Clustering Binary Data with Bernoulli Mixture Models

Neal S. Grantham

Abstract Clustering is an unsupervised learning technique that seeks "natural" groupings in data. One form of data that has not been widely studied in the context of clustering is binary data. A rich statistical framework for clustering binary data is the Bernoulli mixture model for which there exists both Bayesian and non-Bayesian approaches. This paper reviews the development and application of Bernoulli mixture models to clustering binary data.

1 Introduction

Cluster analysis is the separation of heterogeneous data into groups (or clusters) such that data within the same cluster are similar and data between clusters are dissimilar. Clustering marks one approach to unsupervised learning, seeking "natural" groupings in the data without reliance on labeled examples to supervise classification. For example, a large and fast-growing document database may aim to organize documents by topic without relying on a user to supervise the organization process. Clustering methods take many forms (Jain and Dubes, 1988), but for the purposes of this paper we restrict attention to a finite mixture model approach as it allows for "the important question of how many clusters there are in the data to be posed within the framework of standard statistical theory" (Marriott, 1974).

Finite mixture models provide a convenient framework to model population heterogeneity and facilitate clustering (McLachlan and Peel, 2000). Heterogeneity in a population is reframed as arising from the pooling (or mixture) of a finite collection of relatively homogeneous subpopulations. The problem lies, then, in unobserved heterogeneity (Böhning and Seidel, 2003), as it is not known which, or possibly even how many, subpopulations are responsible for producing the data observed. Using a variety of advanced statistical methods, one can estimate parameters of the subpopulations (mixture components), the weighting given to these parameters (mixing weights), and, if not known *a priori*, the number of subpopulations. With these estimates in hand, clustering sim-

ply becomes a matter of using Bayes' rule to classify data as belonging to the mixture components most likely to have produced them.

Parameter estimation in mixture models is not cheap, however. In the first recorded use of mixture models, Pearson (1894) sought method of moments estimators for a mixture of just two normal probability densities by analytically solving for the roots of a ninth degree polynomial; a herculean task in the 19th century. Mixture methodology therefore saw little development until the late 20th century, when the introduction of the EM algorithm (Dempster et al., 1977) and increased accessibility to high-speed computers made demonstrable impacts on the theory and application of mixture models, and consequently cluster analysis, to a wide array of problems (McLachlan and Peel, 2000). Moreover, development of Markov chain Monte Carlo methods have enabled the use of Bayesian approaches to cluster analysis (Marin et al., 2005).

While literature on clustering via mixture models has expanded in recent decades, attention has primarily been paid to modeling continuous data with Gaussian mixtures and applications to binary data remain scarce (Bouguila, 2010). Binary data have a rich history in the areas of text mining (Wang and Kabán, 2005) and topical document classification (Li, 2006), handwritten digit recognition (Bishop, 2006; Grim et al., 2000), sequencing of packets in sensor networks (Kamthe et al., 2011), and the identification of item sale association rules (Agrawal et al., 1994), in addition to numerous biological applications in DNA computing (Fränti et al., 2003), human genetics (Abel et al., 1993), and microbiology (Gyllenberg et al., 1997). Extending from the first application of mixture models to binary data (Celeux and Govaert, 1991), we present a synthesis of the leading classical and Bayesian approaches to unsupervised clustering within the specific context of multivariate binary data.

The paper proceeds as follows. Section 2 presents the finite mixture model framework for multivariate binary data. Section 3 describes standard approaches to parameter estimation when the number of mixture components is known *a priori*. When the number of mixture components

is unknown, methods in Section 4 provide means to select or estimate the number of components in concert with classical and Bayesian estimation procedures. Finally, Section 5 surveys feature selection techniques for high-dimensional binary data, and Section 6 summarizes and addresses areas of future research in clustering binary data.

2 Bernoulli Mixture Model

Let $\boldsymbol{X} = (\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_N)$ be a random sample of D-dimensional binary vectors. We aim to partition $\boldsymbol{X}$ into K (possibly unknown but finite) clusters such that vectors within the same cluster are similar and vectors between clusters are dissimilar. A vector $\boldsymbol{X}_n$ is assumed to arise from a finite mixture density, $f(\boldsymbol{X}_n|K, \boldsymbol{p}, \boldsymbol{\Theta}) = \sum_{k=1}^{K} p_k q(\boldsymbol{X}_n|\boldsymbol{\theta}_k)$, where q is the mixture component density, $\boldsymbol{\Theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K)$ are mixture parameters with $\boldsymbol{\theta}_k$ specific to component $q(\cdot|\boldsymbol{\theta}_k)$, and $\boldsymbol{p} = (p_1, \ldots, p_K)$ are the mixing weights, $\sum_{k=1}^{K} p_k = 1$, $p_k > 0, k = 1, \ldots, K$. For convenience, let $\boldsymbol{\Psi} = (\boldsymbol{p}, \boldsymbol{\Theta})$ and, when K may be unknown, $\boldsymbol{\Psi}_K = (K, \boldsymbol{\Psi})$. As $\boldsymbol{X}_n$ is binary and its components assumed independent, i.e. $\boldsymbol{X}_n = (X_{n1}, \ldots, X_{nD})$ with $X_{nd} \in \{0, 1\}, d = 1, \ldots, D$ and $X_{nd_1} \perp\!\!\!\perp X_{nd_2}$, a natural choice for q is the multivariate Bernoulli distribution with independent components. Letting $\boldsymbol{\theta}_k = (\theta_{k1}, \ldots, \theta_{kD})$ for $k = 1, \ldots, K$ with $0 \leq \theta_{kd} \leq 1, d = 1, \ldots, D$, the Bernoulli mixture model (BMM) from which vectors $\boldsymbol{X}_n$ are independently drawn conditional on $\boldsymbol{\Psi}_K$ is given by

$$f(\boldsymbol{X}_n|\boldsymbol{\Psi}_K) = \sum_{k=1}^{K} p_k \prod_{d=1}^{D} \theta_{kd}^{X_{nd}} (1 - \theta_{kd})^{1 - X_{nd}}, \tag{1}$$

leading to a likelihood over N values of

$$L(\boldsymbol{\Psi}_K|\boldsymbol{X}) = f(\boldsymbol{X}|\boldsymbol{\Psi}_K) = \prod_{n=1}^{N} \sum_{k=1}^{K} p_k \prod_{d=1}^{D} \theta_{kd}^{X_{nd}} (1 - \theta_{kd})^{1 - X_{nd}}. \tag{2}$$

Furthermore, consider addressing the unobserved heterogeneity mentioned in Section 1. One can imagine an unobserved $\boldsymbol{Z} = (\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_N)$ composed of latent membership (or allocation) vectors defined so that for $\boldsymbol{Z}_n = (Z_{n1}, \ldots, Z_{nK})$, $Z_{nk} = 1$ if $\boldsymbol{X}_n$ belongs to component k and 0 otherwise. That is, every $\boldsymbol{Z}_n$ in $\boldsymbol{Z}$ records exactly which mixture component is responsible for

producing $\boldsymbol{X}_n$ in $\boldsymbol{X}$. Assume that $\boldsymbol{Z}_n$'s are independent conditional on K and $\boldsymbol{p}$ and express $Pr(Z_{nk} = 1|K, \boldsymbol{p}) = p_k$ so that $f(\boldsymbol{Z}_n|K, \boldsymbol{p}) = \prod_{k=1}^{K} p_k^{Z_{nk}}$. In other words, $\boldsymbol{Z}_n$ represents a single draw from a Multinomial distribution with proportions $\boldsymbol{p}$. Additionally, $Z_{nk} = 1$ in $\boldsymbol{Z}_n$ records the membership of $\boldsymbol{X}_n$ to component k, so $f(\boldsymbol{X}_n|\boldsymbol{Z}_n, K, \boldsymbol{\Theta}) = \prod_{k=1}^{K} \left[\prod_{d=1}^{D} \theta_{kd}^{X_{nd}}(1-\theta_{kd})^{1-X_{nd}}\right]^{Z_{nk}}$. Combining these two densities together with $f(\boldsymbol{X}_n, \boldsymbol{Z}_n|\boldsymbol{\Psi}_K) = f(\boldsymbol{X}_n|\boldsymbol{Z}_n, K, \boldsymbol{\Theta})f(\boldsymbol{Z}_n|K, \boldsymbol{p})$ for all n, the complete-case likelihood is given by

$$L_C(\boldsymbol{\Psi}_K|\boldsymbol{X}, \boldsymbol{Z}) = f(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\Psi}_K) = \prod_{n=1}^{N}\prod_{k=1}^{K}\left[p_k \prod_{d=1}^{D} \theta_{kd}^{X_{nd}}(1-\theta_{kd})^{1-X_{nd}}\right]^{Z_{nk}}, \tag{3}$$

which will prove useful for clustering. The following two sections explore methods that address parameter estimation and clustering when the number of components K is known (Section 3) and unknown (Section 4).

