Bayesian Analysis (2009) 4, Number 2, pp. 367–392

Improved Criteria for Clustering Based on the Posterior Similarity Matrix

Arno Fritsch* and Katja Ickstadt†

Abstract. In this paper we address the problem of obtaining a single clustering estimate $\widehat{c}$ based on an MCMC sample of clusterings $c^{(1)}, c^{(2)} \ldots, c^{(M)}$ from the posterior distribution of a Bayesian cluster model. Methods to derive $\widehat{c}$ when the number of groups K varies between the clusterings are reviewed and discussed. These include the maximum a posteriori (MAP) estimate and methods based on the posterior similarity matrix, a matrix containing the posterior probabilities that the observations i and j are in the same cluster. The posterior similarity matrix is related to a commonly used loss function by Binder (1978). Minimization of the loss is shown to be equivalent to maximizing the Rand index between estimated and true clustering. We propose new criteria for estimating a clustering, which are based on the posterior expected adjusted Rand index. The criteria are shown to possess a shrinkage property and outperform Binder's loss in a simulation study and in an application to gene expression data. They also perform favorably compared to other clustering procedures.

Keywords: adjusted Rand index, cluster analysis; Dirichlet process mixture model; Markov chain Monte Carlo

1 Introduction

Given a sample of clusterings $c^{(1)}, c^{(2)} \ldots, c^{(M)}$ from the posterior distribution $p(c|y)$ of a Bayesian cluster model, where the sample is the output of a Markov Chain Monte Carlo (MCMC) algorithm, it is often desirable to summarize the sample with a single clustering estimate $\widehat{c}$. Here a clustering is defined as a vector of allocation variables $c = (c_1, \ldots, c_n)'$. The estimation task is complicated by the fact that the intuitive estimator for the posterior probability that observation i belongs to cluster k,

$$P(c_i = k|y) \approx \frac{1}{M} \sum_{m=1}^{M} I_{\{c_i^{(m)}=k\}} \quad , \tag{1}$$

with $i = 1, \ldots, n$ and $k = 1, \ldots, K$, does not lead to sensible results when applied to the unprocessed sample of clusterings. So one cannot simply assign each observation to the cluster k that maximizes (1), as is done in discriminant analysis. This is due to the problem of "label switching", which means that during the MCMC run the labels associated with the clusters change. Methods that deal with label switching, e.g. the

*Department of Statistics, Technische Universität Dortmund, Germany, mailto:arno.fritsch@tu-dortmund.de

†Department of Statistics, Technische Universität Dortmund, Germany, mailto:ickstadt@statistik.tu-dortmund.de

 DOI:10.1214/09-BA414

relabeling algorithm of Stephens (2000), a post-processing algorithm that intends to find a coherent assignment of cluster labels, require K to be fixed. If K varies between the clusterings a possible solution is to choose a clustering based on the posterior similarity matrix $P(c_i = c_j|y)$, an $n \times n$ matrix that contains the pairwise probabilities that two observations belong to the same cluster. This approach is taken, for example, in the Bayesian cluster models for microarray data by Dahl (2006) and Medvedovic, Yeung, and Bumgarner (2004).

In Section 2 a short overview of Bayesian model-based cluster analysis in general is given and in Section 3 possible ways of deriving $\widehat{c}$ with a varying number of clusters K are considered. This includes the MAP and methods based on the posterior similarity matrix. The approach by Medvedovic et al. and the loss function proposed by Binder (1978) are considered. We will point out a relation of the loss function to the Rand index for comparing clusterings and discuss possible problems with its use. New criteria for estimating a clustering from an MCMC sample are introduced, which are based on maximizing the posterior expected adjusted Rand index, thus improving upon Binder's loss. Techniques for optimizing the criteria will also be considered. The new criteria are shown to perform favorably in choosing a clustering estimate close to the true clustering in a simulation study and in an application to gene expression data in Section 4.

2 Background: Bayesian model-based cluster analysis

Cluster analysis, the attempt to group previously unstructured data so that the observations in a group are more similar to each other than to observations from other groups, has been a valuable tool in statistics for a long time. Classical methods such as hierarchical clustering or K-means remain popular, although it is difficult to assess the statistical properties of the solutions provided by these methods. It is, for example, hard to quantify the uncertainty of the allocation of an observation to a specific group or the probability that two observations belong to the same group. Model-based cluster methods offer an alternative. It is assumed that the observations y_i in each cluster are generated by a distribution $p(y_i|\theta_k)$ with a group specific parameter vector θ_k. Problems in model-based cluster analysis concern the estimation of the θ_k's, of the cluster weights $\phi_k = P(c_i = k)$, of the number of clusters K and of the vector of allocations c. Estimation can be done using the EM-algorithm to compute maximum likelihood (ML) estimators (Dempster et al. 1977), or alternatively Bayesian methods can be employed by assigning a prior distribution to all parameters. Inference on the posterior distribution of parameters can again be based on the EM-algorithm to derive maximum a posteriori (MAP) estimates. This approach is taken, for example, in the MCLUST procedure of Fraley and Raftery (2002), who also provide a review of model-based cluster methods in general. Another possibility is to use Markov chain Monte Carlo methods to obtain a sample of the posterior distribution. This has the advantage that, instead of using a single estimate of the parameters, their uncertainty can be taken into account. MCMC approaches also allow to fit rather complex models. It is possible to estimate the number of groups K at the same time as the other parameters, either by using the reversible jump algorithm in a finite mixture model as shown by Richardson and Green

(1997) or by letting the θ_k be (unique) realizations of a Dirichlet process prior (Ferguson 1973) with mass parameter α and base measure G_0, which results in an infinite mixture model. Although Dirichlet process (DP) mixture models have been developed as flexible nonparametric models for random distributions (see, e.g., Dunson (2008) for a variety of applications), they can also be used for clustering, similarly to finite mixture models. For the Dirichlet process each observation y_i has an associated parameter (vector) θ_i which follow the Polya urn scheme

$$\theta_{i+1}|\theta_1, \ldots, \theta_i \sim \frac{1}{\alpha + i} \sum_{j=1}^{i} \delta(\theta_j) + \frac{\alpha}{\alpha + i} G_0 \ , \tag{2}$$

with $\delta(\theta_j)$ indicating the point measure on θ_j (Blackwell and MacQueen 1973). θ_{i+1} is thus either equal to one of the previous θ_i's or is drawn from G_0. The positive probability of sharing the parameter value with previous observations induces a clustering. The Dirichlet process prior can be extended in many ways, for example with the two-parameter Poisson-Dirichlet process (Pitman and Yor 1997), general stick-breaking priors (Ishwaran and James 2001) or generalized gamma process priors (Lijoi et al. 2007). A recent review on prior models for Bayesian cluster analysis is provided by Lau and Green (2007).

Other implementations of complex Bayesian cluster models allow for outlier detection (Quintana and Iglesias 2003), simultaneous clustering and variable selection (Kim et al. 2006; Tadesse et al. 2005), improving the power of multiple testing by clustering correlated observations (Dahl and Newton 2007), or clustering transcription factor binding motifs of varying width (Jensen and Liu 2008).

In the MCMC run label switching occurs if the sampler actually converges to the posterior distribution. If the cluster labels are assumed to be exchangeable, a permutation of cluster labels does not change the likelihood of a clustering. For a model that has priors $p(\theta_k)$ and $p(\phi_k)$ that are equal for all k, each of the $K!$ permutations of labels is associated with a posterior mode of equal height. All of these modes will be visited if the MCMC sampler runs long enough.

It is easy to see that label switching does not affect the posterior similarity matrix with elements $P(c_i = c_j|y)$ which can therefore be estimated from the MCMC sample by

$$\pi_{ij} = P(c_i = c_j|y) \approx \frac{1}{M} \sum_{m=1}^{M} I_{\{c_i^{(m)} = c_j^{(m)}\}} \ . \tag{3}$$

In (3) the number of groups K does not have to be fixed. Because of the symmetry of the posterior similarity matrix it suffices to regard entries with $i < j$.

3 Clustering with a varying number of groups

Some current approaches for clustering with a varying number of components are reviewed in Subsection 3.1 while new criteria are introduced in Subsection 3.2. The new criteria and some of the current approaches require maximization over a set of candidate clusterings c^*, which will be discussed in Subsection 3.3.

