Received: 23 March 2021 | Revised: 14 May 2021 | Accepted: 17 May 2021

DOI: 10.1002/sam.11528

WILEY

RESEARCH ARTICLE

# Modeling and inference for mixtures of simple symmetric exponential families of $p$-dimensional distributions for vectors with binary coordinates

**Abhishek Chakraborty**[1] | **Stephen B. Vardeman**[2,3]

[1]Department of Mathematics, Statistics, and Computer Science, Lawrence University, Appleton, Wisconsin

[2]Department of Statistics, Iowa State University, Ames, Iowa, USA

[3]Department of Industrial and Manufacturing Systems Engineering, Iowa State University, Ames, Iowa, USA

**Correspondence**
Abhishek Chakraborty, 711 E. Boldt Way, SPC 24, Appleton, WI 54911, USA.
Email: abhishek.chakraborty@lawrence.edu

**Abstract**

We propose tractable symmetric exponential families of distributions for multivariate vectors of 0's and 1's in $p$ dimensions, or what are referred to in this paper as binary vectors, that allow for nontrivial amounts of variation around some central value $\boldsymbol{\mu} \in \{0,1\}^p$. We note that more or less standard asymptotics provides likelihood-based inference in the one-sample problem. We then consider mixture models where component distributions are of this form. Bayes analysis based on Dirichlet processes and Jeffreys priors for the exponential family parameters prove tractable and informative in problems where relevant distributions for a vector of binary variables are clearly not symmetric. We also extend our proposed Bayesian mixture model analysis to datasets with missing entries. Performance is illustrated through simulation studies and application to real datasets.

**KEYWORDS**

Bayesian analysis, mixture models, MCMC, pixel flips, missing entries

## 1 | INTRODUCTION

This paper concerns practical modeling and inference for multivariate binary data. The term "binary" refers to the two possible values that each of the single-variate responses can assume, whether naturally occurring in two categories (e.g., presence/absence, success/failure, on/off) or formed by dichotomizing a continuous response (e.g., whether or not systolic blood pressure is <140). Multivariate binary responses are common in many fields, including biological and social sciences. For example, diseases are often diagnosed on the basis of binary data on multiple symptoms. In psychological and educational testing, subjects are often required to give "yes" or "no" answers to multiple questions. Even in engineering fields, conclusions concerning the reliability of complex systems are often based on observation of whether various components are functioning or not. In image processing, objects are often identified based on variables such as area, shape, and so on where data arise from the use of threshold values.

We undertake the task of modeling multivariate vectors of 0's and 1's in $p$ dimensions by first identifying a tractable symmetric family that allows for nontrivial amounts of variation around some central value. We use the fact that for any two $p$-dimensional binary vectors the

squared Euclidean distance between them is given by the number of "pixel flips" between the two vectors. The exponential families of distributions we consider are symmetric in terms of the number of "pixel flips" away from a central binary vector and are characterized by a parameter that controls the amount of variability around the central value allowed by the distribution. We then consider mixture models with such components. Bayesian inference for data modeled by such mixture distributions is then performed via MCMC sampling.

The rest of the paper is arranged as follows. Section 2 describes an exponential family of symmetric distributions on multivariate binary vectors and likelihood-based inference for it in a one-sample model. In Section 3 we present the mixture model and our Bayes analysis for it. Results from simulation studies made to evaluate the performance of Bayes inference for the mixture model are presented in Section 4. We apply the Bayesian mixture model to two real datasets in Section 5. An extension of the Bayes modeling approach to handle incomplete datasets is described in Section 6 and applied to another real problem followed by conclusions in Section 7.

## 2 | A SYMMETRIC EXPONENTIAL FAMILY OF DISTRIBUTIONS ON BINARY VECTORS

Suppose $\mathbf{X} \in \{0,1\}^p$ is a random vector where $p$ is a positive integer. In what follows, realizations of $\mathbf{X}$ will be denoted as $\mathbf{x}$. The number of possible realizations of $\mathbf{X}$ is $2^p$. We first consider a tractable symmetric distribution for the $p$-dimensional binary vector $\mathbf{X}$.

Note that for any two $p$-dimensional binary vectors $\mathbf{x}$ and $\mathbf{z}$ the squared Euclidean distance between them is the number of "coordinate or pixel flips" between the two vectors given by

$$\|\mathbf{x}-\mathbf{z}\|^2 = \sum_{j=1}^{p} 1(x_j \neq z_j),$$

where $1(x_j \neq z_j)$ is the indicator variable that takes the value 1 if the $j$th coordinates of $\mathbf{x}$ and $\mathbf{z}$ do not agree and is 0 otherwise. In information theory, the squared Euclidean distance defined above is known as the Hamming distance. Hereafter, we use the term "pixel flips" and the variable $m$ to denote the squared Euclidean (Hamming) distance between two $p$-dimensional binary vectors.

Let $q(m)$ be a probability mass function (pmf) defined on the set of all possible pixel flip counts between two $p$-dimensional binary vectors, that is, the set $\{0,1,\ldots,p\}$. Then, a sensible symmetric distribution on $\{0,1\}^p$ is

$$f(\mathbf{x}|\boldsymbol{\mu}) = \begin{cases} \dfrac{q(\|\mathbf{x}-\boldsymbol{\mu}\|^2)}{\binom{p}{\|\mathbf{x}-\boldsymbol{\mu}\|^2}}, & \text{for } \mathbf{x} \in \{0,1\}^p \\ 0, & \text{otherwise} \end{cases}. \tag{1}$$

The parameter $\boldsymbol{\mu} \in \{0,1\}^p$ is the central data pattern in form (1). We consider cases where the pmf $q(m)$ is parameterized with a parameter $\alpha$ that can be used to control the amount of variability around $\boldsymbol{\mu}$ represented by $f(\mathbf{x}|\boldsymbol{\mu})$. A particularly tractable version of the basic model (1) is

$$f(\mathbf{x}|\boldsymbol{\mu},\alpha) = \begin{cases} \dfrac{c(\alpha)\cdot\alpha^{(\|\mathbf{x}-\boldsymbol{\mu}\|^2)}}{\binom{p}{\|\mathbf{x}-\boldsymbol{\mu}\|^2}}, & \text{for } \mathbf{x} \in \{0,1\}^p \\ 0, & \text{otherwise} \end{cases}, \tag{2}$$

where $\boldsymbol{\mu} \in \{0,1\}^p$, $\alpha \in (0,1)$, and $c(\alpha)$ is the normalizing constant (required for $f(\cdot)$ to be a pmf). Here, the total probability for all vectors $\mathbf{x} \in \{0,1\}^p$ that are $m$ pixel flips away from $\boldsymbol{\mu}$ decreases geometrically in the number of pixel flips. That is

$$q(m|\alpha) \propto \alpha^m \text{ for } m = 0,1,\ldots,p. \tag{3}$$

The parameter $\alpha$ in form (2) is directly interpretable as a distribution spread parameter. The constant $c(\alpha)$ in form (2) (the constant of proportionality in display (3)), is $(1-\alpha)/(1-\alpha^{p+1})$. Smaller values of $\alpha$ produce realizations $\mathbf{x}$ typically more similar to $\boldsymbol{\mu}$ than do bigger values.