3 Number of Components K is Known

In some situations, it is known ahead of time how many clusters should be formed from the data. For example, Bishop (2006) assigns binary images of handwritten numerals $2, 3,$ and 4 into one of $K = 3$ clusters (one for each digit), and Bouguila (2010) applies text mining to 4,199 university computer science department webpages to cluster pages into one of $K = 4$ possible categories (Course, Faculty, Project, and Student). In situations such as these, joint estimation of $\boldsymbol{\Psi}$ and $\boldsymbol{Z}$ in (3) is achieved through standard applications of the EM algorithm or Gibbs sampling. The CEM algorithm is also an appropriate option.

3.1 EM Algorithm

Mixture models lend themselves nicely to the "complete-case" framework required by the EM algorithm (McLachlan and Krishnan, 1997). The EM algorithm seeks maximization of (2) using the complete-case likelihood (3) and an iterative Expectation (E-step) and Maximization (M-step) procedure. Its application to a BMM is outlined in Algorithm 1. At iteration t of Algorithm 1, Steps 3 & 4 mark the E-step of the algorithm, where $\boldsymbol{Z}_n^{*(t)}$ is updated by $E_{\boldsymbol{\Psi}^{(t-1)}}(\boldsymbol{Z}_n|\boldsymbol{X})$ and Steps

5, 6, & 7 mark the M-step, where $\boldsymbol{\Psi}^{(t)}$ is obtained by maximizing $E_{\boldsymbol{\Psi}^{(t-1)}}(\log L_C(\boldsymbol{\Psi})|\boldsymbol{X})$ with respect to $\boldsymbol{\Psi}$. Because the EM algorithm converges only to a local maximum, it is recommended one initialize the algorithm at many different starting values, iterate to convergence, and select the best maximizer of (2) as one's estimates. Note that $\boldsymbol{Z}_n^{*(t)}$ includes an asterisk because it is no longer binary after the first iteration and instead represents "fuzzy" clustering in that it partially allocates $\boldsymbol{X}_n$ to numerous clusters.

Algorithm 1: EM Algorithm for BMM (K fixed)

initialize Set $t \leftarrow 1$. Choose starting values $\boldsymbol{p}^{(0)}$ and $\boldsymbol{\theta}_k^{(0)} = (\theta_{k1}^{(0)}, \ldots, \theta_{kD}^{(0)})$ (for all k)

repeat

  Compute $s_{nk} = p_k^{(t-1)} \prod_{d=1}^{D} \left[\theta_{kd}^{(t-1)}\right]^{X_{nd}} \left[1 - \theta_{kd}^{(t-1)}\right]^{1-X_{nd}}$ and $s_{n\cdot} = \sum_{k=1}^{K} s_{nk}$ (for all n, k)

  Assign $\boldsymbol{Z}_n^{*(t)} \leftarrow \boldsymbol{S}_n$ with $\boldsymbol{S}_n = (s_{n1}, \ldots, s_{nK})/s_{n\cdot}$ (for all n)

  Compute $u_k = \sum_{n=1}^{N} Z_{nk}^{*(t)}$ and $v_{kd} = \sum_{n=1}^{N} Z_{nk}^{*(t)} X_{nd}$ (for all k, d)

  Assign $\boldsymbol{p}^{(t)} \leftarrow \boldsymbol{u}/N$ with $\boldsymbol{u} = (u_1, \ldots, u_K)$

  Assign $\boldsymbol{\theta}_k^{(t)} \leftarrow \boldsymbol{v}_k/u_k$ with $\boldsymbol{v}_k = (v_{k1}, \ldots, v_{kD})$ (for all k)

  Assign $t \leftarrow t + 1$

until $|\log L(\boldsymbol{\Psi}^{(t)}|\boldsymbol{X}) - \log L(\boldsymbol{\Psi}^{(t-1)}|\boldsymbol{X})| < \varepsilon$, *where* ε *is small*

3.1.1 Bayesian Extension

The EM algorithm is easily altered to converge to *maximum a posteriori* (MAP) estimates rather than maximum likelihood estimates. That is, simply modify the ML approach (Algorithm 1) which maximizes (2) (or, equivalently, maximizes $\log L$) to accommodate a MAP approach by seeking maximization of $\log L + \log \pi(\boldsymbol{\Psi})$, where $\pi(\boldsymbol{\Psi})$ denotes a prior distribution on $\boldsymbol{\Psi}$. For binary data, Ripley (1996) supports use of priors to prevent overfitting in the case of small N.

3.2 CEM Algorithm

A slight modification to the EM algorithm produces the Classification-EM (CEM) algorithm (Celeux and Govaert, 1992) which maximizes the classification maximum likelihood (3) (Symons,

1981). The CEM algorithm introduces Classification (C-step) between Steps 3 & 4 (E-step) and Steps 5–7 (M-step) of Algorithm 1. Here, rather than the "fuzzy" classification of $\boldsymbol{X}_n$ to K clusters designated by $\boldsymbol{Z}_n^*$, a "hard" classification is made to place $\boldsymbol{X}_n$ into its single most likely cluster of origin. Specifically, at iteration t, assign $\boldsymbol{Z}_n^{(t)} \leftarrow (Z_{n1}^{(t)}, \ldots, Z_{nK}^{(t)})$ such that $Z_{nk}^{(t)} = 1$ if $\max(\boldsymbol{Z}_n^{*(t)}) = Z_{nk}^{*(t)}$ and 0 otherwise. Then Step 5 is modified so that $u_k = \sum_{n=1}^N Z_{nk}^{(t)}$ and $v_{kd} = \sum_{n=1}^N Z_{nk}^{(t)} X_{nd}$. For binary data, the CEM algorithm is found to perform better than the EM algorithm for small sample sizes with well-separated mixture components (Govaert and Nadif, 1996) but is outperformed by the EM algorithm in most other scenarios and therefore sees little use.

3.3 Gibbs Sampling

In a Bayesian context, $\boldsymbol{\Psi}$ is treated as random. For a prior distribution $\pi(\boldsymbol{\Psi})$ placed on $\boldsymbol{\Psi}$ and complete-case likelihood (3) the posterior distribution is given by

$$\pi(\boldsymbol{\Psi}|\boldsymbol{X}, \boldsymbol{Z}) \propto \pi(\boldsymbol{\Psi}) f(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\Psi}) = \pi(\boldsymbol{p}) f(\boldsymbol{Z}|\boldsymbol{p}) \cdot \pi(\boldsymbol{\Theta}) f(\boldsymbol{X}|\boldsymbol{Z}, \boldsymbol{\Theta}) \tag{4}$$

where individual parameter priors are taken to be independent such that $\pi(\boldsymbol{\Psi}) = \pi(\boldsymbol{p})\pi(\boldsymbol{\Theta})$.