3.1 Current approaches

Maximum a posteriori (MAP) clustering

As mentioned above the posterior density is invariant to a permutation of the labels. A simple solution to the label switching problem with a varying number of groups is thus to take as $\widehat{c}$ the clustering c^* that maximizes the posterior density, as it is not important from which of the equivalent modes this clustering stems. The MAP estimate is known to minimize the posterior expectation of the 0-1 loss function, where no loss is made if $\widehat{c}$ is exactly equal to c (up to a permutation of the labels) and the loss is 1 in any other case.

Ad hoc approach of Medvedovic et al. (2004)

Medvedovic et al. (2004) employ classical agglomerative hierarchical clustering, as, for example, described in Kaufman and Rousseuw (1990), to obtain an estimate $\widehat{c}$, using $1-\pi_{ij}$, the posterior probability that the observations i and j are not clustered together, as the distance between the observations i and j. If K is known they use average linkage and cut the dendrogram at K groups. For unknown K they use complete linkage and cut the dendrogram at a distance of $1-\varepsilon$, for small, positive ε. For two distinct clusters C_k and $C_{k'}$ then there is at least one pair of observations (i,j) with $c_i = k$ and $c_j = k'$, such that $\pi_{ij} < \varepsilon$. In this paper we will employ $\varepsilon = 0.01$.

Although this approach has been criticized as being rather *ad hoc*, it has to be acknowledged that $1-\pi_{ij}$ is a sensible distance measure. We found that $1-\pi_{ij}$ is a topological *pseudometric* for the space of observations, as it fulfills the conditions

$$\begin{aligned} 1-\pi_{ii} &= 0 && (4)\\ 1-\pi_{ij} &= 1-\pi_{ji} && (5)\\ \text{and}\quad (1-\pi_{ij}) &\leq (1-\pi_{ik})+(1-\pi_{jk})\ . && (6) \end{aligned}$$

While (4) and (5) are straightforward to see, a proof that the triangle inequality (6) is valid can be found in Appendix 6. $1-\pi_{ij}$ is not a *metric* since $1-\pi_{ij}=0$ does not imply that the observations i and j are equal.

Binder's loss function

Binder (1978) was the first to consider loss functions based on pairwise occurences of observations, i.e.

$$L(c^*,c) = \sum_{i<j} \ell_1 \cdot I_{\{c_i^* \neq c_j^*\}} I_{\{c_i = c_j\}} + \ell_2 \cdot I_{\{c_i^* = c_j^*\}} I_{\{c_i \neq c_j\}}\ , \qquad (7)$$

with positive constants ℓ_1 and ℓ_2. c^* is a proposed clustering estimate and the matrix containing $I_{\{c_i^*=c_j^*\}}$ is a known 0-1 matrix, which will be referred to as estimated similarity matrix. The unknown true clustering c has a similarity matrix containing

$I_{\{c_i=c_j\}}$. Since $E(I_{\{c_i=c_j\}}|y) = \pi_{ij}$, the posterior similarity matrix can be seen as the similarity matrix of the posterior expected clustering $E(c|y)$.

The quotient ℓ_1/ℓ_2 determines the preferred kind of clustering. If $\ell_1 \gg \ell_2$ the loss is minimized if all observations are in one cluster, whereas for $\ell_2 \gg \ell_1$ the minimum is attained if all observations are in their own singleton cluster. When there is no particular preference with regard to the two types of errors a pragmatic solution is to set $\ell_1 = \ell_2 = 1$ and thus to penalize the two types of errors equally. This approach is taken by Hurn et al. (2003) in the context of a switching regression model. In the following we will refer to the $\ell_1 = \ell_2 = 1$ case of (7) as Binder's loss.

The posterior expectation of this loss can be written as

$$E(L(c^*, c)|y) = \sum_{i<j} |I_{\{c_i^*=c_j^*\}} - \pi_{ij}| \ , \tag{8}$$

i.e. the sum of absolute deviations of the estimated similarity matrix to the posterior similarity matrix. The estimated clustering $\widehat{c}$ can be taken as the clustering c^* minimizing (8). Because of the linearity of the loss function the same expression is obtained if the loss function is computed between estimated and posterior expected clustering, so that

$$E(L(c^*, c)|y) = L(c^*, E(c|y)) \ . \tag{9}$$

With Binder's loss function a loss of 1 is made whenever a pair of observations is treated differently in the estimated clustering c^* than in the true c. The loss is thus the sum of disagreements in the treatment of pairs of observations between the estimated and true clustering. If the number of disagreements and agreements between two clusterings is denoted by D and A, then $D + A = \binom{n}{2}$. Rand (1971) used $A/\binom{n}{2}$ as a general measure for the similarity of two clusterings, a number that is known as the Rand index. The Rand index of estimated and true clustering $R(c^*, c)$ is thus given by

$$R(c^*, c) = 1 - \frac{L(c^*, c)}{\binom{n}{2}} \ .$$

The clustering $\widehat{c}$ that minimizes the posterior expectation of Binder's loss in equation (8) also maximizes the posterior expected Rand index with the true clustering and, considering equation (9), the Rand index of estimated and posterior expected clustering.

Dahl's criterion

Dahl (2006) used

$$\sum_{i<j} (I_{\{c_i^*=c_j^*\}} - \pi_{ij})^2 \ , \tag{10}$$

as a criterion to be minimized to obtain an estimate $\widehat{c}$. Minimization of (8) and (10) is equivalent, which can be seen by writing

$$\sum_{i<j} |I_{\{c_i^*=c_j^*\}} - \pi_{ij}| = \sum_{i<j} (\pi_{ij} - 2 \cdot I_{\{c_i^*=c_j^*\}}\pi_{ij} + I_{\{c_i^*=c_j^*\}}) \ ,$$

and

$$\sum_{i<j} (I_{\{c_i^*=c_j^*\}} - \pi_{ij})^2 = \sum_{i<j} (\pi_{ij}^2 - 2 \cdot I_{\{c_i^*=c_j^*\}} \pi_{ij} + I_{\{c_i^*=c_j^*\}}) \ .$$

The difference between these sums is $\sum_{i<j} \pi_{ij}(1-\pi_{ij})$, which does not depend on the estimated clustering, so that minimization of Dahl's criterion is equivalent to the minimization of Binder's loss.

3.2 New criteria for clustering

Derivation and Motivation

Table 1: The contingency table of two clusterings.

		Clustering V			
	Class	v_1	$\cdots$	v_L	$\sum$
Clustering U	u_1	n_{11}	$\cdots$	n_{1L}	$n_{1.}$
	$\vdots$	$\vdots$		$\vdots$	$\vdots$
	u_K	n_{K1}	$\cdots$	n_{KL}	$n_{K.}$
	$\sum$	$n_{.1}$	$\cdots$	$n_{.L}$	n

Although the Rand index is not an unsensible measure to compare clusterings one of its drawbacks is that the number of expected chance agreements of the clusterings depends heavily on the number of groups in each clustering, their sizes, and the overall number of observations. To overcome this problem Hubert and Arabie (1985) considered the contingency table of the two clusterings shown in Table 1 and proposed an adjusted Rand index, where the index is corrected for its expected value under the assumption of random sampling of the n_{kl} from fixed marginal sizes $n_{k.}$ and $n_{.l}$, i.e. assuming a generalized hypergeometric distribution for the contingency table. The adjusted Rand has the usual form of an index corrected for chance:

$$\frac{\text{Index - Expected Index}}{\text{Maximum Index - Expected Index}} \ .$$

It has a maximum value of 1 and its value is 0 if the Rand index equals its expected value. Negative values are possible, but uninteresting as they indicate less agreement than expected by chance. Hubert and Arabie (1985) derive the following formula for the adjusted Rand index:

$$\frac{\sum_{k,l} \binom{n_{kl}}{2} - \sum_k \binom{n_{k.}}{2} \sum_l \binom{n_{.l}}{2} / \binom{n}{2}}{\frac{1}{2}[\sum_k \binom{n_{k.}}{2} + \sum_l \binom{n_{.l}}{2}] - \sum_k \binom{n_{k.}}{2} \sum_l \binom{n_{.l}}{2} / \binom{n}{2}} \ . \tag{11}$$

Another problem with the Rand index is its large variance which is considerably lowered by the adjusted Rand index, as found by Milligan and Cooper (1986). Among several

other indices they recommend the adjusted Rand for the comparison of clusterings. It is still one of the most popular measures used for this purpose. In a Bayesian context it is for example used by Dahl (2006) and Medvedovic et al. (2004) for the evaluation of their simulation studies.