A tractable possible alternative to model (2) is

$$f(\mathbf{x}|\boldsymbol{\mu},\alpha) = \begin{cases} \dfrac{\alpha^{(\|\mathbf{x}-\boldsymbol{\mu}\|^2)}}{(1+\alpha)^p}, & \text{for} \mathbf{x} \in \{0,1\}^p \\ 0, & \text{otherwise} \end{cases} \tag{4}$$

for $\boldsymbol{\mu} \in \{0,1\}^p$, and $\alpha \in (0,1)$. Here, the probabilities for vectors $\mathbf{x} \in \{0,1\}^p$ decrease geometrically in the number of pixel flips that $\mathbf{x}$ is away from $\boldsymbol{\mu}$, that is, in terms of the pixel flip count

$$q(m|\alpha) \propto \binom{p}{m}\alpha^m \text{ for } m = 0,1,\ldots,p,$$

where the constant of proportionality is $(1+\alpha)^{-p}$. Note that the total probability for those outcomes $m$ pixel flips away from $\boldsymbol{\mu}$ is $\binom{p}{m}\alpha^m(1+\alpha)^{-p}$. We have found form (2) to be more useful than this second exponential family form (4), so in the following sections we will work with form (2),

and correspondingly display (3). Thus, henceforth $c(\alpha) = (1-\alpha)/(1-\alpha^{p+1})$.

## 2.1 | Properties of the model

We begin by considering a single $p$-dimensional binary random vector $\mathbf{X}$ whose pmf is given in display (2). For a fixed $\boldsymbol{\mu}$, the random pixel flip count, $M$, with possible values in the set $\{0, 1, \ldots, p\}$, has mean

$$E(M) = \frac{\alpha[1-(p+1)\alpha^p + p\alpha^{p+1}]}{(1-\alpha^{p+1})(1-\alpha)}. \tag{5}$$

Furthermore, the Fisher information about the parameter $\alpha$ contained in a single observation $\mathbf{X}$, or equivalently $M$ (for a fixed $\boldsymbol{\mu}$), is

$$\begin{aligned}\mathcal{I}(\alpha) &= -E\left[\frac{d^2}{d\alpha^2}\log\ q(M|\alpha)\right] \\ &= \frac{[1-(p+1)^2\alpha^p + 2p(p+2)\alpha^{p+1} - (p+1)^2\alpha^{p+2} + \alpha^{2p+2}]}{\alpha(1-\alpha^{p+1})^2(1-\alpha)^2},\end{aligned} \tag{6}$$

where $\log(\cdot)$ denotes the natural logarithm and $q(M|\alpha)$ is as in display (3).

## 2.2 | Likelihood-based inference in a one-sample problem

Now consider an independent and identically distributed (iid) training dataset $\{\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_N\}$ where each $\mathbf{X}_i$ is a $p$-dimensional binary random vector. Following from display (2), the joint distribution of the observed vectors is specified by the joint pmf

$$f(\mathbf{x}_1, \ldots, \mathbf{x}_N|\boldsymbol{\mu}, \alpha) = \prod_{i=1}^{N} \frac{c(\alpha)\cdot\alpha^{(\|\mathbf{x}_i-\boldsymbol{\mu}\|^2)}}{\binom{p}{\|\mathbf{x}_i-\boldsymbol{\mu}\|^2}}. \tag{7}$$

An expression equivalent to form (7) in terms of the counts of possible observation vectors $\{\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_{2^p}\}$ (where $\mathbf{z}_j$ denotes the $j$th such binary vector in some ordering of them) is

$$g_N(n_1, n_2, \ldots, n_{2^p}) = \prod_{j=1}^{2^p}\left(\frac{c(\alpha)\cdot\alpha^{(\|\mathbf{z}_j-\boldsymbol{\mu}\|^2)}}{\binom{p}{\|\mathbf{z}_j-\boldsymbol{\mu}\|^2}}\right)^{n_j}, \tag{8}$$

where $n_j$ denotes the number of observations in the training dataset that are identical to $\mathbf{z}_j$ (that is, for $j = 1, \ldots, 2^p$, $n_j = |\mathbf{x}_i : \mathbf{x}_i = \mathbf{z}_j;\ i = 1, 2, \ldots, N|$). From form (7), the one-sample model log-likelihood is

$$\begin{aligned}l(\boldsymbol{\mu}, \alpha) &= \log f(\mathbf{x}_1, \ldots, \mathbf{x}_N|\boldsymbol{\mu}, \alpha) \\ &= N\log(1-\alpha) - N\log(1-\alpha^{p+1}) \\ &\quad + \left(\sum_{i=1}^{N}\|\mathbf{x}_i-\boldsymbol{\mu}\|^2\right)\log\alpha \\ &\quad - \sum_{i=1}^{N}\log\binom{p}{\|\mathbf{x}_i-\boldsymbol{\mu}\|^2}.\end{aligned} \tag{9}$$

We are interested in finding estimators for $\boldsymbol{\mu} \in \{0,1\}^p$ and $\alpha \in (0,1)$. First, a sensible estimator for $\boldsymbol{\mu}$ is an observation vector in the training set with the highest frequency, that is, a (arbitrarily chosen in the case of ties) mode of the relative frequency distribution. That is, we consider

$$\widehat{\boldsymbol{\mu}} = z_{j^*} \quad \text{where} \quad j^* = \arg\max\{n_1, n_2, \ldots, n_{2^p}\}. \tag{10}$$

To develop an estimator for the variability parameter $\alpha$, temporarily fix the central pattern $\boldsymbol{\mu}$ in form (9). Let $m_i$ be the number of pixel flips that $\mathbf{x}_i$ is away from $\boldsymbol{\mu}$. The joint distribution of $\{M_1, M_2, \ldots, M_N\}$ characterized by the parameter $\alpha$ has pmf

$$q(m_1, \ldots, m_N|\alpha) = c(\alpha)^N\alpha^{\sum_{i=1}^{N}m_i}. \tag{11}$$

Both the log-likelihood and the method of moments equations (for $\alpha$ with $\boldsymbol{\mu}$ fixed) are

$$\begin{aligned}&\left(Np - \sum m_i\right)\alpha^{p+2} + \left(\sum m_i - Np - N\right)\alpha^{p+1} \\ &\quad + \left(N + \sum m_i\right)\alpha - \sum m_i = 0.\end{aligned} \tag{12}$$

Solution of this polynomial (in $\alpha$) equation for $p$ of even moderate size is not obvious. But, a plausible estimator for $\alpha$ can be based on the ratio of the sum of relative frequencies of observations one pixel flip away from $\boldsymbol{\mu}$ to those which are identical to $\boldsymbol{\mu}$, that is

$$\widehat{\alpha}(\boldsymbol{\mu}) = \frac{\sum_{j:\|\mathbf{z}_j-\boldsymbol{\mu}\|^2=1} n_j}{\sum_{j:\|\mathbf{z}_j-\boldsymbol{\mu}\|^2=0} n_j} = \frac{|i\ :\ m_i = 1|}{|i\ :\ m_i = 0|}. \tag{13}$$

Note that in the first fraction in display (13) the denominator includes a single term and the numerator is a sum over at most $p$ nonzero terms for a choice of $\boldsymbol{\mu}$. Then in light of forms (10) and (13), one (crude) estimator for $\alpha$ is $\widehat{\alpha}(\widehat{\boldsymbol{\mu}})$.