3.3.1 Conjugate Priors & Posterior Sampling

We seek conjugate priors so that the marginal posterior distributions of $\boldsymbol{p}$ and $\boldsymbol{\Theta}$ share the same distributional form as their respective priors. The vector of mixing weights $\boldsymbol{p}$ lies on a $(K-1)$-dimensional simplex, so a logical prior choice is the Dirichlet distribution (with hyperparameters $\alpha_1, \ldots, \alpha_K$) which is conjugate to the Multinomial distribution assumed for each $\boldsymbol{Z}_n$. Furthermore, for all k and d, θ_{kd} in $\boldsymbol{\theta}_k$ is chosen to follow a Beta distribution (with hyperparameters γ_{kd} and δ_{kd}) to achieve conjugacy with the univariate Bernoulli distributions of each X_{nd}. Thus, over all k and

d, the priors chosen are

$$\pi(\boldsymbol{p}) = \frac{\Gamma(\alpha_0)}{\prod_{k=1}^{K}\Gamma(\alpha_k)}\prod_{k=1}^{K} p_k^{\alpha_k-1} \quad \text{and} \quad \pi(\theta_{kd}) = \frac{\Gamma(\gamma_{kd}+\delta_{kd})}{\Gamma(\gamma_{kd})\Gamma(\delta_{kd})}\theta_{kd}^{\gamma_{kd}-1}(1-\theta_{kd})^{\delta_{kd}-1}, \tag{5}$$

where $\alpha_0 = \sum_{k=1}^{K}\alpha_k$, yielding a posterior distribution

$$\pi(\boldsymbol{\Psi}|\boldsymbol{X},\boldsymbol{Z}) \propto \prod_{n=1}^{N}\prod_{k=1}^{K}\left[p_k^{Z_{nk}+\alpha_k-1}\prod_{d=1}^{D}\theta_{kd}^{Z_{nk}X_{nd}+\gamma_{kd}-1}(1-\theta_{kd})^{Z_{nk}(1-X_{nd})+\delta_{kd}-1}\right] \tag{6}$$

$$\propto \prod_{k=1}^{K}\left[p_k^{u_k+\alpha_k-1}\prod_{d=1}^{D}\theta_{kd}^{v_{kd}+\gamma_{kd}-1}(1-\theta_{kd})^{u_k-v_{kd}+\delta_{kd}-1}\right] \tag{7}$$

where $u_k = \sum_{n=1}^{N} Z_{nk}$ and $v_{kd} = \sum_{n=1}^{N} Z_{nk}X_{nd}$. If there is no prior information to help distinguish the K components from one another, it is common to select $\alpha_1 = \cdots = \alpha_K = 1$ and $\gamma_{11} = \cdots = \gamma_{KD} = \delta_{11} = \cdots = \delta_{KD} = 1$ so that the priors are deemed symmetric.

Algorithm 2: Gibbs Sampling for BMM (K fixed)

initialize Set $t \leftarrow 1$. Choose starting values $\boldsymbol{p}^{(0)}$ and $\boldsymbol{\theta}_k^{(0)} = (\theta_{k1}^{(0)}, \ldots, \theta_{kD}^{(0)})$ (for all k)

repeat

- Compute $s_{nk} = p_k^{(t-1)}\prod_{d=1}^{D}\left[\theta_{kd}^{(t-1)}\right]^{X_{nd}}\left[1-\theta_{kd}^{(t-1)}\right]^{1-X_{nd}}$ and $s_{n\cdot} = \sum_{k=1}^{K} s_{nk}$ (for all n, k)
- Generate $\boldsymbol{Z}_n^{(t)}$ from $Multinomial(1, \boldsymbol{S}_n)$ with $\boldsymbol{S}_n = (s_{n1}, \ldots, s_{nK})/s_{n\cdot}$ (for all n)
- Compute $u_k = \sum_{n=1}^{N} Z_{nk}^{(t)}$ and $v_{kd} = \sum_{n=1}^{N} Z_{nk}^{(t)}X_{nd}$ (for all k, d)
- Generate $\boldsymbol{p}^{(t)}$ from $Dirichlet(\alpha_1 + u_1, \ldots, \alpha_K + u_K)$
- Generate $\theta_{kd}^{(t)}$ from $Beta(\gamma_{kd} + v_{kd}, \delta_{kd} + u_k - v_{kd})$ (for all k, d)
- Assign $t \leftarrow t + 1$

until t *is suitably large*

Joint samples are drawn from the posterior distribution (7) using Markov chain Monte Carlo (MCMC) methods popularized in the seminal paper by Gelfand and Smith (1990) and preceded by "data augmentation" in Tanner and Wong (1987). Diebolt and Robert (1994) proposed a Gibbs sampler for mixture models which we make use of here for a BMM. The sampling scheme is

presented in Algorithm 2. After obtaining a suitably large number of samples from the posterior, point estimates for $\boldsymbol{p}$ and each θ_{kd} are obtained by evaluating their posterior mean and each $\boldsymbol{Z}_n$ is estimated via its posterior mode.

3.3.2 Label-Switching Problem

If symmetric priors are selected for the parameters, then a major issue that must be tackled in Bayesian estimation under the mixture model framework is the label-switching problem. Let ν denote a permutation of $1, \ldots, K$ with corresponding parameter permutation given by $\nu(\boldsymbol{\Psi}) = ((p_{\nu(1)}, \ldots, p_{\nu(K)}), (\boldsymbol{\theta}_{\nu(1)}, \ldots, \boldsymbol{\theta}_{\nu(K)}))$. Then the mixture likelihood (2) is the same for all $K!$ possible $\nu(\boldsymbol{\Psi})$, or, stated another way, is invariant to switching of the component labels (Redner and Walker, 1984). If there is no prior information to distinguish between the parameters of the K mixture components, then the resulting posterior surface will retain this symmetry in the form of $K!$ modes, none of which uniquely identify the mixing weights or mixture parameters of the individual components. Therefore, as the MCMC sampler traverses the posterior surface and visits many different modes the labels of the components permute. Without accounting for these permutations, naïve analysis of the resulting posterior sample yields posterior means $\boldsymbol{p} \approx (1/K, \ldots, 1/K)$ and $\boldsymbol{\theta}_1 \approx \ldots \approx \boldsymbol{\theta}_K$ which are likely far from the truth.

There are relatively few satisfactory solutions to this problem (Stephens, 2000b). An ad hoc approach is to simply impose ordering constraints on the parameters (e.g. $p_1 \leq \ldots \leq p_k$) to foster identifiability during the sampling process. Despite their intuitive appeal, ordering constraints have proven largely ineffective because they produce biased posterior samples and significantly slow convergence of the MCMC sampler (West, 1997; Celeux et al., 2000). To avoid hampering convergence, alternate approaches advocate running the MCMC sampler on the unconstrained parameter space and applying a relabeling algorithm to the posterior samples.

Stephens (2000b) defines a relabeling algorithm for clustering inference by means of the Kullback-Leibler divergence which measures the difference between two probability distributions. Mini-

mizing the expected loss from reporting one probability distribution over another yields optimal permutations (or relabelings) $\nu_1, \ldots, \nu_T$ corresponding to posterior samples $\boldsymbol{\Psi}^{(1)}, \ldots, \boldsymbol{\Psi}^{(T)}$ which solve the label-switching problem by placing $\nu_1(\boldsymbol{\Psi}^{(1)}), \ldots, \nu_T(\boldsymbol{\Psi}^{(T)})$ in the same symmetric mode.

Although easy to implement, the algorithm can be demanding on both storage and computation time. Every sweep $t = 1, \ldots, T$ of Steps 3–8 of Algorithm 2 must store not only the sampled $\boldsymbol{\Psi}^{(t)}$ but also the respective classification probabilities $\boldsymbol{S}_1(\boldsymbol{\Psi}^{(t)}), \ldots, \boldsymbol{S}_N(\boldsymbol{\Psi}^{(t)})$ (as defined in Algorithm 2). To circumvent this issue, Stephens (2000b) formulates an "on-line" version of the algorithm following Celeux (1998) that, at iteration t, optimally relabels $\boldsymbol{\Psi}^{(t)}$ with ν_t and discards $\boldsymbol{S}_1(\boldsymbol{\Psi}^{(t)}), \ldots, \boldsymbol{S}_N(\boldsymbol{\Psi}^{(t)})$ before continuing to iteration $t + 1$. With respect to computation time, the most straightforward way to find ν_t in Step 5 is to simply try all possible $K!$ permutations and select the best minimizer, but this method is not feasible for decent-sized K. Thankfully, the problem can be reformulated as an integer programming problem (Stephens, 2000b, Appendix A) for which quicker solutions exist (Taha, 2007).