If, like Dahl and Medvedovic et al., one deems the adjusted Rand index to be preferable as a measure of association to the Rand index it might be preferable to try to maximize the adjusted Rand index of estimated and true clustering $AR(c^*, c)$ instead of $R(c^*, c)$, as is done with the minimization of Binder's loss. We thus take $AR(c^*, c)$ as an utility function to be maximized. Suppose that the estimated and true clustering corresponds to clustering U and V in Table 1, respectively. In that case the equations

$$\begin{aligned} \textstyle\sum_k \binom{n_{k\cdot}}{2} &= \sum_{i<j} I_{\{c_i^*=c_j^*\}} \ , \\ \textstyle\sum_l \binom{n_{\cdot l}}{2} &= \sum_{i<j} I_{\{c_i=c_j\}} \quad \text{and} \\ \textstyle\sum_{k,l} \binom{n_{kl}}{2} &= \sum_{i<j} I_{\{c_i^*=c_j^*\}} I_{\{c_i=c_j\}} \ , \end{aligned}$$

hold. $AR(c^*, c)$ can then be written as

$$\frac{\sum_{i<j} I_{\{c_i^*=c_j^*\}} I_{\{c_i=c_j\}} - \sum_{i<j} I_{\{c_i^*=c_j^*\}} \sum_{i<j} I_{\{c_i=c_j\}} / \binom{n}{2}}{\frac{1}{2}\left[\sum_{i<j} I_{\{c_i^*=c_j^*\}} + \sum_{i<j} I_{\{c_i=c_j\}}\right] - \sum_{i<j} I_{\{c_i^*=c_j^*\}} \sum_{i<j} I_{\{c_i=c_j\}} / \binom{n}{2}} \ .$$

This expression depends of course on the unknown true clustering c. When taking the posterior expectation to obtain an expression that can be computed given an potential clustering estimate c^* and the MCMC sample $c^{(1)}, c^{(2)} \ldots, c^{(M)}$ we can utilize either side of equation (9) leading to either $E(AR(c^*, c)|y)$ or $AR(c^*, E(c|y))$. Unlike Binder's loss $AR(c^*, c)$ is not a linear function of the $I_{\{c_i=c_j\}}$ so these two expressions are related but not the same.

Maximizing $E(AR(c^*, c)|y)$ over c^* leads to a clustering that has maximum posterior expected adjusted Rand index with the true clustering. This can be approximated from the MCMC sample by

$$\frac{1}{M} \sum_{m=1}^{M} AR(c^*, c^{(m)}) \ , \tag{12}$$

with $AR(c^*, c^{(m)})$ being computed by equation (11).
The adjusted Rand index with the posterior expected clustering $AR(c^*, E(c|y))$ is given by the expression

$$\frac{\sum_{i<j} I_{\{c_i^*=c_j^*\}} \pi_{ij} - \sum_{i<j} I_{\{c_i^*=c_j^*\}} \sum_{i<j} \pi_{ij} / \binom{n}{2}}{\frac{1}{2}\left[\sum_{i<j} I_{\{c_i^*=c_j^*\}} + \sum_{i<j} \pi_{ij}\right] - \sum_{i<j} I_{\{c_i^*=c_j^*\}} \sum_{i<j} \pi_{ij} / \binom{n}{2}} \ , \tag{13}$$

where the π_{ij} are estimated from the MCMC sample via equation (3). These two slightly different criteria have been introduced since both have distinctive advantages.

Expression (13) requires the computation of the posterior similarity matrix, but can then be evaluated a lot faster than (12), which is advantageous if the criterion needs to be calculated for many different c^*. We also found (13) to be more amenable to a theoretical study. Expression (12) on the other hand does not require the computation of the posterior similarity matrix, which can be preferable for large n, where this matrix gets too large to be stored. And one can argue that maximizing $E(AR(c^*, c)|y)$ is a more standard approach than maximizing $AR(c^*, E(c|y))$. Practically, however, we found in our applications that maximization of either criterion leads to nearly identical results. For simplicity we will in the following refer to both criteria as PEAR for Posterior Expected Adjusted Rand.

A possible disadvantage of using PEAR as an optimality criterion is that the adjusted Rand index is 0 if one of the compared clusterings consists of only one cluster or of all singletons. As there will always be clusterings c^* leading to positive values of (12) or (13) these extreme clusterings will never be chosen as $\widehat{c}$.

A shrinkage property

It is instructive to consider the behavior of Binder's loss and PEAR, if there were no restrictions for the $I_{\{c_i^*=c_j^*\}}$, i.e. all $I_{\{c_i^*=c_j^*\}}$ could be set individually to 0 or 1 without regard of the other indicator functions. From equation (8) it is clear that for Binder's loss the optimal solution in that case is simply to set $I_{\{c_i^*=c_j^*\}} = 1$, if $\pi_{ij} \geq 0.5$. In the case of PEAR it can be seen that for the expression (13) if ℓ of the $I_{\{c_i^*=c_j^*\}}$ are 1 and the rest 0, the maximum is attained if the $I_{\{c_i^*=c_j^*\}} = 1$ correspond to the ℓ highest π_{ij}. Denoting with $\pi_{(i)}$ the ith largest π_{ij} and letting $\sum_{i<j} \pi_{ij} / \binom{n}{2} = \bar{\pi}_{..}$, the maximum of (13) is then given by the maximum of

$$PEAR^*(\ell) = \frac{\sum_{i=1}^{\ell} \pi_{(i)} - \ell\, \bar{\pi}_{..}}{\frac{1}{2}\ell\,(1 - 2\bar{\pi}_{..}) + \frac{1}{2}\binom{n}{2}\bar{\pi}_{..}} \tag{14}$$

for $\ell = 0, 1, \ldots, \binom{n}{2}$. It can be shown that at the value ℓ^* for which (14) is maximal there is a threshold t such that $\pi_{(\ell^*)} = \min\{\pi_{(i)} : \pi_{(i)} \geq t\}$ and that the following relation holds:

$$\begin{aligned} \bar{\pi}_{..} < t < 0.5 \quad & \text{if} \quad \bar{\pi}_{..} < 0.5 \\ t = 0.5 \quad & \text{if} \quad \bar{\pi}_{..} = 0.5 \\ \bar{\pi}_{..} > t > 0.5 \quad & \text{if} \quad \bar{\pi}_{..} > 0.5 \ . \end{aligned} \tag{15}$$

Compared to Binder's loss, where t is always 0.5, the threshold for setting $I_{\{c_i=c_j\}}$=1 is shrunk towards the mean of the π_{ij}. PEAR thus adjusts to the amount of clustering found in the data. If overall only few π_{ij} are large, the conditions on putting two observations in one cluster are less strict, e.g. two observations i and j might be clustered together if π_{ij} is only 0.4. The opposite applies if overall many π_{ij} are large. A proof of (15) can be found in Appendix 7.

3.3 Optimization of criteria

Unlike the method of Medvedovic et al. (2004) the criteria MAP, the expectation of Binder's loss and PEAR have to be optimized over several c^* to obtain a clustering estimate $\widehat{c}$. As it is not feasible to compute the criteria for all possible clusterings of the n observations, a small set of candidate clusterings c^* that will lead to a close to optimal solution is needed. A simple solution taken for example by Dahl (2006) is to take the MCMC sample $c^{(1)}, c^{(2)} \ldots, c^{(M)}$ as this set.

Another possibility of a small set of clusterings with potentially good values that will be considered here is given by the clusterings obtained by the hierarchical clustering approach with distances $1 - \pi_{ij}$ of Medvedovic et al.. The criteria can be computed for the clusterings on every level of the hierarchy and $\widehat{c}$ taken to be the optimal among these. For the special case of the expectation of Binder's loss and the clusterings c^* from the hierarchical clustering with average linkage it is not difficult to show that the optimal $\widehat{c}$ among the c^*'s is always obtained by cutting the dendrogram at 0.5, so that in this case it is not necessary to compute the criterion for all c^*.

More sophisticated optimization methods could of course be applied. For the expectation of Binder's loss Lau and Green (2007) realized that it suffices to minimize the linear functional

$$\sum_{i<j} I_{\{c_i^*=c_j^*\}}(1 - 2\pi_{ij}) \ .$$

Using the constraints that for all triples (i, j, k), if $I_{\{c_i^*=c_j^*\}} = 1$ then $I_{\{c_i^*=c_k^*\}} = I_{\{c_j^*=c_k^*\}}$, they formulate the minimization of the expected loss as a binary integer programming problem, which can be solved exactly. Practically this is only feasible for small n as the number of variables grows quadratically and the number of constraints cubically in n. Even the algorithm to obtain an approximate solution given by Lau and Green involves already solving n binary integer programming problems with $O(n)$ variables and $O(n^2)$ constraints in each iteration.