From the Law of Large Numbers applied to the relative frequencies of various $\mathbf{x} \in \{0,1\}^p$, it follows that the estimator for $\boldsymbol{\mu}$ defined in form (10) is consistent in probability. That is, if $\{\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_N\}$ are iid observations from the

distribution specified by form (2) and $\widehat{\boldsymbol{\mu}}^N$ is as in display (10) (based on $N$ observations), then

$$\widehat{\boldsymbol{\mu}}^N \xrightarrow{\text{prob}} \boldsymbol{\mu} \quad \text{as} \quad N \to \infty \tag{14}$$

and in fact

$$P[\widehat{\boldsymbol{\mu}}^N \neq \boldsymbol{\mu}] \to 0 \quad \text{as} \quad N \to \infty. \tag{15}$$

It then follows that

$$P[\widehat{\alpha}^N(\widehat{\boldsymbol{\mu}}^N) \neq \widehat{\alpha}^N(\boldsymbol{\mu})] \to 0 \quad \text{as} \quad N \to \infty, \tag{16}$$

where $\widehat{\alpha}^N(\widehat{\boldsymbol{\mu}}^N)$ and $\widehat{\alpha}^N(\boldsymbol{\mu})$ are based on $N$ observations and constructed according to form (13). It follows from the Law of Large Numbers applied to the observed relative frequencies of observations of outcomes at $\boldsymbol{\mu}$ and one pixel flip away from it, and the continuous mapping theorem that the (unrealizable) estimator $\widehat{\alpha}^N(\boldsymbol{\mu})$ is consistent for $\alpha$ in probability. So, combining forms (15) and (16), we ultimately have the consistency of $\widehat{\alpha}^N(\widehat{\boldsymbol{\mu}}^N)$ for $\alpha$.

Possible improvement to $\widehat{\alpha}^N(\widehat{\boldsymbol{\mu}}^N)$ can be obtained by adopting a one-step Newton correction. Using the notation

$$l'(\boldsymbol{\mu}, \alpha) = \frac{\partial}{\partial \alpha} l(\boldsymbol{\mu}, \alpha) = \frac{\sum_{i=1}^N \|\mathbf{x}_i - \boldsymbol{\mu}\|^2}{\alpha} + \frac{N(p+1)\alpha^p}{(1-\alpha^{p+1})} - \frac{N}{(1-\alpha)}, \text{ and}$$

$$l''(\boldsymbol{\mu}, \alpha) = \frac{\partial^2}{\partial \alpha^2} l(\boldsymbol{\mu}, \alpha) = \frac{-\sum_{i=1}^N \|\mathbf{x}_i - \boldsymbol{\mu}\|^2}{\alpha^2} + N(p+1)\frac{p\alpha^{p-1} + \alpha^{2p}}{(1-\alpha^{p+1})^2} - \frac{N}{(1-\alpha)^2}$$

and

$$\widetilde{\alpha}^N(\boldsymbol{\mu}) = \widehat{\alpha}^N(\boldsymbol{\mu}) - \frac{l'(\boldsymbol{\mu}, \alpha)}{l''(\boldsymbol{\mu}, \alpha)} \tag{17}$$

this is

$$\widetilde{\alpha}^N(\widehat{\boldsymbol{\mu}}^N) = \widehat{\alpha}^N(\widehat{\boldsymbol{\mu}}^N) - \frac{l'(\widehat{\boldsymbol{\mu}}^N, \alpha)}{l''(\widehat{\boldsymbol{\mu}}^N, \alpha)}. \tag{18}$$

From display (15), it follows that

$$P[\widetilde{\alpha}^N(\widehat{\boldsymbol{\mu}}^N) \neq \widetilde{\alpha}^N(\boldsymbol{\mu})] \to 0 \quad \text{as} \quad N \to \infty. \tag{19}$$

Thus, the limiting behavior of $\widetilde{\alpha}^N(\widehat{\boldsymbol{\mu}}^N)$ is the same as that of $\widetilde{\alpha}^N(\boldsymbol{\mu})$. Standard arguments for the asymptotic normality of one-step Newton corrections of consistent estimators applied here show that since $\widehat{\alpha}^N(\boldsymbol{\mu})$ is consistent for $\alpha$,

$$\sqrt{N}(\widetilde{\alpha}^N(\boldsymbol{\mu}) - \alpha) \xrightarrow{d} \text{N}(0, \mathcal{I}^{-1}(\alpha)), \tag{20}$$

where $\mathcal{I}(\alpha)$ is given in display (6). Hence,

$$\sqrt{N}(\widetilde{\alpha}^N(\widehat{\boldsymbol{\mu}}^N) - \alpha) \xrightarrow{d} \text{N}(0, \mathcal{I}^{-1}(\alpha)). \tag{21}$$

Furthermore,

$$\frac{\sqrt{N}(\widetilde{\alpha}^N(\widehat{\boldsymbol{\mu}}^N) - \alpha)}{\sqrt{\mathcal{I}^{-1}(\widetilde{\alpha}^N(\widehat{\boldsymbol{\mu}}^N))}} \xrightarrow{d} \text{N}(0,1) \tag{22}$$

and

$$\frac{\sqrt{N}(\widetilde{\alpha}^N(\widehat{\boldsymbol{\mu}}^N) - \alpha)}{\sqrt{-l''(\widehat{\boldsymbol{\mu}}^N, \widetilde{\alpha}^N(\widehat{\boldsymbol{\mu}}^N))}} \xrightarrow{d} \text{N}(0,1) \tag{23}$$

and these can be applied to produce large $N$ confidence limits and tests of $\text{H}_0 : \alpha = \alpha_0$ in the one-sample model.

## 3 | MIXTURE MODELS WITH SYMMETRIC EXPONENTIAL FAMILY COMPONENTS

The symmetric single component one-sample model is typically too simple for applications and mixture distribution models are a way to obtain more realistic models. Mixture distributions are commonly used tools for modeling data which is thought to be generated by multiple underlying mechanisms, or, is sampled from multiple populations [5, 19]. In this section we consider a mixture model for component distributions of the form (2). The finite mixture model with $K$ components for a single observation $\mathbf{X}$ has pmf

$$h(\mathbf{x}|\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K, \alpha_1, \ldots, \alpha_K, \pi_1, \ldots, \pi_K) = \sum_{j=1}^K \pi_j f(\mathbf{x}|\boldsymbol{\mu}_j, \alpha_j), \tag{24}$$

where $\boldsymbol{\mu}_j$ is the $p$-dimensional central binary vector and $\alpha_j$ is the spread parameter for component $j$, the $\pi_j$ are the mixing proportions $\left(\pi_j > 0 \text{ and } \sum_{j=1}^K \pi_j = 1\right)$, and $f(\mathbf{x}|\boldsymbol{\mu}_j, \alpha_j)$ is the pmf in display (2) evaluated at $\mathbf{x}$ with central pattern $\boldsymbol{\mu}_j$ and variability parameter $\alpha_j$.