4 Number of Components K is Unknown

Often it is not known how many clusters K are inherent in the available data. The methods available to handle unknown K can be broadly classified into three categories: hypothesis testing, information criteria, and fully Bayesian estimation. Many of these methods build directly off of the methods available for fixed K in Section 3.

4.1 Hypothesis Testing

The following two methods pose the selection of K in a familiar hypothesis testing framework.

4.1.1 Likelihood Ratio Test

The standard solution to finding the number of clusters in a classical setting is to conduct a series of likelihood ratio tests (LRTs) of the sufficiency of k clusters vs. $k+1$ clusters to adequately group the data. Concretely, for $k = 1, 2, \ldots$, hypotheses $H_0 : K = k$ vs. $H_1 : K = k + 1$, and LRT test statistic λ, we examine $-2 \log \lambda = 2[\log L(\hat{\boldsymbol{\Psi}}_{k+1}|\boldsymbol{X}) - \log L(\hat{\boldsymbol{\Psi}}_k|\boldsymbol{X})]$, where $\hat{\boldsymbol{\Psi}}_{k+1}$ and

$\hat{\boldsymbol{\Psi}}_k$ denote the maximum likelihood estimates of $\boldsymbol{\Psi}_K$ obtained by the EM algorithm under H_1 and H_0 respectively. Large values of $-2\log\lambda$ suggest k clusters are not sufficient to properly model the heterogeneity observed in $\boldsymbol{X}$. Unfortunately, "large values" is not so easy to define because $-2\log\lambda$ violates regularity conditions under the mixture model framework and thus does not retain its usual asymptotic χ^2 distribution (Titterington et al., 1985).

A widely adopted adaptation to this problem is use of a parametric bootstrapping approach to approximate the distribution of $-2\log\lambda$ under H_0 (McLachlan, 1987). The set of values $\{-2\log\lambda^{(1)}, \ldots, -2\log\lambda^{(T)}\}$ obtained by Algorithm 3 approximate the true null distribution of $-2\log\lambda$ and lead to an approximate p-value of $\sum_{t=1}^{T} I(-2\log\lambda < -2\log\lambda^{(t)})/T$, where $I(\cdot)$ denotes the indicator function. If H_0 is rejected, k is incremented by one and the LRT + Bootstrap procedure is repeated; otherwise, k is deemed sufficient to group the observed data and $K = k$ is adopted. Clearly, this method carries with it significant computational complexity, especially when Algorithm 3 is required to run for several hypothesis tests over increasing k. On this note, Smyth (2000) suggests exploiting the independence between bootstrap samples by evaluating Steps 3–6 of Algorithm 3 in parallel.

Algorithm 3: Parametric Bootstrap Sampling for Mixture Models (McLachlan, 1987)

1 **initialize** Set $t \leftarrow 1$. Denote by $\hat{\boldsymbol{\Psi}}_0$ the ML estimate of $\boldsymbol{\Psi}$ under $H_0 : K = k$ and data $\boldsymbol{X}$.

2 **repeat**

3 Generate a bootstrap sample $\boldsymbol{X}^{(t)} = \{\boldsymbol{X}_1^{(t)}, \ldots, \boldsymbol{X}_n^{(t)}\}$ from (1) with $\boldsymbol{\Psi} = \hat{\boldsymbol{\Psi}}_0$.

4 Obtain ML estimates $\hat{\boldsymbol{\Psi}}_k^{(t)}$ and $\hat{\boldsymbol{\Psi}}_{k+1}^{(t)}$ for $\boldsymbol{X}^{(t)}$ via the EM algorithm (Algorithm 1).

5 Compute $-2\log\lambda^{(t)} = 2[\log L(\hat{\boldsymbol{\Psi}}_{k+1}^{(t)}|\boldsymbol{X}^{(t)}) - \log L(\hat{\boldsymbol{\Psi}}_k^{(t)}|\boldsymbol{X}^{(t)})]$

6 Assign $t \leftarrow t + 1$

7 **until** $t > (100/\alpha) - 1$ *for* α*% significance level (McLachlan and Peel, 1997)*

4.1.2 Bayes Factors

From a Bayesian perspective, testing $H_0 : K = k$ vs. $H_1 : K = k + 1$ is accomplished by use of Bayes factors (Kass and Raftery, 1995). Let $B = f(\boldsymbol{X}|K = k + 1)/f(\boldsymbol{X}|K = k)$ denote the

Bayes factor, where the marginal likelihood is given by

$$f(\boldsymbol{X}|K) = \int L(\boldsymbol{\Psi}|\boldsymbol{X}, K)\pi(\boldsymbol{\Psi}|K)d\boldsymbol{\Psi}. \tag{8}$$

Cursory inspection of (2) and (5) suggest evaluating this integral is analytically intractable and must be approximated (see Sections 4.2.2, 4.2.3). Large values of B (roughly $2\log B > 2$) provide evidence in favor of H_1 over H_0 (Raftery, 1996). Lau and Green (2007) question the validity of Bayes factors as a viable Bayesian approach to selecting K, stressing that no informative prior is placed on K (H_0 and H_1 are favored equally *a priori*) and that reliable estimation of $f(\boldsymbol{X}|K)$ is difficult and may be unrepresentative of the posterior surface in high-dimensional settings.

4.2 Information Criteria

Hypothesis testing methods and to-be-discussed fully Bayesian estimation methods can be very computationally demanding, particularly in the machine learning fields of computer vision and pattern recognition (Bouguila and Ziou, 2007). A response to this has been use of information criteria to inform selection of a model. Criteria operate by producing some measure of model quality, promoting models by their balance between maximizing the log-likelihood function and minimizing the number of parameters included. The criteria presented here are diverse, but can be shown to be closely related in the case of binary data (Li, 2006). Use of information criteria is cautioned, however, as they violate regularity conditions under the mixture model framework and therefore do not retain asymptotic optimality (Titterington et al., 1985).

4.2.1 Penalized Likelihood

A straightforward approach fits separate models for a collection of fixed K and selects the "best" model with respect to some penalized likelihood criterion. For many criteria, a "best" model is one that minimizes $C(K) = -2\mathcal{L}_{\max}(K) + 2\psi\eta(K)$ with respect to K, where $\mathcal{L}_{\max}(K) = \log L(\hat{\boldsymbol{\Psi}}_K|\boldsymbol{X})$, $\eta(K)$ is the degrees of freedom of a model with K clusters, and ψ is the penalty term. Criteria for mixture models are plentiful and primarily differ by their choice of ψ (McLach-

lan and Peel, 2000). Notable examples include the application of Akaike's Information Criterion (AIC) (Akaike, 1972) to mixture models by Bozdogan and Sclove (1984) with $\psi = 1$, and Bayesian Information Criterion (BIC) (Schwarz, 1978) with $\psi = \frac{1}{2}\log N$. Another criterion is AIC3 (Bozdogan, 1983) with $\psi = 3/2$. In the specific case of binary data modeled by BMMs, Nadif and Govaert (1998) rigorously compare these three criteria and the Normalized Entropy Criterion (NEC) (Celeux and Soromenho, 1996) on their ability to accurately identify K and find that AIC consistently performs the best of the four.