Other optimization methods that could be used for all criteria include for example greedy search or simulated annealing algorithms.

4 Applications

We test the performance of the discussed approaches to find an estimated clustering $\widehat{c}$ that is close to the true clustering by fitting a simple Dirichlet process mixture model with normal components to simulated data and gene expression data.

The minimization of the posterior expectation of Binder's loss (Eqn. (8)) will be referred to as MinBinder, the maximization of PEAR (Eqn. (13)) as MPEAR and the complete linkage method for unknown K of Medvedovic et al. as MedvComp. Using Equation (12) for PEAR leads to nearly identical results. As a benchmark we also show results obtained with the MCLUST procedure mentioned in Chapter 2 which is fitted as implemented in the `R` package `mclust` using the default settings. Computation times

refer to a desktop computer with 3 GHz and 2 GB RAM.

4.1 Prior setting of Dirichlet process mixture model

Here we briefly discuss the prior for the important hyperparameter α of the DP mixture model. Further details on the model, the prior settings and implementation can be found in Appendix 9.

From (2) it can be seen that α controls the prior clustering behavior and should therefore be assigned a prior distribution. A variety of priors have been proposed for α, including Gamma (Escobar and West 1995), inverse Gamma (Medvedovic et al. 2004) and inverse Beta distributions (Griffin and Steel 2006). The choice of the prior is usually guided by considering the induced prior on the number of components K, which depends only on α and n, see Antoniak (1974) for details. Since our inference is based on $\pi_{ij} = P(c_i = c_j|y)$ we rather consider the induced prior on $P(c_i = c_j)$, given by $P(c_i = c_j) = 1/(1+\alpha)$. We will use the $Ga(\delta_1, \delta_2)$ prior of Escobar and West (1995), if values of δ_1 and δ_2 that induce a suitable prior on $1/(1+\alpha)$ can be found. Alternatively a $Beta(v_1, v_2)$ distribution can be assigned directly to $1/(1+\alpha)$, leading to a Beta distribution of the second kind for α (Johnson et al. 1995, p.248).

4.2 Simulation study

Setup

The simulated data are 3-dimensional with 8 clusters, where the cluster means are given by the 8 possible values of $(\pm\delta, \pm\delta, \pm\delta)^T$. Observations are obtained by adding independent standard normal errors to the cluster means. As δ determines how well the clusters are separated, we use values of $\delta = 0.5, 1.0, 1.5, 2.0$ to get data sets that range from ones with largely overlapping to ones with fairly well separated clusters. One scenario with equal cluster sizes is simulated, where each cluster contains 50 observations and one with unequal sizes where half of the clusters contain 20 and the other half 80 observations. For each combination of δ and cluster size 10 data sets are generated. To investigate how the clustering methods perform in extreme cases, data sets are also generated for $\delta = 0$, so that all observations come from the same normal component. A scenario were each observation comes from its own component is simulated by setting $\delta = 2$ and extending the dimensionality of the data to 9, giving 512 observations with distinct cluster means.

The model described in Appendix 9 is fitted to the data sets with the hyperparameters a, b and v fixed at 1. A $Ga(4, 2)$ prior is used for α. If there are cluster structures in the data most pairs of observations will not share a cluster and the $Ga(4, 2)$ prior assigns most of the prior mass for $1/(1+\alpha)$ to values between 0.1 and 0.6. As judged by trace plots of the number of clusters burn-in seems to be rather quick and after discarding the first 1000 iterations the algorithm is run for 50,000 iterations of which every 100th is used for the estimation of the posterior similarity matrix. It takes about

8 minutes to run the model for one data set.

Results

Table 2: Mean minimal value of posterior expectation of Binder's loss found with different optimization methods for the equal cluster size data.

	All observations (n=400)			Half of observations (n=200)			
	Draws	Comp	Avg	Draws	Comp	Avg	Lau&Green
δ=0.5	21527	21561	21544	5368	5378	5366	5363
δ=1.0	24477	24457	24315	6112	6095	6075	6062
δ=1.5	10087	8842	8639	2487	2175	2143	2141
δ=2.0	4465	3843	3739	1016	913	895	894

Draws refers to minimization over the MCMC sample, Comp and Avg to the minimization over all levels of the hierarchical clustering with complete/average linkage and Lau&Green to the optimization method proposed by Lau and Green (2007), which could not be applied to all observations.

First we take a look at the performance of the different optimization procedures mentioned in Section 3.3. Table 2 shows the average minimal value found for the posterior expectation of Binder's loss with different approaches. Using the `R` package `lpSolve` as done by Lau and Green (2007) to implement their algorithm it was not possible to apply the algorithm to all 400 observations, as the optimization problems required at each iteration got too large to be handled by the software. We therefore tested the approach by only considering a part of the posterior similarity matrix corresponding to the first half of observations from each true cluster. For these 200 observations it took the algorithm of Lau and Green about 30-40 minutes to finish for each data set. The algorithm did succeed in finding clusterings with the lowest value, but is closely followed by the hierarchical clustering with average linkage, which took less than a second to compute. Note also that minimization of the criterion over the drawn clusterings does lead to minimal values that are much higher than for the other methods. Similar results have also been found for the unequal cluster size data and for the optimization of the criteria MAP and MPEAR (results not shown). In the following we take the best clustering over all optimization methods as $\widehat{c}$ for each criterion.

To evaluate how close the $\widehat{c}$ of the different criteria are to the true clustering c adjusted Rand indices of estimated and true clustering are computed. Figure 1 shows these for the data with equal and unequal cluster sizes. In the case of $\delta = 2$, i.e. for well separated clusters, all methods perform comparably. When the clusters are more overlapping MPEAR can be seen to give estimates closer to the true clustering than the other approaches. It gives notably better results than MinBinder, which it is meant to improve. This might be the case because the mentioned shrinkage effect comes into play. It is probably most beneficial if there are many π_{ij}'s close to 0.5, which is not the case for $\delta = 2$. For $\delta = 0.5$ most of the MCMC clusterings consist of only one cluster

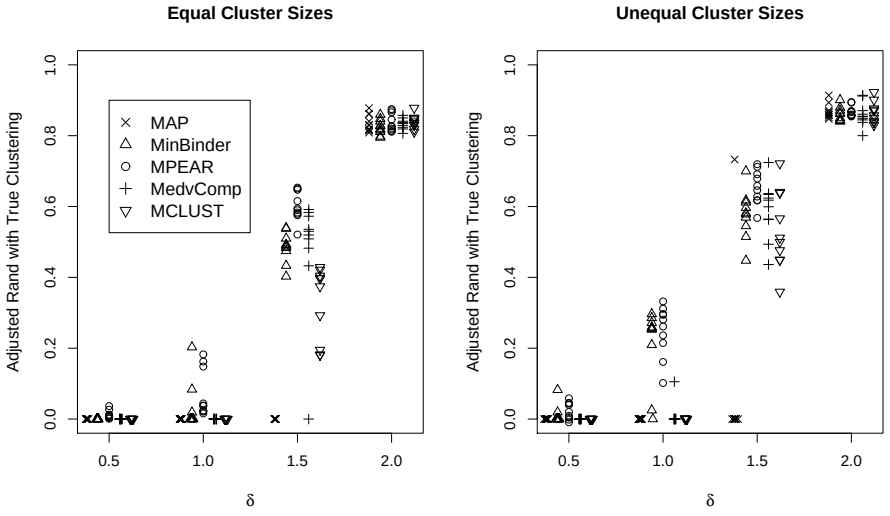

Figure 1: Adjusted Rand Index with true clustering for clusterings estimated with five different methods. Left: Data with clusters of equal sizes. Right: Data with clusters of unequal sizes.

so that none of the estimation criteria can find meaningful cluster structure. The other methods also put all observations into one cluster for higher values of δ. MedvComp and MCLUST do this for δ equal to 1.0, MAP already for $\delta = 1.5$. In the case of MAP this can be explained by the high prior probability that the Dirichlet process places on allocating all observations to one component as can be inferred from formula (22). The likelihood has to be quite high to counteract this.

