability matrix decomposition models. *Psychometrika* **61,** 7–29.

Meulders, M., De Boeck, P., Van Mechelen, I., Gelman, A., and Maris, E. (1997). *Bayesian inference with probability matrix decomposition models.* Unpublished manuscript, Department of Psychology, University of Leuven, Belgium.

Raftery, A. E. (1996). Hypothesis testing and model selection via posterior simulation. In *Practical Markov Chain Monte Carlo*, ed. W. Gilks, S. Richardson, and D. Spiegelhalter, 163–187. New York: Chapman & Hall.

Rubin, D. B. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *Annals of Statistics* **12,** 1151–1172.

Tanner, M. A., and Wong, W. H. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association* **82,** 528–540.

Van Mechelen, I., and De Boeck, P. (1990). Projection of a binary criterion into a model of hierarchical classes. *Psychometrika* **55,** 677–694.

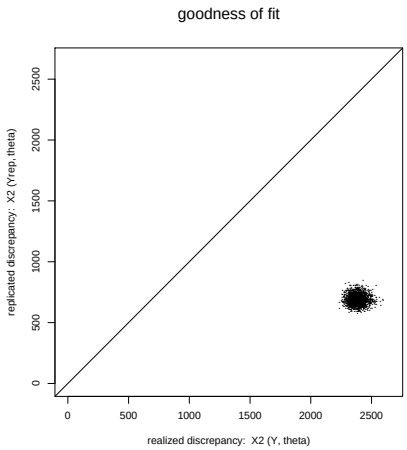

Figure 1: Scatterplot of replicated χ^2 discrepancy, $X^2(\mathbf{Y}^{\text{rep},l}, \theta^{(l)})$ vs. realized discrepancy, $X^2(\mathbf{Y}, \theta^{(l)})$, for 2000 random draws of $(\theta^{(l)}, \mathbf{Y}^{\text{rep},l})$ from the posterior distribution of the three-bundle PMD model fit to the psychiatric diagnosis data. The diagonal line corresponds to equality of the discrepancies. The realized discrepancies are consistently much larger, indicating that the discrepancy between data and model is much greater than would be predicted under the model.

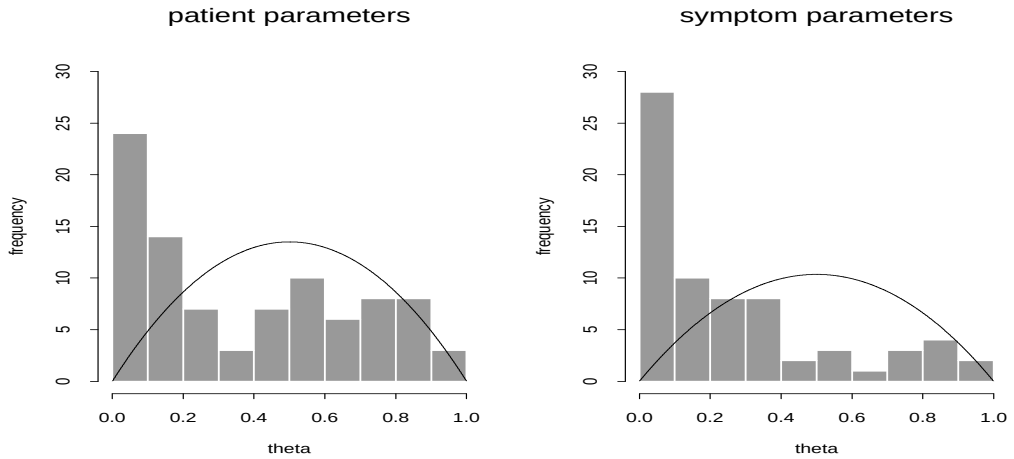

Figure 2: Histograms of (a) 90 patient parameters θ_j and (b) 69 symptom parameters θ_j, as estimated from the three-bundle PMD model fit to the psychiatric diagnosis data. These histograms of posterior estimates contradict the assumed Beta$(\theta|2,2)$ prior densities (plotted on top of the histograms) for each set of θ_j's, and motivated us to switch to mixture prior distributions. This implicit comparison to the values of θ_j under the prior distribution can be viewed as a posterior predictive check in which the replicated data include 30 new patients and 23 new symptoms, and thus new values for the θ_j's.

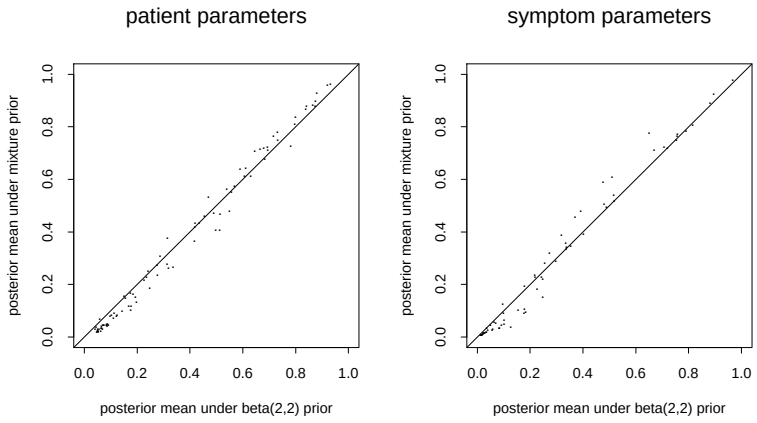

Figure 3: Estimates of parameters θ_j from the expanded PMD model (with mixture prior distributions on the θ_j's) compared to the original model (with Beta$(\theta|2,2)$ prior distributions). The diagonal lines correspond to equality of the estimates. The main change in fitting the mixture model is, for the parameters estimated near zero, to pull their estimates even closer to zero.

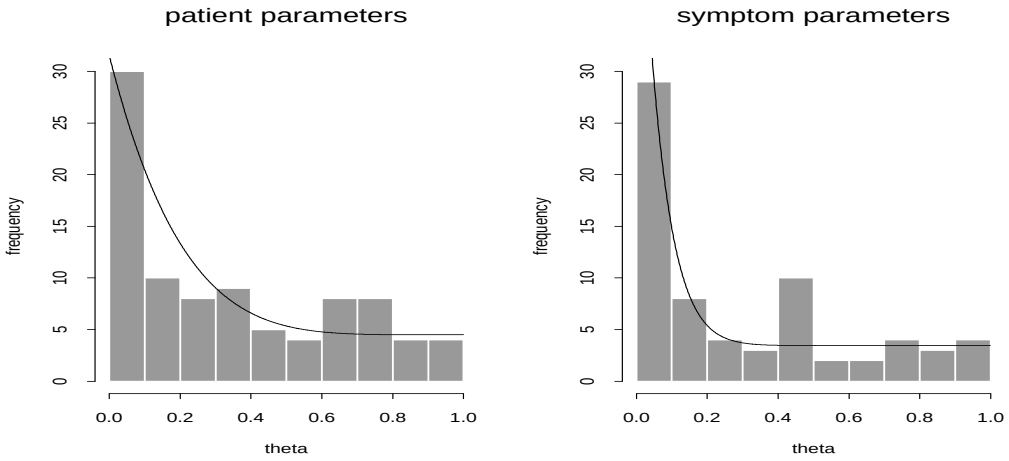

Figure 4: Histograms of (a) 90 patient parameters θ_j and (b) 69 symptom parameters θ_j, as estimated from the expanded model fit to the psychiatric diagnosis data. The mixture prior densities of the θ_j's (plotted on top of the histograms) are not perfect, but they approximate the corresponding histograms much better than the Beta$(\theta|2,2)$ densities in Figure 2.