4.2.2 Coding-Based

A related approach can be found in the coding theory literature, where model selection is reframed as finding the number of clusters "which minimizes the amount of information (measured in bits, if base-2 logarithm is used, or in nits, if natural logarithm is adopted (Wallace et al., 2005)) needed to transmit $\boldsymbol{X}$ efficiently from a sender to a receiver" (Bouguila and Ziou, 2007). This information compression is captured by the expected minimum message length (MML) (Baxter and Oliver, 2000) and is commonly approximated by a variant of Laplace's method (Kass and Raftery, 1995) given by

$$-\log f(\boldsymbol{X}|K) \approx -\log L(\hat{\boldsymbol{\Psi}}) - \log \pi(\hat{\boldsymbol{\Psi}}) + \frac{1}{2}\log|\boldsymbol{I}(\hat{\boldsymbol{\Psi}}, \boldsymbol{X})| - \frac{1}{2}M\log(2\pi) \tag{9}$$

where $f(\boldsymbol{X}|K)$ is as in (8), $\boldsymbol{I}(\hat{\boldsymbol{\Psi}}, \boldsymbol{X})$ is the observed Fisher information matrix, and $M = K(D + 1)$ is the number of parameters to be estimated for a BMM. A "best" model with respect to K is one that minimizes the approximate expected MML (9). Baxter and Oliver (2000) note that $\boldsymbol{I}(\hat{\boldsymbol{\Psi}}, \boldsymbol{X})$ is difficult to compute for the form of the likelihood (2) and recommend substituting it by the complete-data expected Fisher information matrix $\boldsymbol{I}_C(\boldsymbol{\Psi}) = E\left\{\left[\frac{\partial}{\partial \boldsymbol{\Psi}}\log L_C(\boldsymbol{\Psi}, \boldsymbol{X})\right]^2\right\}$ via (3). Similar coding-based criteria include minimum description length (MDL) (Rissanen, 1978) and mixture minimum description length (MMDL) (Figueiredo et al., 1999).

4.2.3 Stochastic Complexity & AutoClass

Approximated in (9) by Laplace's method, $-\log f(\boldsymbol{X}|K)$ is called stochastic complexity (Rissanen, 1987) and measures the degree of information contained in the data $\boldsymbol{X}$ for a model with K components. A competing approximation of stochastic complexity that has been widely used for mixture models is the so-called AutoClass system developed by Cheeseman and Stutz (1996). With conjugate priors selected (Section 3.3.1), their Bayesian classification system substitutes the complete-case likelihood (3) for (2) in (8) to produce $f(\boldsymbol{X}, \boldsymbol{Z}|K)$ given by

$$\int L_C(\boldsymbol{\Psi}_K|\boldsymbol{X}, \boldsymbol{Z})\pi(\boldsymbol{\Psi})d\boldsymbol{\Psi} = \int f(\boldsymbol{Z}|K, \boldsymbol{p})\pi(\boldsymbol{p}|K)d\boldsymbol{p} \int f(\boldsymbol{X}|\boldsymbol{Z}, K, \boldsymbol{\Theta})\pi(\boldsymbol{\Theta}|K)d\boldsymbol{\Theta} \tag{10}$$

$$= f(\boldsymbol{Z}|K)f(\boldsymbol{X}|\boldsymbol{Z}, K) \tag{11}$$

where

$$f(\boldsymbol{Z}|K) = \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_0 + N)} \prod_{k=1}^{K} \frac{\Gamma(\alpha_k + u_k)}{\Gamma(\alpha_k)}, \tag{12}$$

$$f(\boldsymbol{X}|\boldsymbol{Z}, K) = \prod_{k=1}^{K} \left[\prod_{d=1}^{D} \frac{\Gamma(\gamma_{kd} + \delta_{kd})}{\Gamma(\gamma_{kd})\Gamma(\delta_{kd})} \frac{\Gamma(v_{kd} + \gamma_{kd})\Gamma(u_k - v_{kd} + \delta_{kd})}{\Gamma(u_k + \gamma_{kd} + \delta_{kd})} \right]. \tag{13}$$

The Cheeseman-Stutz approximation is therefore $-\log f(\boldsymbol{X}|K) \approx -\log f(\boldsymbol{X}, \boldsymbol{Z}|K)$ and AutoClass selects the "best" model as that which minimizes (11) with respect to K. As it relates to binary data, Gyllenberg et al. (1997) apply AutoClass to binary bacterial taxa data and find it performs exceptionally well in selecting K.

Cheeseman and Stutz (1996) note that AutoClass benefits from a natural form of Occam's razor, as "priors always favor classifications with smaller numbers of classes, and do so overwhelmingly, once the number of classes exceeds some small fraction of the database size." This is a consequence of the way the parameters influence the posterior through a tug-of-war between their multiplicative priors and the marginal likelihood. Specifically, a new cluster k^* under consideration introduces

D new multiplicative priors $\pi(\theta_{k^*1}), \ldots, \pi(\theta_{k^*D})$ to the posterior and increases the dimensionality of $\pi(\boldsymbol{p})$ by 1 due to the inclusion of p_{k^*}. These multiplicative priors always lower the marginal, but the new parameters may raise the marginal. Thus, a cluster is deemed unnecessary if the parameters do not raise the marginal by more than the priors lower the marginal. This behavior favors classifications with smaller numbers of classes and therefore naturally prevents overfitting, a feature lacking in the frequentist approach to clustering where the best partition achieved by maximum likelihood is one in which every observation belongs to its own individual cluster.

4.3 Fully Bayesian Estimation

Several methods presented thus far operate within the Bayesian paradigm (Bayes factors, MML, AutoClass), but are not considered fully Bayesian because they fix K and perform model selection deterministically. Fully Bayesian estimation treats K too as random and supplements (4) with a prior $\pi(K)$ placed on K yielding the posterior distribution

$$\pi(\boldsymbol{\Psi}_K|\boldsymbol{X}, \boldsymbol{Z}) \propto \pi(K)\pi(\boldsymbol{p}|K)f(\boldsymbol{Z}|K, \boldsymbol{p})\pi(\boldsymbol{\Theta}|K)f(\boldsymbol{X}|\boldsymbol{Z}, K, \boldsymbol{\Theta}) \tag{14}$$

from which joint samples are drawn. The prior $\pi(K)$ is typically chosen as a $Poisson(\beta)$ density with mean $\beta = 1$ (Nobile, 2004), $\beta = 3$ (Phillips and Smith, 1996), or even $\beta = 6$ (Stephens, 2000a). Inference on K is made through its marginal posterior probabilities. In the following sections, $K_{\max}$ marks a pre-specified upper bound for K.

4.3.1 Reversible Jump MCMC

Proposed in Green (1995) and developed for mixture models in Richardson and Green (1997), the reversible jump MCMC (RJMCMC) obtains joint posterior samples of $\boldsymbol{Z}$, $\boldsymbol{\Psi}$, and K by appending two trans-dimensional moves to each sweep of the standard mixture model Gibbs sampler (Section 3.3, Algorithm 2): Combine/Split (CS) and Birth/Death (BD). In CS, an attempt is made at random to decrease or increase K by combining two existing components into one or by splitting an existing component into two new components. In BD, an attempt is made to either create

or destroy an empty component (i.e. a component to which no $\boldsymbol{X}_n$ are allocated by $\boldsymbol{Z}$), but is not dicussed here.

Suppose the CS step is initiated and the sampler randomly selects a combination move (provided $K \neq 1$). A pair of components (k_1, k_2) is chosen at random to combine together. In univariate settings, the components must be chosen to be adjacent in their means, but dealing with the adjacency condition is "very annoying" in multivariate settings and is largely ignored (Zhang et al., 2004). A new component k^* with mixing weight p_{k^*} and parameter vector $\boldsymbol{\theta}_{k^*}$ is created by preserving the first two moments of the component distributions. For multivariate Bernoulli, this simply equates to solving p_{k^*} and $\boldsymbol{\theta}_{k^*}$ satisfying

$$p_{k^*} = p_{k_1} + p_{k_2}, \tag{15}$$

$$p_{k^*}\boldsymbol{\theta}_{k^*} = p_{k_1}\boldsymbol{\theta}_{k_1} + p_{k_2}\boldsymbol{\theta}_{k_2}. \tag{16}$$