For the data sets with only one true cluster all criteria except MPEAR correctly assign all observations to one cluster. As noted PEAR will never be maximal at one cluster and thus places some observations into other clusters, with most observations still being in one large cluster. For the data where all observations come from distinct components the MCMC output of the DP mixture model consists again mostly of clusterings with only one cluster and the criteria based on its output accordingly put all observations into one cluster. While this might look like a flaw of the DP mixture model, it can be argued that all observations in one cluster or all in singleton clusterings are just different ways of expressing that no cluster structures have been found in the data.

The mean number of clusters found by the different methods are shown in Table 3. It can be seen that for the higher values of δ the estimate obtained with MinBinder has a lot more clusters than the 8 truly present. This is also the case for MPEAR, but not as extreme. When the other methods do not put all observations in one cluster their

Table 3: Mean number of clusters found in the simulation study.

	Equal Cluster Sizes				
	MAP	MinBinder	MPEAR	MedvComp	MCLUST
δ=0.5	1.0	1.4	6.5	1.0	1.0
δ=1.0	1.0	5.2	3.4	1.0	1.0
δ=1.5	1.0	89.1	12.7	6.6	3.4
δ=2.0	8.1	23.7	12.2	8.0	8.0
	Unequal Cluster Sizes				
	MAP	MinBinder	MPEAR	MedvComp	MCLUST
δ=0.5	1.0	2.5	5.9	1.0	1.0
δ=1.0	1.0	35.1	4.8	1.1	1.1
δ=1.5	1.7	72.2	14.3	5.2	4.6
δ=2.0	8.1	23.6	13.4	7.5	8.2
	Extreme Case Data				
	MAP	MinBinder	MPEAR	MedvComp	MCLUST
One Cluster	1.0	1.0	7.8	1.0	1.0
Singletons	1.0	1.0	7.7	1.0	4.3

estimates have less than the true number of clusters for $\delta = 1.5$ and are on average approximately correct for $\delta = 2$.

The tendency of MinBinder and MPEAR to overestimate the number of clusters can be explained by the fact that many observations are put in singleton or in very small clusters. Table 4 shows the mean number of singletons and larger clusters for the two criteria. It can be seen that on average both methods are close to the correct number of (large) clusters for δ equal to 1.5 and 2. MinBinder produces many singletons. An example of this is given in Figure 2 in Appendix 8. It shows the π_{ij} for an observation i that is put in a singleton cluster by both MinBinder and MPEAR and for an observation that is clustered by itself only by MinBinder. In the latter case it can be seen that the shrunken threshold t on π_{ij} of MPEAR leads to the observation being assigned to the correct cluster.

Table 4: Mean number of singletons and large clusters (more than 10 observations) for equal cluster size data.

	MPEAR		MinBinder	
	Singletons	Large Clusters	Singletons	Large Clusters
δ=0.5	2.5	2.2	0.1	1.0
δ=1.0	0.2	3.0	1.6	2.2
δ=1.5	2.0	8.1	66.5	8.9
δ=2.0	3.8	8.0	12.1	8.0

A sensitivity analysis of the simulation study concerning the number of MCMC iterations, different prior settings of the DP mixture model and different performance measures is given in Appendix 10.

4.3 Yeast galactose data

Table 5: Adjusted Rand index with true clustering for yeast galactose data. (Results for replicates are averages over the four replications.)

	MAP	MinBinder	MPEAR	MedvComp	MCLUST
All data	0.952	0.952	0.952	0.952	0.937
Replicates	0.903	0.906	0.908	0.907	0.841

The different estimation approaches are compared on gene expression data from a study on the galactose pathway (Ideker et al. 2001). Microarrays were used to measure mRNA concentrations under 20 different conditions in growing yeast, with the experiment being replicated four times. We use the same subset of 205 genes already employed by Medvedovic et al. (2004) and Qin (2006). The subset reflects four functional categories of the Gene Ontology (Gene Ontology Consortium 2000), which will be assumed to represent the true clustering. This true clustering consists of two large groups containing 83 and 93 genes and two small ones with 14 and 15 genes. While Medvedovic et al. and Qin used the original data for clustering we employ the first two principal components instead. The first two PCs explain a large part of the variation of the data and show clear cluster structures when plotted. Overall we obtain better results than reported by Medvedovic et al. and Qin. Since principal components are uncorrelated and the variances are comparable in our case the simple DP mixture model seems to be appropriate and is fitted to the PCs derived from each single experimental replicate and from the average expression over the four replicates.

Table 5 gives the adjusted Rand indices with the true clustering for MCLUST and the DP mixture model. MPEAR gives the highest indices, while MCLUST performs a bit worse indicating that the DP mixture model is indeed appropriate for the data. Examining the produced clustering estimates $\widehat{c}$ it turns out that, besides wrong allocation of some single observations, one of the small groups is either joined with the other small group or with one of the large groups by all methods.

A sensitivity analysis similar to the one for the simulation study described in Appendix 10 is also conducted. Fitting the DP mixture model with a vague prior on α does not change the results of Table 5 much (data not shown), however with additional Gamma priors on the hyperparameters b and v different results are obtained as shown in Table 6. MPEAR and MinBinder give higher adjusted Rand indices when the model is fit to all data but the indices are lower for the single replicates, a bit more so for MinBinder than for MPEAR. The results for MAP and MedvComp are not so much affected. Looking at the estimates $\widehat{c}$ this seems to be the case because MPEAR and

Table 6: Adjusted Rand index with true clustering for yeast galactose data fitted with additional hyperpriors on the b, v parameters. (Results for replicates are averages over the four replications.)

	MAP	MinBinder	MPEAR	MedvComp	MCLUST
All data	0.960	0.966	0.966	0.960	0.937
Replicates	0.908	0.802	0.815	0.895	0.841

MinBinder can now successfully separate all four groups, only assigning some observations to small or singleton clusters, which as in the simulation study does happen more often for MinBinder than MPEAR. The two criteria however also split one of the large groups for some of the replicates, leading to the lower adjusted Rand indices. MAP and MedvComp still join one small group with the other or a large group but do not split one of the large groups. To see whether the employed model with diagonal covariances is not overly simplistic we fit a DP mixture model with a general covariance matrix as an additional sensitivity analysis. We used the function `DPdensity` of the `R` package `DPpackage` (Jara et al. 2009) with a $Ga(4, 2)$ prior on α and the other hyperparameters set according to Bensmail et al. (1997). The results (data not shown) are comparable to the ones of Table 6, indicating that the simple DP mixture is appropriate for the data.

5 Conclusions and outlook

In this paper we considered ways of choosing a clustering estimate $\widehat{c}$ in Bayesian model-based cluster analysis based on an MCMC sample of clusterings with a varying number of groups K. We proposed new criteria that maximize the adjusted Rand index with the true clustering. The PEAR criteria could be shown to possess a shrinkage property and performed well in a simulation study and in a real data set application, where estimated clusterings closer to the truth than the ones resulting from minimizing Binder's loss could be found. They also compared favorably to the clusterings obtained using MAP, an ad hoc criterion of Medvedovic et al. (2004) and MCLUST. A disadvantage of the PEAR criteria is that they never select a clustering with only one group or with all singleton groups as $\widehat{c}$. Therefore we would generally recommend to use them instead of Binder's loss unless a large fraction of the MCMC clusterings $c^{(1)}, c^{(2)} \ldots, c^{(M)}$ consist of either of the extreme clusterings, in which case minimization of Binder's loss seems more appropriate.

For optimization of the criteria it turned out that hierarchical clustering with $1 - \pi_{ij}$ as distance and average linkage is a quick way to get a good approximation to the optimum. In the minimization of the posterior expectation of Binder's loss it gave $\widehat{c}$'s with almost as low values as obtained with the optimization algorithm of Lau and Green (2007) but took less than one second to compute instead of 30 minutes. Besides the algorithm of Lau and Green one of the clusterings obtained by cutting the hierarchical

cluster tree with average linkage at different levels was chosen as best by the different criteria in almost all cases. Optimizing the criteria only over the clusterings observed in the MCMC sample performed worse and seems not to be a very good way to obtain an estimate $\widehat{c}$. Some theoretical justification for employing the hierarchical approach is given by the fact that $(1 - \pi_{ij})$ could be shown to be a pseudometric for the space of observations.