Now consider $N$ independent observations $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_N$ that comprise a training set of size $N$ and dimension $p$. The corresponding log-likelihood function is

$$\log h(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N|\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K, \alpha_1, \ldots, \alpha_K, \pi_1, \ldots, \pi_K) = \sum_{i=1}^N \log \sum_{j=1}^K \pi_j f(\mathbf{x}_i|\boldsymbol{\mu}_j, \alpha_j). \tag{25}$$

We introduce a latent $N$-dimensional component assignment vector $\boldsymbol{k}$ with entries in $\{1, 2, \ldots, K\}$ to encode

the mixture components from which each observation arises. Each element $k(i)$ in $\boldsymbol{k}$ is a stochastic indicator variable corresponding to observation $\mathbf{x}_i$ and $k(i)$ can take values $\{1, 2, \ldots, K\}$ for $i = 1, 2, \ldots, N$. The numbering of components and thus the exact mapping provided by $\boldsymbol{k}$ is arbitrary as long as it faithfully represents which observations belong to the same classes.

### 3.1 | Prior distributions for Bayes analysis

We give the mixing proportions $\pi_j$ (which must be positive and sum to 1) a symmetric Dirichlet prior distribution with concentration parameter $\rho/K$, that is

$$p(\pi_1, \ldots, \pi_K|\rho) \sim \text{Dirichlet}(\rho/K, \ldots, \rho/K) = \frac{\Gamma(\rho)}{(\Gamma(\rho/K))^K}\prod_{j=1}^{K}\pi_j^{\rho/K-1}. \quad (26)$$

Let $N_j$ be the number of observations assigned to component $j$ (the so-called occupation number). Given the mixing proportions $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)$, the indicator variables follow a distribution given by the pmf

$$p(k(1), \ldots, k(N)|\pi_1, \ldots, \pi_K) = \prod_{j=1}^{K}\pi_j^{N_j}. \quad (27)$$

Then the occupation numbers have joint pmf

$$p(N_1, \ldots, N_K|\pi_1, \ldots, \pi_K) = \frac{N!}{N_1! \ \ldots \ N_K!}\prod_{j=1}^{K}\pi_j^{N_j}. \quad (28)$$

The mixing proportions can be integrated out of form (27) [15] and the prior distribution for the indicator variables is

$$p(k(1), \ldots, k(N)|\rho) = \frac{\Gamma(\rho)}{\Gamma(N+\rho)}\prod_{j=1}^{K}\frac{\Gamma(N_j+\rho/K)}{\Gamma(\rho/K)}. \quad (29)$$

The conditional prior distribution for a single indicator variable given all others is obtained from form (29) by holding all but the particular variable fixed and is specified by the pmf on $\{1, 2, \ldots, K\}$

$$p(k(i) = j|\boldsymbol{k}_{-i}, \rho) = \frac{N_{-i,j} + \rho/K}{(N-1)+\rho}, \quad (30)$$

where the subscript $-i$ refers to all indices except $i$, and $N_{-i,j}$ denotes the number of observations other than $\mathbf{x}_i$ that are associated with component $j$. In deriving form (30) we use the fact that the observations are a priori exchangeable. If we consider the limit of this structure as the number of components $K \to \infty$, the conditional prior for $k(i)$ has the limits [7, 15] specified by

$$\text{components for which } N_{-i,j} > 0: \quad p(k(i) = j|\boldsymbol{k}_{-i}, \rho) = \frac{N_{-i,j}}{(N-1)+\rho} \quad (31)$$

$$\text{all other components combined}: \quad p(k(i) \neq k(i') \text{ for all } i \neq i'|\boldsymbol{k}_{-i}, \rho) = \frac{\rho}{(N-1)+\rho}. \quad (32)$$

These probabilities in displays (31) and (32) are the probabilities for seating a new customer in a Chinese Restaurant Process (CRP) [1, 14] with infinitely many tables. The CRP is a discrete-time stochastic process and is used to specify a distribution over partitions of $N$ observations. Hence, it can be used as a prior distribution for the latent variables $k(i)$ which determine the component assignments of the $\mathbf{x}_i$.

The limit of the model described above is equivalent to a Dirichlet process mixture model [2, 6] with a countably infinite number of components. It has been argued in [18] that with the particular parameterization of the Dirichlet prior over $\{\pi_1, \ldots, \pi_K\}$ used here, that as $N \to \infty$ the number of components typically required to model $N$ observations becomes independent of $K$ and is approximately $\mathcal{O}(\rho \log(N))$. Thus, the mixture model stays well defined as $K \to \infty$, leading to what is known as an infinite mixture model [13, 15]. Infinite mixtures are known to work well even when there are only a small finite number of components in the true mixture. As a result, the infinite mixture models are commonly used models whenever the true number of components is unknown. For our purposes we truncate the Dirichlet process mixture model to have $K < \infty$ components. $K$ is chosen to be large and the parameter $\rho$ controls the number of nonempty components in a direct manner, larger $\rho$ implying larger numbers of nonempty components a priori. The parameter $\rho > 0$ can be thought of as a prior sample size (in terms of the Pólya urn scheme [4]) or a total prior mass for the component assignment vector $\boldsymbol{k}$.

Plausible prior distributions for the component parameters $\boldsymbol{\mu}_j$ and $\alpha_j$ of the mixture model are discrete uniform priors on central patterns $\boldsymbol{\mu}_j$; $j = 1, 2, \ldots, K$ and independently the Jeffreys priors [8] for variability parameters $\alpha_j$; $j = 1, 2, \ldots, K$, that is

$$p(\boldsymbol{\mu}_j, \alpha_j) \propto p_{\boldsymbol{\mu}}(\boldsymbol{\mu}_j)p_{\alpha}(\alpha_j); \quad j = 1, \ldots, K,$$

where

$$p_{\boldsymbol{\mu}}(\boldsymbol{\mu}_j) = \frac{1}{2^p}; \quad j = 1, \ldots, K \quad (33)$$

and

$$p_{\alpha}(\alpha_j) \propto \sqrt{\mathcal{I}(\alpha_j)}; \quad j = 1, \ldots, K \quad (34)$$

for $\mathcal{I}(\alpha)$ given in display (6). One might also consider *beta* priors for the variability parameters $\alpha_j$, however for our simulation studies and applications to real datasets discussed later we consider the Jeffreys priors given in display (34). The following section details two MCMC sampling procedures we employ for generating samples from the posterior distributions of the mixture component parameters and the indicator variables.

## 3.2 | MCMC samplers

For inference in the mixture model described in the previous section we employ two MCMC algorithms where the Markov chains use Gibbs and Metropolis updates. Gibbs sampling and the Metropolis algorithm are two well-known MCMC sampling techniques for complicated multivariate probability distributions when direct sampling is difficult. Based on whether the mixing proportions are included in the model (and hence, the sampling scheme) or not, we propose two variants of the MCMC sampler.