If a split were instead randomly selected (provided $K \neq K_{\max}$), then a component k^* is selected at random from the K available and split into components k_1 and k_2 with respective mixing weights p_{k_1} and p_{k_2} given by (15), and parameter vectors $\boldsymbol{\theta}_{k_1}$ and $\boldsymbol{\theta}_{k_2}$ given by (16). As it stands, however, the problem is ill-posed; the number of unknowns outweighs the number of equations. Richardson and Green (1997) address this by filling the available degrees of freedom with random draws from a Beta distribution. Hence, we draw b_1 and b_2 from $Beta(2, 2)$ and define

$$p_{k_1} = b_1 p_{k^*} \qquad p_{k_2} = (1 - b_1)p_{k^*} \tag{17}$$

$$\boldsymbol{\theta}_{k_1} = \boldsymbol{\theta}_{k^*} - b_2\sqrt{\boldsymbol{\theta}_{k^*}(1 - \boldsymbol{\theta}_{k^*})\frac{p_{k^*}}{p_{k_1}}} \quad \boldsymbol{\theta}_{k_2} = \boldsymbol{\theta}_{k^*} + b_2\sqrt{\boldsymbol{\theta}_{k^*}(1 - \boldsymbol{\theta}_{k^*})\frac{p_{k^*}}{p_{k_2}}}. \tag{18}$$

The proposed split or combine is accepted with probability $\min\{1, R\}$ and $\min\{1, 1/R\}$ re-

spectively, with R given by

$$R = \frac{\pi(\Delta_K^{'}|\boldsymbol{X})Pr(\Delta_K^{'} \rightarrow \Delta_K)}{\pi(\Delta_K|\boldsymbol{X})f(b_1)f(b_2)Pr(\Delta_K \rightarrow \Delta_K^{'})} \left| \frac{\partial \Delta_K^{'}}{\partial(\Delta_K, b_1, b_2)} \right| \tag{19}$$

where $f(b_i)$ is the $Beta(2,2)$ pdf evaluated at b_i, $i = 1, 2$, Δ_K is the state of $(\boldsymbol{Z}, \boldsymbol{\Psi}_K)$ in the lower dimensional space (pre-split/post-combined), and $\Delta_K^{'}$ is the state of $(\boldsymbol{Z}, \boldsymbol{\Psi}_K)$ in the higher dimensional space (post-split/pre-combined). An inherent difficulty with RJMCMC is developing the jumping moves ($\Delta_K^{'} \rightarrow \Delta_K$ or $\Delta_K \rightarrow \Delta_K^{'}$) between dimensional spaces for updates of K.

4.3.2 Allocation Sampler

Nobile and Fearnside (2007) present a MCMC scheme in the spirit of RJMCMC but having integrated out $\boldsymbol{\Psi}$ from the model so that joint posterior draws are made on $\boldsymbol{Z}$ and K only. This modification successfully avoids the need to invent "good" jumping moves as in RJMCMC. Moreover, it removes direct dependence on the dimensionality of the data D, making the so-called allocation sampler particularly attractive for high-dimensional settings. It does, however, restrict the choice of available priors, requiring priors on $\boldsymbol{p}$ and θ_{kd} such that they may be safely integrated out of (14). For conjugate priors on $\boldsymbol{p}$ and θ_{kd} (Section 3.3.1), the posterior is

$$\pi(K|\boldsymbol{Z}, \boldsymbol{X}) = \int \pi(K, \boldsymbol{\Psi}|\boldsymbol{X}, \boldsymbol{Z}) d\boldsymbol{\Psi} \propto \pi(K) f(\boldsymbol{Z}|K) f(\boldsymbol{X}|\boldsymbol{Z}, K) \tag{20}$$

where $f(\boldsymbol{Z}|K)$ and $f(\boldsymbol{X}|\boldsymbol{Z}, K)$ are given by (12) and (13) respectively. The sampling scheme is outlined in Algorithm 4 (page 20).

One type of move (AE) is responsible for changing the number of components K as well as allocation $\boldsymbol{Z}$. AE (Absorb/Eject) is very similar to the reversible jump moves of Richardson and Green (1997). Ejection (split) or absorption (combination) is randomly selected (provided $K \neq K_{\max}$ or $K \neq 1$, respectively) and accepted with respective probabilities $\min\{1, R\}$ or

$\min\{1, 1/R\}$, where

$$R = \frac{\pi(K', \boldsymbol{Z}'|\boldsymbol{X})}{\pi(K, \boldsymbol{Z}|\boldsymbol{X})} \cdot \frac{Pr(\{K', \boldsymbol{Z}'\} \to \{K, \boldsymbol{Z}\})}{Pr(\{K, \boldsymbol{Z}\} \to \{K', \boldsymbol{Z}'\})}. \tag{21}$$

and additional terms as defined in Algorithm 4.

The remaining moves (GS, M1, M2, M3) detail different ways one might re-allocate $\boldsymbol{X}_n$ in $\boldsymbol{X}$ between components. GS is as given in Algorithm 4. M1, M2, and M3, begin by randomly selecting two distinct components k_1, k_2 from $\{1, \ldots, K\}$. The observations currently allocated to these two components are then re-allocated in one of three ways to produce a new allocation $\boldsymbol{Z}'$. M1 re-allocates individual observations with constant probability, with b_{k_1} drawn from $Beta(\alpha_{k_1}, \alpha_{k_2})$ and every $\boldsymbol{X}_n$ in components k_1 and k_2 re-allocated to component k_1 with probability b_{k_1} and to component k_2 with probability $1 - b_{k_1}$. M2 re-allocates grouped observations. If component k_1 is not empty ($u_{k_1} > 0$), m is drawn randomly from $\{1, \ldots, u_{k_1}\}$ and m observations are randomly selected from component k_1 and re-allocated to component k_2. Finally, M3 re-allocates individual observations sequentially with observation-dependent probabilities. Its implementation is more complex than the former two moves and its details left to Appendix A.2 of Nobile and Fearnside (2007). A move is accepted with probability $\min\{1, R\}$ with

$$R = \frac{\pi(K, \boldsymbol{Z}'|\boldsymbol{X})}{\pi(K, \boldsymbol{Z}|\boldsymbol{X})} \cdot \left[\frac{Pr(\boldsymbol{Z}' \to \boldsymbol{Z})}{Pr(\boldsymbol{Z} \to \boldsymbol{Z}')}\right]_j \tag{22}$$

where $\boldsymbol{Z}'$ marks the proposed reallocation and $j = 1, 2, 3$ indicate the form of $Pr(\boldsymbol{Z}' \to \boldsymbol{Z})/Pr(\boldsymbol{Z} \to \boldsymbol{Z}')$ taken for moves M1, M2, or M3, respectively.

5 Feature Weighting for High-Dimensional Data

For binary data, the components of $\boldsymbol{X}_n = (X_{n1}, \ldots, X_{nD})$ represent the presence (1) or absence (0) of features $d = 1, \ldots, D$. Many irrelevant features may harm clustering by introducing excessive noise. The methods detailed in Sections 3 & 4 do not account for this and treat all

features as informative. Inspired by feature weighting for Gaussian mixture models (Law et al., 2004), Wang and Kabán (2005) address feature weighting for binary data by modifying the BMM to account for features that are deemed uninformative. A feature d is uninformative if $\theta_{1d}, \ldots, \theta_{Kd}$ do not tend to vary and, to reduce unnecessary noise, are better represented by a single parameter λ_d (Novovičová et al., 1996). The likelihood (2) is then repurposed as

$$L(\boldsymbol{\Psi}_K, \boldsymbol{\lambda}, \rho_1, \ldots, \rho_D | \boldsymbol{X}) = \prod_{n=1}^{N} \sum_{k=1}^{K} p_k \prod_{d=1}^{D} \left[\rho_d \theta_{kd}^{X_{nd}} (1-\theta_{kd})^{1-X_{nd}} + (1-\rho_d) \lambda_d^{X_{nd}} (1-\lambda_d)^{1-X_{nd}} \right], \tag{23}$$

where $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_D)$ and $\rho_d = Pr(\phi_d = 1)$ is the feature saliency of feature d with latent variables $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_D)$ marking the relevant features ($\phi_d = 1$ if feature d is relevant, 0 otherwise). As developed in Bouguila (2010), the complete-case likelihood (3) is modified to include the latent $\boldsymbol{\phi}$ so that $L_C(\boldsymbol{\Psi}_K, \boldsymbol{\lambda}, \rho_1, \ldots, \rho_D | \boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{\phi}) =$

$$\prod_{n=1}^{N} \prod_{k=1}^{K} \left\{ p_k \prod_{d=1}^{D} \left[\rho_d \theta_{kd}^{X_{nd}} (1-\theta_{kd})^{1-X_{nd}} \right]^{\phi_d} \left[(1-\rho_d) \lambda_d^{X_{nd}} (1-\lambda_d)^{1-X_{nd}} \right]^{1-\phi_d} \right\}^{Z_{nk}}. \tag{24}$$

Feature selection is shown to improve clustering of binary data when K is taken as known (Wang and Kabán, 2005). Placing conjugate Beta priors on $\rho_1, \ldots, \rho_D, \lambda_1, \ldots, \lambda_D$, Bouguila (2010) uses (24), the Bayesian extension to the EM algorithm, and AutoClass to cluster high-dimensional binary data with feature weighting when K is unknown, finding that downweighting uninformative features can help better identify K. Most recently, Elguebaly and Bouguila (2013) incorporate feature weighting into clustering under a fully Bayesian approach, albeit for continuous data.