Although the PEAR criterion was derived having in mind the adjusted Rand index as a performance measure it also gives good results for other criteria for comparing estimated and true clustering, as found in the sensitivity analysis of the simulation study (see Appendix 10). Minimization of the posterior expectation of the distance between estimated and true clustering could nevertheless be done with other measures for comparing clusterings, e.g. the entropy based "variation of information"-distance proposed by Meilă (2007) and we plan to do so in future research.

The post-processing methods based on the posterior similarity matrix can be used on the MCMC output of any Bayesian clustering model not just on the simple DP mixture model used in this paper for illustrative purposes. The results of the sensitivity analyses indicate that the methods are fairly robust to changes in the underlying model, at least if there are clear cluster structures in the data, i.e. for the cases $\delta = 1.5$ and $\delta = 2$ in the simulation study and the galactose data. If there a no clear cluster structures we have found the results to be sensitive to the prior distribution of α.

We are currently preparing an `R` package `mcclust` containing methods for post-processing an MCMC sample of clusterings and code for the simple Dirichlet process mixture model.

6 Appendix: Proof of triangle inequality (6)

It is to show that

$$\begin{array}{rrcl} & (1-\pi_{ij}) & \leq & (1-\pi_{ik})+(1-\pi_{jk}) \\ \Longleftrightarrow & \pi_{ij} & \geq & \pi_{ik}+\pi_{jk}-1 \ . \end{array}$$

Proof: In every possible clustering the observations i,j and k are grouped according to one of the patterns

$$\begin{array}{lll} \text{I} : \{i,j,k\} & \text{II} : \{i,j\},\{k\} & \text{III} : \{i\},\{j,k\} \\ \text{IV} : \{i,k\},\{j\} & \text{V} : \{i\},\{j\},\{k\} \ . & \end{array}$$

Then the following equations hold

$$\begin{aligned} \pi_{ij} &= P(\text{I}|y) + P(\text{II}|y) && (16) \\ \pi_{jk} &= P(\text{I}|y) + P(\text{III}|y) && (17) \\ \pi_{ik} &= P(\text{I}|y) + P(\text{IV}|y) && (18) \\ 1 &= P(\text{I}|y) + \cdots + P(\text{V}|y) \ , && (19) \end{aligned}$$

and with (17) and (18) one obtains

$$\begin{aligned}
\pi_{jk} + \pi_{ik} - 1 &= 2 \cdot P(\mathrm{I}|y) + P(\mathrm{III}|y) + P(\mathrm{IV}|y) - 1 \\
&\overset{(16)}{=} P(\mathrm{I}|y) + P(\mathrm{III}|y) + P(\mathrm{IV}|y) - P(\mathrm{II}|y) + \pi_{ij} - 1 \\
&\overset{(19)}{=} -2 \cdot P(\mathrm{II}|y) - P(\mathrm{V}|y) + \pi_{ij} \\
&\leq \pi_{ij} \qquad \square
\end{aligned}$$

7 Appendix: Proof of shrinkage property (15)

The conditions on $\pi_{(\ell^*)}$ for $PEAR^*$ of equation (14) to take its maximum are considered. This is done by considering under what conditions $PEAR^*(\ell)$ is greater than $PEAR^*(\ell-1)$. Recall that $\pi_{(i)}$ is the ith largest π_{ij}. Then

$$\begin{aligned}
PEAR^*(\ell) &\geq PEAR^*(\ell-1) \\
\iff \frac{\sum_{i=1}^{\ell} \pi_{(i)} - \ell\,\bar{\pi}_{..}}{\frac{1}{2}\ell\,(1-2\bar{\pi}_{..}) + \frac{1}{2}\binom{n}{2}\bar{\pi}_{..}} &\geq \frac{\sum_{i=1}^{\ell-1} \pi_{(i)} - (\ell-1)\,\bar{\pi}_{..}}{\frac{1}{2}(\ell-1)\,(1-2\bar{\pi}_{..}) + \frac{1}{2}\binom{n}{2}\bar{\pi}_{..}} \\
\iff \pi_{(\ell)}[(\ell-1)(1-2\bar{\pi}_{..})] + \binom{n}{2}\bar{\pi}_{..} &\geq (1-2\bar{\pi}_{..}) \sum_{i=1}^{\ell-1} \pi_{(i)} + \binom{n}{2}\bar{\pi}_{..}^2 \\
\iff \binom{n}{2}\bar{\pi}_{..}(\pi_{(\ell)} - \bar{\pi}_{..}) &\geq (1-2\bar{\pi}_{..}) \sum_{i=1}^{\ell-1} (\pi_{(i)} - \pi_{(\ell)}) \ . \qquad (20)
\end{aligned}$$

For $\bar{\pi}_{..} \leq 0.5$ the last inequality is decreasing on the left side with rising ℓ and increasing on the right, so that $PEAR^*$ has a unique maximum (or two maxima at adjacent values). The uniqueness of the maximum can also be shown for $\bar{\pi}_{..} > 0.5$, but the proof is rather technical and is omitted here.
The threshold t can be determined by setting (20) to equality

$$\binom{n}{2}\bar{\pi}_{..}(t - \bar{\pi}_{..}) = (1-2\bar{\pi}_{..}) \sum_{i=1}^{\ell-1} (\pi_{(i)} - t) \ . \qquad (21)$$

Since $\sum_{i=1}^{\ell-1} (\pi_{(i)} - t) > 0$ it follows that

$$\begin{aligned}
t > \bar{\pi}_{..} \quad &\text{if} \quad \bar{\pi}_{..} < 0.5 \\
t = \bar{\pi}_{..} \quad &\text{if} \quad \bar{\pi}_{..} = 0.5 \\
t < \bar{\pi}_{..} \quad &\text{if} \quad \bar{\pi}_{..} > 0.5 \ .
\end{aligned}$$

To prove the relation of t to 0.5 we solve (21) for t and show that for $\bar{\pi}_{..} < 0.5$ it is smaller than 0.5:

$$\begin{aligned} \frac{(1-2\bar{\pi}_{..})\sum_{i=1}^{\ell-1}\pi_{(i)} + \binom{n}{2}\bar{\pi}_{..}^2}{(1-2\bar{\pi}_{..})(\ell-1) + \binom{n}{2}\bar{\pi}_{..}} &< 0.5 \\ \iff (1-2\bar{\pi}_{..})\sum_{i=1}^{\ell-1} 2\pi_{(i)} &< (\ell-1) + \binom{n}{2}(1-2\bar{\pi}_{..}) \\ \iff \sum_{i=1}^{\ell-1}(2\pi_{(i)}-1) &< \binom{n}{2}\bar{\pi}_{..} \ . \end{aligned}$$

The maximum that the term on the left can take is $(\ell-1)$, this is the case if the first $(\ell-1)$ $\pi_{(i)}$ are equal to 1. As $\binom{n}{2}\bar{\pi}_{..} = \sum_i \pi_{(i)}$, the term on the right is at least as large. Equality holds, if $\pi_{(i)} = 1$ for $i \leq \ell-1$ and all other $\pi_{(i)} = 0$.
For $\bar{\pi}_{..} > 0.5$ the direction of the inequality is changed when dividing by $(1-2\bar{\pi}_{..})$, so that the proof is complete.

8 Appendix: Additional graph

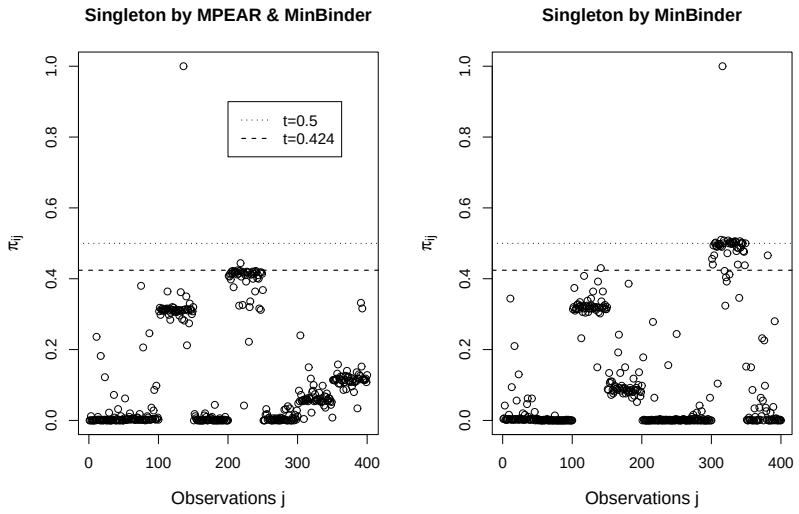

Figure 2: Pairwise posterior probabilities π_{ij} for two observations i. Equal cluster size data with $\delta = 2$ and $\bar{\pi}_{..}$=0.119. Left: Observation is put into its own cluster by MPEAR and MinBinder. Right: Observation is put into its own cluster only by MinBinder. Lines indicate thresholds t for MPEAR (- - -) and MinBinder ($\cdots$), t is described in Section 3.2.