The mixing proportions are integrated out in the first version. In that case, it follows from Equation (29) that the prior distribution for the $N$-dimensional component assignment vector $\boldsymbol{k}$ is

$$g(\boldsymbol{k}) \propto \left( \prod_{j\, s.t.\, N_j>0} \left(\frac{\rho}{K}\right)\left(\frac{\rho}{K}+1\right) \ldots \left(\frac{\rho}{K}+(N_j-1)\right) \right).$$

So, under the present model assumptions, for a chosen $K$, the joint distribution of the $\mathbf{X}_i$, the component parameters $\boldsymbol{\mu}_j$ and $\alpha_j$, and the indicator vector $\boldsymbol{k}$ in the first version of the sampler has density proportional to

$$g(\boldsymbol{k}) \prod_{j=1}^{K} \left( \prod_{i:k(i)=j} f(\mathbf{x}_i|\boldsymbol{\mu}_j, \alpha_j) \right) p_{\boldsymbol{\mu}}(\boldsymbol{\mu}_j) p_{\alpha}(\alpha_j), \qquad (35)$$

where $f(\cdot)$ is as given in display (2), $p_{\boldsymbol{\mu}}(\cdot)$ and $p_{\alpha}(\cdot)$ are given in Equations (33) and (34) respectively.

At each iteration of the MCMC algorithm, one updates the component parameters and the indicator variables based on sampling schemes determined by values of all other variables. For the central patterns $\boldsymbol{\mu}_j$, we use Metropolis updates. A Gibbs step for $\alpha_j$ is not obvious and so a Metropolis step is used. To make such a Metropolis step it is convenient to employ normal proposals, but that requires a parameterization with parameter space $\mathbb{R}$. Thus, we consider the logit transformation of the variability parameter $\alpha_j$ to $\theta_j = \log\left(\frac{\alpha_j}{1-\alpha_j}\right) \in \mathbb{R}$, which allows proposal of a value from a normal distribution centered at the value of the current iterate with fixed standard deviation $\sigma$. The assignment indicators are updated one at a time following relationship (30) using conventional Gibbs steps. The first version of the MCMC sampler, a Metropolis-within-Gibbs sampler, is then summarized below.

1. Choose appropriate values of $K$, $\rho$, $\sigma$, and set the number of iterations $T$.
2. Initialize $\boldsymbol{\mu}_j$ and $\alpha_j$ for $j = 1, \ldots, K$, and $\boldsymbol{k}$.
3. Repeat $T$ times.
   - **(a)** Conditioned on the data $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$, $\boldsymbol{k}$, and corresponding $\alpha_j$, update the central pattern $\boldsymbol{\mu}_j$ for component $j$ ($j = 1, \ldots, K$) using a density proportional to

   $$\left( \prod_{i:k(i)=j} f(\mathbf{x}_i|\boldsymbol{\mu}_j, \alpha_j) \right) p_{\boldsymbol{\mu}}(\boldsymbol{\mu}_j)$$

   via a Metropolis step. A proposal $\boldsymbol{\mu}_j^*$ is obtained from a uniform distribution over vectors one pixel flip away from the current $\boldsymbol{\mu}_j$. The acceptance ratio then employed is

   $$\frac{\alpha_j^{\sum_{i:k(i)=j} \|\mathbf{x}_i - \boldsymbol{\mu}_j^*\|^2} \Big/ \prod_{i:k(i)=j} \binom{p}{\|\mathbf{x}_i - \boldsymbol{\mu}_j^*\|^2}}{\alpha_j^{\sum_{i:k(i)=j} \|\mathbf{x}_i - \boldsymbol{\mu}_j\|^2} \Big/ \prod_{i:k(i)=j} \binom{p}{\|\mathbf{x}_i - \boldsymbol{\mu}_j\|^2}}.$$

   - **(b)** Conditioned on the data $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$, $\boldsymbol{k}$, and corresponding $\boldsymbol{\mu}_j$, update the logit of the variability parameter, $\theta_j$ for component $j$ ($j = 1, \ldots, K$), using a density proportional to

   $$\left( \prod_{i:k(i)=j} f\left(\mathbf{x}_i|\boldsymbol{\mu}_j, \frac{\exp(\theta_j)}{1+\exp(\theta_j)}\right) \right) p_{\alpha}\left(\frac{\exp(\theta_j)}{1+\exp(\theta_j)}\right)\left(\frac{\exp(\theta_j)}{(1+\exp(\theta_j))^2}\right)$$

   via a Metropolis step by proposing a value $\theta_j^*$ from a normal distribution centered at the current $\theta_j$ with fixed standard deviation $\sigma$.
   - **(c)** Conditioned on the data $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$, $\boldsymbol{\mu}_j$, and $\alpha_j$ (or, equivalently $\theta_j$), for each $i = 1, 2, \ldots, N$, update the $k(i)$'s one at a time by sampling from discrete distributions on $\{1, 2, \ldots, K\}$ with probabilities proportional to

   $$(N_{-i,j} + \rho/K) f(\mathbf{x}_i|\boldsymbol{\mu}_j, \alpha_j)$$

A more time-efficient Metropolis-within-Gibbs algorithm can be formulated by including the mixing proportions $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)$ in the model rather than

integrating them out. In that case, considering a symmetric Dirichlet prior for $\boldsymbol{\pi}$ given in display (26), the joint density for the $\mathbf{X}_i$, the parameters $\boldsymbol{\mu}_j$ and $\alpha_j$, the mixing proportions $\boldsymbol{\pi}$, and the indicator vector $\boldsymbol{k}$ is proportional to

$$\left[\prod_{i=1}^{N}\pi_{k(i)}f(\mathbf{x}_i|\boldsymbol{\mu}_{k(i)},\alpha_{k(i)})\right]\left[\prod_{l=1}^{K}p_{\boldsymbol{\mu}}(\boldsymbol{\mu}_l)p_{\alpha}(\alpha_l)\right]\left[\left(\frac{\Gamma(\rho)}{(\Gamma(\rho/K))^K}\right)\prod_{m=1}^{K}\pi_m^{\rho/K-1}\right], \quad (36)$$

where $f(\cdot)$ is as given in form (2), $p_{\boldsymbol{\mu}}(\cdot)$ and $p_{\alpha}(\cdot)$ are given in Equations (33) and (34) respectively. At each iteration of this algorithm we sample the component parameters, the indicator variables, and (unlike the previous algorithm) the mixing proportions jointly from their corresponding posterior distribution using a standard Gibbs step. The second Metropolis-within-Gibbs sampler is summarized below.

1. Choose appropriate values of $K$, $\rho$, $\sigma$, and set the number of iterations $\boldsymbol{T}$.
2. Initialize $\boldsymbol{\mu}_j$ and $\alpha_j$ (for $j = 1, \ldots, K$), $\boldsymbol{k}$, and $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K) \sim$ Dirichlet $\left(\frac{\rho}{K}, \ldots, \frac{\rho}{K}\right)$.
3. Repeat $T$ times.
   - **(a)** The central pattern $\boldsymbol{\mu}_j$ for component $j$ ($j = 1, \ldots, K$) is updated following step 3(a) in the previous algorithm (first version of the sampler).
   - **(b)** The logit of the variability parameter, $\theta_j$ for component $j$ ($j = 1, \ldots, K$), is updated using a Metropolis step identical to step 3(b) in the first version of the sampler.
   - **(c)** Conditioned on the data $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$, $\boldsymbol{k}$, and parameters $\boldsymbol{\mu}_j$ and $\alpha_j$ (or equivalently $\theta_j$) for $j = 1, \ldots, K$, update $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)$ by sampling from Dirichlet $\left(\frac{\rho}{K} + N_1, \ldots, \frac{\rho}{K} + N_K\right)$ where $N_j = |\mathbf{x}_i : k(i) = j;\ i = 1, \ldots, N|$ for $j = 1, \ldots, K$.
   - **(d)** Conditioned on the data $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$, $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)$, $\boldsymbol{\mu}_j$, and $\alpha_j$ (or equivalently $\theta_j$), the $k(i)$'s are sampled independently, each from $\{1, 2, \ldots, K\}$, with probabilities proportional to