6 Discussion

Bernoulli mixture models (BMMs) are immensely useful for clustering binary data when the number of clusters is or is not known. Classical approaches make use of maximum likelihood methods to estimate the parameters of the BMM and capture the most-likely partition of $\boldsymbol{X}$.

Bayesian approaches arrive at comparable partitions and provide an added flexibility in their ability to incorporate prior information and prevent overfitting, but face computational hurdles due to the label-switching problem and potentially slow MCMC convergence.

Despite these issues, recent literature in clustering with mixture models has focused almost exclusively on Bayesian methods for their generality and power. A Bayesian method semi-related to those presented here is Bayesian Hierarchical Clustering (BHC) (Heller and Ghahramani, 2005; Heard et al., 2006). BHC selects a "best" partition $\boldsymbol{Z}$ of $\boldsymbol{X}$ as one that maximizes the posterior distribution and is chosen from partitions developed by starting with N individual clusters and sequentially merging the most alike observations together until only one cluster is produced. However, for BHC and the other Bayesian methods discussed up to this point which estimate $\boldsymbol{Z}$ via its posterior mode, it is stressed that "the maximum a posteriori partition has no objective status as a best estimate of the clustering of the data" as it is never made explicit what constitutes a "best" partition (Lau and Green, 2007). To avoid this criticism, Lau and Green (2007) develop a Bayesian method that seeks a partition of the data that minimizes the expected pairwise coincidence loss function responsible for penalizing partitions that group unlike observations together. Though particularly computationally demanding, the method it is guaranteed to arrive at an optimal partition $\boldsymbol{Z}$ of the data $\boldsymbol{X}$.

Of course, while all these methods have their own pros and cons, as do those described in Sections 3 & 4, the increasing prevalance of high-dimensional binary data suggests the most important direction for clustering procedures involves the incorporation of feature weighting (Section 5). Only very recent attempts have placed feature weighting in a fully Bayesian estimation setting (Elguebaly and Bouguila, 2013), but none so far in the particular case of binary data. The marriage of fully Bayesian estimation and feature weighting may prove extremely effective in clustering binary data.

Algorithm 4: Allocation Sampler (Nobile and Fearnside, 2007)

initialize Set $t \leftarrow 1$. Choose starting values $K^{(0)}$ and $\boldsymbol{Z}^{(0)} = (\boldsymbol{Z}_1^{(0)}, \ldots, \boldsymbol{Z}_N^{(0)})$.

repeat

  Assign $K^{(t)} \leftarrow K^{(t-1)}$ and $\boldsymbol{Z}^{(t)} \leftarrow \boldsymbol{Z}^{(t-1)}$

  Randomly select move from $\{AE, GS, M1, M2, M3\}$

  if *AE* **then** attempt to change $K^{(t)}$ with absorption/ejection Metropolis-Hastings

    Choose a component k at random to eject from / absorb from.

    if *ejection* **then**

      Create a $(K^{(t)} + 1)$-th component k^* and let $K^{'} = K^{(t)} + 1$

      Obtain $\boldsymbol{Z}^{'}$ by reassigning vectors from k to k^* based on draw from $Beta(a, a)$

      Assign $K^{(t)} \leftarrow K^{'}$ and $\boldsymbol{Z}^{(t)} \leftarrow \boldsymbol{Z}^{'}$ with probability $\min\{1, R\}$, R as in (21)

    else $absorption$

      Choose another component k^* to absorb into and let $K = K^{(t)} - 1$

      Obtain re-allocations $\boldsymbol{Z}$ by reassigning all vectors in k to k^*

      Assign $K^{(t)} \leftarrow K$ and $\boldsymbol{Z}^{(t)} \leftarrow \boldsymbol{Z}$ with probability $\min\{1, 1/R\}$, R as in (21)

  else do not attempt to change $K^{(t)}$

    if *GS* **then** change $\boldsymbol{Z}_n^{(t)}$ in $\boldsymbol{Z}^{(t)}$ one-at-a-time with systematic sweep Gibbs sampler

      for $n = 1, \ldots, N$ **do**

        Generate $\boldsymbol{Z}_n^{'}$ from its full conditional distribution and assign $\boldsymbol{Z}_n^{(t)} \leftarrow \boldsymbol{Z}_n^{'}$

    else change $\boldsymbol{Z}_n^{(t)}$ in $\boldsymbol{Z}^{(t)}$ simultaneously with Metropolis-Hastings M1, M2, or M3

      Randomly select two distinct k_1, k_2 from $\{1, \ldots, K^{(t)}\}$

      Perform M1, M2, or M3 (whichever was chosen) to obtain new allocation $\boldsymbol{Z}'$

      Assign $\boldsymbol{Z}^{(t)} \leftarrow \boldsymbol{Z}^{'}$ with probability $\min\{1, R\}$, R as in (22)

  Assign $t \leftarrow t + 1$

until *t is suitably large*

References

Abel, L., Golmard, J.-L., and Mallet, A. (1993), "An autologistic model for the genetic analysis of familial binary data." *American journal of human genetics*, 53, 894.

Agrawal, R., Srikant, R., et al. (1994), "Fast algorithms for mining association rules," in *Proc. 20th int. conf. very large data bases, VLDB*, vol. 1215, pp. 487–499.

Akaike, H. (1972), *Information theory and an extension of the Maximum Likelihood Principle*, Budapest: Akademiai Kiado.

Baxter, R. A. and Oliver, J. J. (2000), "Finding overlapping components with MML," *Statistics and Computing*, 10, 5–16.

Bishop, C. (2006), *Pattern Recognition and Machine Learning*, Information Science and Statistics, Springer.

Böhning, D. and Seidel, W. (2003), "Editorial: recent developments in mixture models," *Computational Statistics & Data Analysis*, 41, 349 – 357, recent Developments in Mixture Model. Recent Developments in Mixture Model.

Bouguila, N. (2010), "On multivariate binary data clustering and feature weighting," *Computational Statistics & Data Analysis*, 54, 120–134.

Bouguila, N. and Ziou, D. (2007), "High-Dimensional Unsupervised Selection and Estimation of a Finite Generalized Dirichlet Mixture Model Based on Minimum Message Length," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 29, 1716–1731.

Bozdogan, H. (1983), "Determining the Number of Component Clusters in the Standard Multivariate Normal Mixture Model Using Model-Selection Criteria." Tech. rep., DTIC Document.

Bozdogan, H. and Sclove, S. (1984), "Multi-sample cluster analysis using Akaike's Information Criterion," *Annals of the Institute of Statistical Mathematics*, 36, 163–180.

Celeux, G. (1998), "Bayesian Inference for Mixture: The Label Switching Problem," in *COMPSTAT*, eds. Payne, R. and Green, P., Physica-Verlag HD, pp. 227–232.

Celeux, G. and Govaert, G. (1991), "Clustering criteria for discrete data and latent class models," *Journal of Classification*, 8, 157–176.

— (1992), "A classification EM algorithm for clustering and two stochastic versions," *Computational statistics & Data analysis*, 14, 315–332.