9 Appendix: Simple Dirichlet process mixture model

The used model is similar to the one proposed by Qin (2006). It assumes

$$\begin{aligned} y|\mu,\tau &\sim N_d(\mu,\tau^{-1}I_d) \\ \mu,\tau|G &\sim G \\ G &\sim DP(\alpha,p(\mu,\tau)) \ , \end{aligned}$$

with random probability measure G. Clustering is induced on common values of (μ,τ). The centering distribution is chosen as a conjugate Normal-Gamma

$$\begin{aligned} p(\mu,\tau) &= p(\mu|\tau)p(\tau) \\ &= N_d(0,\tau^{-1}v^{-1}I_d)Ga(a,b) \ , \end{aligned}$$

which considerably simplifies Gibbs sampling of the class indicators. An iteration of the MCMC algorithm consists of one conjugate Gibbs scan and three split-merge proposals as described by Dahl (2005). We found that the latter are beneficial in reducing the autocorrelation of the chain.

Setting of the hyperparameter α is discussed in Section 4.1. Priors can be assigned to v and either a or b as well. Convenient choices that lead to full conditionals of known form are Gamma priors for v and b.

Since the model is conjugate it is possible to analytically integrate out the parameters μ and τ and use the marginal posterior of the allocations $p(c^*|y) \propto p(y|c^*)p(c^*)$ in computing the posterior density to find the MAP. For the Dirichlet process the prior probability on the allocations is given by

$$p(c^*) = \frac{\prod_{k=1}^{K} \alpha\,\Gamma(n_k)}{\prod_{i=1}^{n}(\alpha+i-1)} \ , \tag{22}$$

with n_k being the number of observations in cluster k (Quintana and Iglesias 2003), for formulas for $p(y|c^*)$ see Dahl (2005). If priors have been assigned to the hyperparameters a draw of their value given c^* is also taken into account in computing the posterior density.

Application of the model is of course only sensible if one is looking for spherical clusters. In other cases the model could be extended in various fashions by using a more flexible likelihood or centering distribution as might be necessary in certain clustering problems. It suffices for our purpose of comparing various ways of post-processing an MCMC sample of clusterings.

The model is implemented in the statistical software `R` (R Development Core Team 2009) where `C` functions are called for the time-demanding Gibbs sampling and split-merge steps.

9.1 Details on MCMC sampler

The conditional distribution of c_i given all other indicators c_{-i} and y is given by

$$P(c_i = k|c_{-i}, y) \propto \frac{n_{k,-i}}{\alpha + n - 1} \int N(y_i|\mu, \tau) p(\mu, \tau|y_{k,-i})\, d\mu\, d\tau \tag{23}$$

$$P(c_i = K + 1|c_{-i}, y) \propto \frac{\alpha}{\alpha + n - 1} \int N(y_i|\mu, \tau) p(\mu, \tau)\, d\mu\, d\tau \ , \tag{24}$$

where K is the number of clusters in c_{-i}, $n_{k,-i}$ is the number of $c_{-i} = k$ and $y_{k,-i}$ are the corresponding observations. Conditioning on current draws of any of the hyperparameters α, b and v that is assigned a prior distribution is also assumed. Since $p(\mu|\tau)p(\tau) = N_d(0, \tau^{-1}v^{-1}I)Ga(a, b)$ the integral in (24) can be solved by first integrating with respect to μ, leading to $p(y_i|c_i = K + 1, \tau) = N_d(0, \tau^{-1}(1 + v^{-1})I)$ and then using results of Bernardo and Smith (1994, p.140) to obtain

$$p(y_i|c_i = K + 1) = t_d(0, \frac{b}{a}(1 + v^{-1})I, 2a) \ ,$$

where $t_d(\eta, \Sigma, \nu)$ is the d-dimensional Student t-distribution with expectation η (for $\nu > 1$) and variance $\frac{\nu}{\nu-2}\Sigma$ (for $\nu > 2$).
The integral in (23) can be solved by first employing standard results in Bayesian inference to give $p(\mu|\tau, y_{k,-i})p(\tau|y_{k,-i}) = N_d(\mu^*, \tau^{-1}v^{*-1}I)Ga(a^*, b^*)$, where

$$\begin{aligned} \mu^* &= \frac{n_{k,-i}}{v + n_{k,-i}} \bar{y}_{k,-i} \\ v^* &= v + n_{k,-i} \\ a^* &= a + dn_{k,-i}/2 \\ b^* &= b + \frac{1}{2}[\sum_{\substack{j \neq i \\ c_j = k}} y_j^T y_j - v^* \mu^{*T} \mu^*] \ . \end{aligned}$$

Applying the same reasoning as above one then obtains

$$p(y_i|c_i = k, c_{-i}, y_{k,-i}) = t_d(\mu^*, \frac{b^*}{a^*}(1 + v^{*-1})I, 2a^*) \ .$$

If the hyperparameter α is assigned a $Ga(\delta_1, \delta_2)$ prior the Gibbs sampling scheme of Escobar and West (1995) is employed.
A $Beta(v_1, v_2)$ prior on $1/(1 + \alpha)$ induces

$$p(\alpha) = \frac{\Gamma(v_1 + v_2)}{\Gamma(v_1)\Gamma(v_2)} \alpha^{v_2 - 1}(1 + \alpha)^{-(v_1 + v_2)} \ ,$$

leading to a full conditional depending only on K and n

$$p(\alpha|K, n) \propto \frac{\alpha^{K + v_2 - 1}}{(1 + \alpha)^{(v_1 + v_2)}} \frac{\Gamma(\alpha)}{\Gamma(\alpha + n)}$$

which is updated with a random walk Metropolis algorithm on $\log(\alpha)$.
If v is assigned a $Ga(\gamma_1, \gamma_2)$ prior its full conditional is given by a

$$Ga(\gamma_1 + \frac{K \cdot d}{2}, \gamma_2 + \frac{1}{2}\sum_{k=1}^{K} \tau_k \cdot \mu_k^T \mu_k) \ ,$$

where μ_k, τ_k are sampled for each cluster from the $N_d(\mu^*, \tau^{-1} v^{*-1} I) Ga(a^*, b^*)$ distribution given above. Similarly, if b is assigned a $Ga(\eta_1, \eta_2)$ prior the its full conditional is a $Ga(\eta_1 + K \cdot a, \eta_2 + \sum_{k=1}^{K} \tau_k)$.

10 Appendix: Sensitivity analysis of simulation study

To evaluate the effect of the number of MCMC iterations some of the data sets have been fitted with twice the number of iterations, which did not improve the results (data not shown). More iterations seemed only to be beneficial if the optimization of the criteria is done solely over the drawn clusterings $c^{(1)}, c^{(2)} \ldots, c^{(M)}$, where in some case better scoring clusterings could be found.

Table 7: Average adjusted Rand index with the true clustering for equal cluster size data and different prior settings.

		$\delta = 1.5$		
Prior	MAP	MinBinder	MPEAR	MedvComp
Standard	0.000	0.485	0.601	0.476
Gamma b,v	0.000	0.393	0.568	0.456
Vague α	0.000	0.414	0.547	0.326
Fixed α	0.000	0.428	0.592	0.550
		$\delta = 2$		
Prior	MAP	MinBinder	MPEAR	MedvComp
Standard	0.836	0.824	0.837	0.833
Gamma b,v	0.784	0.793	0.821	0.827
Vague α	0.838	0.826	0.840	0.835
Fixed α	0.831	0.820	0.831	0.836

To investigate the effect that the underlying clustering model has on the estimated clusterings the DP mixture model is also fit to the simulated data with different priors for the hyperparameters. The first modification consists of placing additional $Ga(1, 1)$ priors on both b and v. For the second setting the $Ga(4, 2)$ on α is replaced by a vague prior. In a clustering context a good choice for a vague prior seems to be to let $P(c_i = c_j) = 1/(1 + \alpha) \sim Beta(1, 1)$, leading to $p(\alpha) = (1 + \alpha)^{-2}$. The last setting has a fixed $\alpha = 4$.