$$\pi_j f(\mathbf{x}_i|\boldsymbol{\mu}_j, \alpha_j).$$

Note that the $k(i)$'s can be sampled all at once, unlike the one-at-a-time update in the first algorithm, avoiding the recomputation of $N_{-i,j}$'s for each $i$. Moreover, since for a given $p$ the number of possible observation vectors is at most $2^p$, the probabilities in (d) above need to be calculated for at most $2^p$ vectors instead of $N$ observations. This is helpful for datasets where $N >> 2^p$. We use this second sampler for the Newspaper Reading Survey dataset discussed later where $N = 10{,}858$ and $p = 7$.

# 4 | SIMULATION STUDIES

## 4.1 | Method performance

We evaluated the performance of Bayes analysis of the mixture model under the following simulation settings. We considered different dataset sizes $N$, different dimensions of the problem $p$, and variants of the "true" number of components $K_{true}$ to generate observations from a $\pi_1 = \pi_2 = \cdots = \pi_{K_{true}}$ mixture with:

1. $N = 500$, $p = 4$, and $K_{true} = 2$;
2. $N = 1000$, $p = 6$, and $K_{true} = 7$; and
3. $N = 2000$, $p = 8$, and $K_{true} = 10$.

For each of the above specifications we considered five distinct sets of central vectors $\boldsymbol{\mu}$ and spread parameters $\alpha$. Details regarding the sets of central vectors and variability parameters used to generate the datasets under each setting can be found in the Supporting information. We simulated 10 datasets from each set of model parameters $\boldsymbol{\mu}$ and $\alpha$. For each dataset, we compared the probabilities from the "true" mixture model with the probabilities estimated after $T = 10{,}000$ complete iterations of the MCMC sampler (the probabilities estimated considering all 10,000 iterations and those estimated after a burn-in of the first 1000 iterations are almost identical). Both versions of the MCMC sampler were implemented for each of the above specifications. The results were found to be similar. We report the results from the first version for specification (i) and from the second sampler for specifications (ii) and (iii).

The primary focus of this simulation study was to evaluate the proposed mixture model in terms of its performance in estimating the probabilities of observations $\mathbf{x}$ from the simulated datasets. Note that the number of distinct observations in a dataset of dimension $p$ can be at most $2^p$. The probability of an observation $\mathbf{x}$ from the "true" mixture model with equal mixture weights is

$$h(\mathbf{x}|\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_{K_{true}}, \alpha_1, \ldots, \alpha_{K_{true}}) = \sum_{j=1}^{K_{true}}\frac{1}{K_{true}}f(\mathbf{x}|\boldsymbol{\mu}_j, \alpha_j). \quad (37)$$

The estimated probability for the corresponding observation is computed by averaging across iterations the probabilities obtained from the mixture model identified at

each iterate as

$$\widehat{h}(\mathbf{x}|\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K, \alpha_1, \ldots, \alpha_K) = \left( \sum_{t=1}^{T} \sum_{j=1}^{K} \frac{N_j^t}{N} f(\mathbf{x}|\boldsymbol{\mu}_j^t, \alpha_j^t) \right) / T, \tag{38}$$

where the superscript $t$ denotes the $t$th iteration of the sampler.

We used the Kullback–Leibler divergence [11, 12] (written as KL divergence) to quantify the difference between the two probability distributions given by $h(\mathbf{x})$ and $\widehat{h}(\mathbf{x})$. KL divergence originated in probability theory and information theory. It is a nonsymmetric, nonnegative measure of how well the "true" distribution of observations, or a precisely calculated theoretical distribution is approximated by the model in question. A KL divergence of 0 indicates that the two distributions in question are identical. The KL divergence between $h(\mathbf{x})$ and $\widehat{h}(\mathbf{x})$ is given by

$$D_{KL}(h||\widehat{h}) = \sum_{\mathbf{x}\in\{0,1\}^p} h(\mathbf{x}) \log \left( \frac{h(\mathbf{x})}{\widehat{h}(\mathbf{x})} \right). \tag{39}$$

Note that the sum in Equation (39) is over all $2^p$ possible choices of observation vectors. For the Bayes analysis of the mixture model, the value of $K$ was chosen to be at least twice the value of $K_{true}$ for each specification ($K$ is chosen to be 5, 15, and 20 for specifications (i), (ii), and (iii) respectively). The model parameter $\rho$ and the MCMC algorithm parameter $\sigma$ were chosen to be 20 and 0.5 respectively. The starting values for the logits of the variability parameters were chosen randomly from a normal distribution with mean zero and standard deviation $\sigma$. The $p$-dimensional central binary vectors were randomly initialized from a uniform distribution. The initial indicator variables were chosen as iid from a discrete uniform distribution on $\{1, 2, \ldots, K\}$.

For each dataset, we compared the KL divergence computed between the "true" model probabilities and the probabilities estimated from the proposed mixture model with the KL divergence computed between the "true" model probabilities and the relative frequency distribution of the observed dataset. This was then repeated for every combination of model parameters $\boldsymbol{\mu}$ and $\alpha$ under each simulation setting. The comparisons show clearly the advantage of the Bayes analysis of the mixture model over completely nonparametric inference for the distributions on $\{0,1\}^p$. It can also be observed that the method's performance gets better with increasing values of $K_{true}$.

Although details are not reported in the results here, the proposed Bayes analysis of the mixture model also accurately identified, under each setting, the sets of central vectors $\boldsymbol{\mu}$ used to generate the respective datasets in terms of the frequencies of vectors observed across all the MCMC iterations.

### 4.2 | Computational complexity

The runtimes for both versions of the MCMC sampler scale approximately linearly in the number of iterations ($T$) and the number of observations ($N$). The dimension $p$ of the dataset has negligible effect on the compute times. Additionally, with increase in the number of components $K$ specified for the proposed model the runtimes increase almost linearly. The computational complexity (in terms of time) of the proposed Bayes analysis of the mixture model can be represented as approximately $\mathcal{O}(T \cdot N \cdot K)$. The second MCMC sampler is found to be approximately four times as efficient (in terms of compute time) as the first version. Thus, the second sampler is recommended when $N >> 2^p$, as is implemented for the Newspaper Reading Survey study in Section 5.2. The timing results are based on the implementations of the MCMC samplers coded in R on a Condo Cluster compute node with two 2.6 GHz 8-core Intel E5-2640 v3 Haswell processors.

## 5 | APPLICATION TO REAL DATASETS

In this section we demonstrate the performance of the proposed Bayesian mixture model analysis on two real datasets. These datasets were shared with us by Dr. A.C. Tamhane and his coauthors from their research study in [17]. We extend our sincere gratitude to the authors for their contribution. The first dataset is from a teaching style study conducted by [3]. The second dataset comes from a research project at the Media Management Center at Northwestern University.