Celeux, G., Hurn, M., and Robert, C. P. (2000), "Computational and Inferential Difficulties with Mixture Posterior Distributions," *Journal of the American Statistical Association*, 95, pp. 957–970.

Celeux, G. and Soromenho, G. (1996), "An entropy criterion for assessing the number of clusters in a mixture model," *Journal of Classification*, 13, 195–212.

Cheeseman, P. and Stutz, J. (1996), "Advances in Knowledge Discovery and Data Mining," Menlo Park, CA, USA: American Association for Artificial Intelligence, chap. Bayesian Classification (AutoClass): Theory and Results, pp. 153–180.

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. Series B*, 39, 1–38.

Diebolt, J. and Robert, C. P. (1994), "Estimation of Finite Mixture Distributions through Bayesian Sampling," *Journal of the Royal Statistical Society. Series B (Methodological)*, 56, pp. 363–375.

Elguebaly, T. and Bouguila, N. (2013), "Simultaneous Bayesian clustering and feature selection using RJMCMC-based learning of finite generalized Dirichlet mixture models," *Signal Processing*, 93, 1531 – 1546, special issue on Machine Learning in Intelligent Image Processing. Special issue on Machine Learning in Intelligent Image Processing.

Figueiredo, M. A., Leitão, J. M., and Jain, A. K. (1999), "On fitting mixture models," in *Energy minimization methods in computer vision and pattern recognition*, Springer, pp. 54–69.

Fränti, P., Xu, M., and Kärkkäinen, I. (2003), "Classification of binary vectors by using ΔSC distance to minimize stochastic complexity," *Pattern Recognition Letters*, 24, 65 – 73.

Gelfand, A. E. and Smith, A. F. (1990), "Sampling-based approaches to calculating marginal densities," *Journal of the American statistical association*, 85, 398–409.

Govaert, G. and Nadif, M. (1996), "Comparison of the mixture and the classification maximum likelihood in cluster analysis with binary data," *Computational Statistics & Data Analysis*, 23, 65 – 81, classification. Classification.

Green, P. J. (1995), "Reversible jump Markov chain Monte Carlo computation and Bayesian model determination," *Biometrika*, 82, 711–732.

Grim, J., Pudil, P., and Somol, P. (2000), "Multivariate structural Bernoulli mixtures for recognition of handwritten numerals," in *Pattern Recognition, 2000. Proceedings. 15th International Conference on*, vol. 2, pp. 585–589 vol.2.

Gyllenberg, M., Koski, T., and Verlaan, M. (1997), "Classification of Binary Vectors by Stochastic Complexity," *Journal of Multivariate Analysis*, 63, 47 – 72.

Heard, N. A., Holmes, C. C., and Stephens, D. A. (2006), "A Quantitative Study of Gene Regulation Involved in the Immune Response of Anopheline Mosquitoes: An Application of Bayesian Hierarchical Clustering of Curves," *Journal of the American Statistical Association*, 101, pp. 18–29.

Heller, K. A. and Ghahramani, Z. (2005), "Bayesian Hierarchical Clustering," *Twenty-second International Conference on Machine Learning (ICML-2005)*.

Jain, A. K. and Dubes, R. C. (1988), *Algorithms for clustering data*, Prentice-Hall, Inc.

Kamthe, A., Carreira-Perpinán, M. A., and Cerpa, A. (2011), "Adaptation of a mixture of multivariate Bernoulli distributions," in *IJCAI Proceedings-International Joint Conference on Artificial Intelligence*, vol. 22, p. 1336.

Kass, R. E. and Raftery, A. E. (1995), "Bayes factors," *Journal of the american statistical association*, 90, 773–795.

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

Law, M., Figueiredo, M., and Jain, A. (2004), "Simultaneous feature selection and clustering using mixture models," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 26, 1154–1166.

Li, T. (2006), "A Unified View on Clustering Binary Data," *Machine Learning*, 62, 199–215.

Marin, J.-M., Mengersen, K., and Robert, C. P. (2005), "Bayesian modelling and inference on mixtures of distributions," *Handbook of statistics*, 25, 459–507.

Marriott, F. (1974), *The interpretation of multiple observations*, Academic Press.

McLachlan, G. and Krishnan, T. (1997), *The EM Algorithm and Extensions*, John Wiley & Sons, New York.

McLachlan, G. and Peel, D. (2000), *Finite Mixture Models, Willey Series in Probability and Statistics*, John Wiley & Sons, New York.

McLachlan, G. J. (1987), "On Bootstrapping the Likelihood Ratio Test Stastistic for the Number of Components in a Normal Mixture," *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 36, pp. 318–324.

McLachlan, G. J. and Peel, D. (1997), "On a Resampling Approach to Choosing the Number of Components in Normal Mixture Models," in *Proceedings of Interface 96, 28th Symposium on the Interface*, pp. 260–266.

Nadif, M. and Govaert, G. (1998), "Clustering for binary data and mixture models – choice of the model," *Applied Stochastic Models and Data Analysis*, 13, 269–278.

Nobile, A. (2004), "On the posterior distribution of the number of components in a finite mixture," *The Annals of Statistics*, 32, 2044–2073.

Nobile, A. and Fearnside, A. (2007), "Bayesian finite mixtures with an unknown number of components: The allocation sampler," *Statistics and Computing*, 17, 147–162.

Novovičová, J., Pudil, P., and Kittler, J. (1996), "Divergence based feature selection for multimodal class densities," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 18, 218–223.

Pearson, K. (1894), "Contributions to the theory of mathematical evolution," *Philosophical Transactions of the Royal Society of London A*, 185, pp. 71–110.

Phillips, D. B. and Smith, A. F. (1996), "Bayesian model comparison via jump diffusions," in *Markov chain Monte Carlo in practice*, Springer, pp. 215–239.

Raftery, A. E. (1996), "Hypothesis testing and model selection," in *Markov chain Monte Carlo in practice*, Springer US, pp. 163–187.

Redner, R. A. and Walker, H. F. (1984), "Mixture Densities, Maximum Likelihood and the EM Algorithm," *SIAM Review*, 26, pp. 195–239.

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

Ripley, B. D. (1996), *Pattern recognition and neural networks*, Cambridge university press.

Rissanen, J. (1978), "Modeling by shortest data description," *Automatica*, 14, 465–471.

— (1987), "Stochastic Complexity," *Journal of the Royal Statistical Society. Series B (Methodological)*, 49, pp. 223–239.

Schwarz, G. (1978), "Estimating the Dimension of a Model," *The Annals of Statistics*, 6, 461–464.

Smyth, P. (2000), "Model selection for probabilistic clustering using cross-validated likelihood," *Statistics and Computing*, 10, 63–72.

Stephens, M. (2000a), "Bayesian analysis of mixture models with an unknown number of components – an alternative to reversible jump methods," *The Annals of Statistics*, 28, 40–74.

— (2000b), "Dealing with label switching in mixture models," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62, 795–809.

Symons, M. J. (1981), "Clustering criteria and multivariate normal mixtures," *Biometrics*, 35–43.

Taha, H. A. (2007), *Operations research: an introduction*, Pearson/Prentice Hall.

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.

Titterington, D., Smith, A., and Makov, U. (1985), *Statistical analysis of finite mixture distributions*, Wiley series in probability and mathematical statistics: Applied probability and statistics, Wiley.

Wallace, C. S. et al. (2005), *Statistical and inductive inference by minimum message length*, Springer.

Wang, X. and Kabán, A. (2005), "Finding Uninformative Features in Binary Data," in *Intelligent Data Engineering and Automated Learning - IDEAL 2005*, eds. Gallagher, M., Hogan, J., and Maire, F., Springer Berlin Heidelberg, vol. 3578 of *Lecture Notes in Computer Science*, pp. 40–47.

West, M. (1997), "Hierarchical Mixture Models in Neurological Transmission Analysis," *Journal of the American Statistical Association*, 92, pp. 587–606.

Zhang, Z., Chan, K., Wu, Y., and Chen, C. (2004), "Learning a multivariate Gaussian mixture model with the reversible jump MCMC algorithm," *Statistics and Computing*, 14, 343–355.