Table 7 shows the average adjusted Rand indices that resulted from applying the estimation methods to the output of the model with the different priors for the equal

cluster size data and δ equal to 1.5 and 2. For $\delta = 2$ the results are relatively robust for all criteria. In the case of $\delta = 1.5$ the average adjusted Rand indices are generally a bit worse for the "Gamma b, v" and "Vague α" setting, indicating that here the added prior flexibility makes it harder to find the true cluster structure. The results of MPEAR are the least affected by the different priors and still lead to the best average result. The results for δ equal to 0.5 and 1 do not change much compared to the standard prior (data not shown), except that for $\delta = 1$ and the "Vague α" setting none of the criteria can find any cluster structure and that the results are better for MinBinder and MPEAR for the "Fixed α" setting.

A similar pattern is found for the unequal cluster size data while the results for the extreme case data sets are not affected by the different priors.

We also looked at the results for other performance measures than the adjusted Rand. Using instead the original Rand index both MinBinder and MPEAR give better results than the other criteria, with the two criteria performing very similar. If a performance measure derived from different ideas like the entropy based "variation of information"-distance proposed by Meilă (2007) is employed the results of the simulation study still do not change much (data not shown).

References

Antoniak, C. E. (1974). "Mixtures of Dirichlet Processes with Applications to Bayesian Nonparametric Problems." *Annals of Statistics*, 2: 1152–1174. 376

Bensmail, H., Celeux, G., Raftery, A. E., and Robert, C. P. (1997). "Inference in Model-Based Cluster Analysis." *Statistics and Computing*, 7: 1–10. 381

Bernardo, J. M. and Smith, A. F. M. (1994). *Bayesian Theory*. New York: Wiley. 386

Binder, D. A. (1978). "Bayesian Cluster Analysis." *Biometrika*, 65: 31–38. 367, 368, 370

Blackwell, D. and MacQueen, J. B. (1973). "Ferguson Distributions via Polya Urn Schemes." *Annals of Statistics*, 1: 353–355. 369

Dahl, D. B. (2005). "Sequentially-Allocated Merge-Split Sampler for Conjugate and Nonconjugate Dirichlet Process Mixture Models." *Journal of Computational and Graphical Statistics*, under revision (preprint available from author's web page). 385

— (2006). "Model-Based Clustering for Expression Data via a Dirichlet Process Mixture Model." In Do, K. A., Müller, P., and Vannucci, M. (eds.), *Bayesian Inference for Gene Expression and Proteomics*, 201–218. Cambridge University Press. 368, 371, 373, 375

Dahl, D. B. and Newton, M. A. (2007). "Multiple Hypothesis Testing by Clustering Treatment Effects." *Journal of the American Statistical Association*, 102: 517–526. 369

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). "Maximum Likelihood from Incomplete Data via the EM Algorithm." *Journal of the Royal Statistical Society, Ser. B*, 39: 1–38. 368

Dunson, D. B. (2008). "Nonparametric Bayes Applications to Biostatistics." *Department of Statistical Science, Duke University, Durham, NC, USA*, (Technical Report 6). 369

Escobar, M. D. and West, M. (1995). "Bayesian Density Estimation and Inference Using Mixtures." *Journal of the American Statistical Association*, 90: 577–588. 376, 386

Ferguson, T. S. (1973). "A Bayesian Analysis of Some Nonparametric Problems." *Annals of Statistics*, 1: 209–230. 369

Fraley, C. and Raftery, A. E. (2002). "Model-Based Clustering, Discriminant Analysis and Density Estimation." *Journal of the American Statistical Association*, 97: 611–631. 368

Gene Ontology Consortium (2000). "Gene Ontology: Tool for the Unification of Biology." *Nature Genetics*, 25: 25–29. 380

Griffin, J. E. and Steel, M. F. J. (2006). "Order-Based Dependent Dirichlet Processes." *Journal of the American Statistical Association*, 101: 179–194. 376

Hubert, L. and Arabie, P. (1985). "Comparing Partitions." *Journal of Classification*, 2: 193–218. 372

Hurn, M., Justel, A., and Robert, C. P. (2003). "Estimating Mixtures of Regressions." *Journal of Computational and Graphical Statistics*, 12: 55–79. 371

Ideker, T., Thorsson, V., Ranish, J. A., Christmas, R., Buhler, J., Eng, J. K., Bumgarner, R., Goodlett, D. R., Aebersold, R., and Hood, L. (2001). "Integrated Genomic and Proteomic Analyses of a Systematically Perturbed Metabolic Network." *Science*, 292: 929–934. 380

Ishwaran, H. and James, L. F. (2001). "Gibbs Sampling Methods for Stick-Breaking Priors." *Journal of the American Statistical Association*, 96: 161–173. 369

Jara, A., Hanson, T., Quintana, F. A., Müller, P., and Rosner, G. L. (2009). "DPpackage: Bayesian Nonparametric and Semiparametric Analysis." R package: 1.0–7. 381

Jensen, S. T. and Liu, J. S. (2008). "Bayesian Clustering of Transcription Factor Binding Motifs." *Journal of the American Statistical Association*, 103: 188–200. 369

Johnson, N. L., Kotz, S., and Balakrishan, N. (1995). *Continuous Univariate Distributions, Volume 2*. New York: Wiley, 2nd edition. 376

Kaufman, L. and Rousseuw, P. J. (1990). *Finding Groups in Data*. New York: Wiley. 370

Kim, S., Tadesse, M. G., and Vannucci, M. (2006). "Variable Selection in Clustering via Dirichlet Process Mixture Models." *Biometrika*, 93: 877–893. 369

Lau, J. W. and Green, P. J. (2007). "Bayesian Model-Based Clustering Procedures." *Journal of Computational and Graphical Statistics*, 16: 526–558. 369, 375, 377, 381

Lijoi, A., Mena, R. H., and Prünster, I. (2007). "Controlling the Reinforcement in Bayesian Non-Parametric Mixture Models." *Journal of the Royal Statistical Society, Ser. B*, 69: 715–740. 369

Medvedovic, M., Yeung, K., and Bumgarner, R. (2004). "Bayesian Mixture Model Based Clustering of Replicated Microarray Data." *Bioinformatics*, 20: 1222–1232. 368, 370, 373, 375, 376, 380, 381

Meilă, M. (2007). "Comparing Clusterings – an Information Based Distance." *Journal of Multivariate Analysis*, 98: 873–895. 382, 388

Milligan, G. W. and Cooper, M. C. (1986). "A Study of the Comparability of External Criteria for Hierarchical Cluster Analysis." *Multivariate Behavioral Research*, 21: 441–458. 372

Pitman, J. and Yor, M. (1997). "The Two-Parameter Poisson-Dirichlet Distribution Derived from a Stable Subordinator." *Annals of Probability*, 25: 855–900. 369

Qin, Z. S. (2006). "Clustering Microarray Gene Expression Data Using Weighted Chinese Restaurant Process." *Bioinformatics*, 22: 1988–1997. 380, 385

Quintana, F. A. and Iglesias, P. L. (2003). "Bayesian Clustering and Product Partition Models." *Journal of the Royal Statistical Society, Ser. B*, 65: 557–574. 369, 385

R Development Core Team (2009). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.
URL `http://www.R-project.org` 385

Rand, W. M. (1971). "Objective Criteria for the Evaluation of Clustering Methods." *Journal of the American Statistical Association*, 66: 846–850. 371

Richardson, S. and Green, P. J. (1997). "On Bayesian Analysis of Mixtures with an Unknown Number of Components." *Journal of the Royal Statistical Society, Ser. B*, 59: 731–792. 368

Stephens, M. (2000). "Dealing with Label Switching in Mixture Models." *Journal of the Royal Statistical Society, Ser. B*, 62: 795–809. 368

Tadesse, M. G., Sha, N., and Vannucci, M. (2005). "Bayesian Variable Selection in Clustering High-Dimensional Data." *Journal of the American Statistical Association*, 100: 602–617. 369

Acknowledgments

The authors wish to thank Björn Bornkamp and Robert Wolpert as well as the Editor, the Associate Editor and two anonymous referees for helpful comments. Katja Ickstadt gratefully acknowledges financial support of the Deutsche Forschungsgemeinschaft (SFB 475, "Reduction of Complexity in Multivariate Data Structures").