### 5.1 | Teaching style study

A survey of 468 fourth grade teachers from the counties of Lancashire and Cumbria was conducted by [3] in which each teacher was asked 38 questions, each of which were binary in nature (yes–no questions), about the way classes are handled. The dataset reports the answers to the following 6 questions from the 38.

- **Q.1.** Pupils not allowed to move around? (Y = 1, N = 0)
- **Q.2.** Pupils not allowed to talk? (Y = 1, N = 0)
- **Q.3.** Pupils expected to be quiet? (Y = 1, N = 0)
- **Q.4.** Explore concepts (1) or develop numerical skills (0)?

**Q.5.** Emphasis on separate subject teaching? (Y = 1, N = 0)
**Q.6.** Emphasis on integrated teaching? (Y = 1, N = 0)

For this dataset $N = 468$, and $p = 6$. The number of distinct binary vectors present in this dataset is 49 out of the possible 64 ($2^p$). Both versions of the MCMC sampler were implemented for this study. The second version was approximately four times as fast per iteration. The results were similar in terms of the most frequent central vectors identified, the mixing behavior of iterates of the variability parameters, the estimated probabilities of vectors, and the KL divergences. The results reported below are for the first version of the sampler.

We ran five MCMC chains initialized at different sets of $\boldsymbol{\mu}$ and $\alpha$ chosen randomly. The initial set of indicator variables for each MCMC run was chosen randomly from a discrete uniform distribution. For each chain we set $K = 30$, $\rho = 20$, $\sigma = 0.5$, and $T = 10{,}000$ iterations. The results reported for this analysis are considered over all 10,000 complete iterations. A comparison of the relative frequencies for the 10 most frequent binary vectors observed in the dataset with the respective probabilities estimated from the mixture model for each MCMC run suggests that the Bayesian mixture model approach performs well in approximating the observed relative frequencies with the model probabilities. The estimated probability for a vector is computed according to formula (38). Similar performance results could also be inferred from the KL divergence measure. For the five MCMC runs, the KL divergence values were found to be 0.072, 0.071, 0.074, 0.074, and 0.072 respectively. For reference, the KL divergence measure for estimation with uniform probability is 0.664. When computing the KL divergence according to Equation (39), $h(\mathbf{x})$ is considered to be the relative frequency of an outcome $\mathbf{x}$ observed in the dataset and $\widehat{h}(\mathbf{x})$ is the probability estimated from the mixture model for the corresponding $\mathbf{x}$. (When $h(\mathbf{x})$ is zero the contribution of the corresponding term is interpreted as zero because $\lim_{y\to 0^+} y\log(y) = 0$.)

The initialization of the parameters of the mixture model does not seem to have any effect on the method's performance. The model tends to favor a high number of components as can be observed from the distribution of number of nonempty components chosen over $T = 10{,}000$ iterations at each run of the MCMC algorithm. This is also an effect of the parameter $\rho$ which is set at 20 for this study. The high value of $\rho$ is equivalent to setting a uniform prior on the component assignment vector $\boldsymbol{k}$ and thus results in a large number of nonempty components. $K = 30$ seems to be a good choice since the relative frequencies start dropping off after 27 or 28 components. The nine most common central vectors identified by the Bayesian mixture model analysis in terms of their frequencies across all iterations (all five MCMC runs combined) are reported in Table 1.

**TABLE 1** Nine most frequent central vectors identified by the Bayesian mixture model analysis in the Teaching Style Study dataset reported in descending order of frequencies

| Q.1. | Q.2. | Q.3. | Q.4. | Q.5. | Q.6. |
|---|---|---|---|---|---|
| 1 | 1 | 1 | 0 | 1 | 0 |
| 1 | 1 | 1 | 0 | 0 | 1 |
| 1 | 1 | 1 | 1 | 1 | 0 |
| 1 | 1 | 0 | 0 | 1 | 0 |
| 1 | 1 | 0 | 1 | 1 | 0 |
| 1 | 1 | 1 | 1 | 0 | 1 |
| 0 | 0 | 0 | 1 | 0 | 1 |
| 0 | 0 | 0 | 0 | 1 | 0 |
| 1 | 1 | 1 | 1 | 1 | 1 |

These nine vectors identified by the analysis are among the 11 most frequent outcomes observed in the dataset. It seems that there exist multiple groups of teachers whose answers to Q.1., Q.2., and Q.3. are "yes," suggesting a strict approach to teaching. There also exist two groups of teachers who are lenient in terms of their "no" answers to these three questions. As suggested by the central vectors mentioned above, teachers are almost equally split with their responses to Q.4., Q.5., and Q.6. irrespective of their strict or lenient approach. Note that the answers to Q.5. and Q.6 are supposed to be different as indicated by the central vectors. The model also identifies a group of teachers who have responded "yes" to all six questions indicating possible inattention to the survey.

The plots for iterates of the variability parameters $\alpha$ that go with the above central vectors and their corresponding histograms suggest that the MCMC chains seem to have mixed well. The most frequent vector observed in the dataset, that is (1,1,1,0,1,0), is associated with the smallest estimate (average of the iterates) of the spread parameter in each of the five MCMC runs. This indicates that the probability of occurrence of this particular outcome is mostly explained by the component with $\boldsymbol{\mu} = (1,1,1,0,1,0)$, contributions to probabilities of other outcomes which are nonzero pixel flip counts away from (1,1,1,0,1,0) being small. For the rest of the central vectors identified and reported here, the estimates of the corresponding spread parameters are larger thereby accounting for considerable contributions towards probabilities of other observed outcomes. The component with $\boldsymbol{\mu} = (1,1,0,1,1,0)$ seems to give appreciable probability to the outcome (1,1,0,0,1,0) which is one pixel flip away, and vice-versa. The components with centers (1,1,1,1,1,0) and (1,1,1,1,0,1) seem to assign substantial probability to outcomes at the other center (being

two pixel flips away) as well as to (1,1,1,1,1,1) which is a single pixel flip away from both the central vectors.

## 5.2 | Newspaper Reading Survey

This dataset was generated from a mail survey conducted by the Newspaper Association of America on $N = 10{,}858$ readers. The subjects responded to seven questions (Q.i., i = 1, … , 7) on whether they read a particular newspaper on each day (i) of the week starting from Monday. There are 118 distinct multivariate binary vectors observed out of the possible 128 ($2^p$, where $p$ is 7). For this dataset, we again implemented both versions of the MCMC sampler. Since $N \gg 2^p$, the second Metropolis-within-Gibbs sampler (with the mixing proportions) was considerably faster per iteration than the first (where the proportions are integrated out). The results for the two versions were similar in terms of the most frequent central vectors identified, the mixing behavior of variability parameter iterates, the estimated probabilities of vectors, and the KL divergences. Reported below are the results for the second version of the MCMC sampler.

As for the previous analysis, we ran five MCMC chains with different starting values for the mixture component parameters, the indicator variables, and the mixing proportions. We ran each chain with $K = 50$ components, $\rho = 20$, $\sigma = 0.5$, and for $T = 10{,}000$ iterations. The results reported are considered over all 10,000 complete iterations. The effective performance of the proposed modeling approach is evident from the KL divergence measure and the comparison of relative frequencies for the 10 most frequent vectors observed in the dataset with their respective probabilities estimated from the Bayesian mixture model for each MCMC run. The KL divergence values for the five MCMC runs were 0.005 each. For reference, the KL divergence for estimation with uniform probability is 2.407. The KL divergence for each run is computed according to Equation (39), where $h(\mathbf{x})$ is the relative frequency of an outcome $\mathbf{x}$ observed in the dataset and $\widehat{h}(\mathbf{x})$ is the probability estimated from the mixture model for the corresponding $\mathbf{x}$. (Again, when $h(\mathbf{x})$ is zero the contribution of the corresponding term is interpreted as zero because $\lim_{y\to 0^+} y\log(y) = 0$.)

The initialization of the parameters does not seem to have any effect on the method's performance. As for the analysis with the Teaching Style Study dataset, similar observations can be inferred from the distribution of number of components with nonempty sets of related observations identified over $T = 10{,}000$ iterations for each run of the algorithm. For clarity we only mention those numbers of nonempty components which have occurred for more than 100 iterations out of the 10,000. The value of $\rho$ is set to be 20 in this study as well. Again, for such a high value of $\rho$ the mixture model identifies a large number of nonempty components. $K = 50$ seems to be an optimal choice for the number of components since the relative frequencies start dropping off after 46 or 47 components.

**TABLE 2** Nine most highly probable central vectors identified by the Bayesian mixture model analysis in the Newspaper Reading Survey dataset reported in descending order of frequencies

| Q.1. | Q.2. | Q.3. | Q.4. | Q.5. | Q.6. | Q.7. |
|---|---|---|---|---|---|---|
| 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| 0 | 0 | 0 | 0 | 0 | 1 | 1 |
| 0 | 0 | 1 | 0 | 0 | 0 | 1 |
| 1 | 1 | 1 | 1 | 1 | 0 | 1 |
| 0 | 0 | 0 | 1 | 1 | 1 | 1 |
| 1 | 0 | 0 | 0 | 1 | 1 | 1 |
| 0 | 0 | 0 | 0 | 1 | 1 | 1 |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 |

Across all five MCMC runs combined, the nine most frequent central vectors identified by the analysis for the Newspaper Reading Survey dataset are reported in Table 2. Seven of these nine vectors comprise the seven most frequent outcomes observed in the dataset.

The identified central vectors indicate that there are groups of readers who usually read the particular newspaper more on weekends, especially on Sundays, as compared to the middle days of the week. There are readers who read the newspaper on all seven days of the week. This group of readers can be thought of as "subscribers" to the newspaper. The analysis also points out that there is a sizable group of readers who do not read the newspaper at all.

From the plots and histograms of the corresponding variability parameters, it can be seen that the estimate of $\alpha$ corresponding to $\boldsymbol{\mu} = (1,1,1,1,1,1,1)$, the most frequent outcome in the dataset, is close to zero for each MCMC run. This suggests that the probability of the most frequent outcome is mostly contributed by the component with $\boldsymbol{\mu} = (1,1,1,1,1,1,1)$. Similar conclusions can be drawn for the outcomes (0,0,0,0,0,0,1), (0,0,0,0,0,0,0), and (0,0,0,0,0,1,1), the second, third, and fourth most frequent outcome in the dataset respectively. These four outcomes make up close to 72% of the observations in the entire dataset. The model seems to identify almost degenerate distributions at these four components, although, there seems to be some contribution of a component with central vector $\boldsymbol{\mu} = (0,0,0,0,0,0,1)$ towards the probabilities of outcomes (0,0,0,0,0,1,1) and (0,0,0,0,1,1,1), one and two pixel

flips away respectively, as well as from $\boldsymbol{\mu} = (0,0,0,0,0,0,0)$ towards the probability of (0,0,0,0,0,0,1). From the variability parameter estimates corresponding to the other central vectors reported, their respective components do seem to give appreciable probability to the rest of the outcomes in the dataset.

# 6 | HANDLING MULTIVARIATE BINARY VECTORS WITH MISSING ENTRIES

The occurrence of missing values, and hence, incomplete datasets is not uncommon in applications. In this section we extend our proposed Bayesian mixture model analysis to datasets with missing entries under an implicit assumption that missingness is not informative, that is, occurs at random.

## 6.1 | Modeling entries missing at random

Suppose that a $p$-dimensional binary vector $\mathbf{X} \in \{0,1\}^p$ has the pmf given in display (2). Now assume that some of the $p$ entries of an observed $\mathbf{x}$ are missing. Specifically, $l$ entries of $\mathbf{x}$ are assumed to be observed and the remaining $(p-l)$ of them are assumed to be missing. The observed and the missing entries of $\mathbf{x}$ can be denoted by $\mathbf{x}_1^l = (x_1, x_2, \ldots, x_l)$ and $\mathbf{x}_{l+1}^p = (x_{l+1}, x_{l+2}, \ldots, x_p)$ respectively. Note that the indices of the missing elements and their ordering need to be maintained while referencing the component central vectors. We consider the conditional distribution of $\mathbf{x}_{l+1}^p = (x_{l+1}, x_{l+2}, \ldots, x_p)$ given $\mathbf{x}_1^l = (x_1, x_2, \ldots, x_l)$ which is specified by

$$f(\mathbf{x}_{l+1}^p | \mathbf{x}_1^l, \boldsymbol{\mu}, \alpha) \propto \frac{\alpha^{(\|\mathbf{x}_{l+1}^p - \boldsymbol{\mu}_{l+1}^p\|^2)}}{\binom{p}{\|\mathbf{x}_{l+1}^p - \boldsymbol{\mu}_{l+1}^p\|^2 + \|\mathbf{x}_1^l - \boldsymbol{\mu}_1^l\|^2}}. \quad (40)$$

For using this in a Gibbs step for generating $\mathbf{x}_{l+1}^p \in \{0,1\}^{p-l}$ at an iterate of an MCMC sampler for a Bayes mixture model, we first generate a number of pixel flips away from $\boldsymbol{\mu}_{l+1}^p$ and then distribute them randomly to coordinates of $\boldsymbol{\mu}_{l+1}^p$. The conditional pmf of the number of pixel flips is

$$q(m | \mathbf{x}_1^l, \boldsymbol{\mu}, \alpha) \propto \frac{\binom{p-l}{m} \alpha^m}{\binom{p}{m + \|\mathbf{x}_1^l - \boldsymbol{\mu}_1^l\|^2}},$$

$$\text{where} \quad m = 0, 1, \ldots, (p-l). \quad (41)$$

## 6.2 | An MCMC sampler

In the context of multivariate binary vectors with missing entries we treat unobserved entries of data vectors as auxiliary or latent variables and an appropriate MCMC sampler is presented below.

1. Choose appropriate values of $K$, $\rho$, $\sigma$, and set the number of iterations $T$.
2. Initialize $\boldsymbol{\mu}_j$ and $\alpha_j$ for $j = 1, \ldots, K$ and $\boldsymbol{k}$. Generate starting values for the missing entries in an observation conditioned on the mixture component parameters and the indicator variables according to display (41).
3. Repeat $T$ times.
   - **(a)** To update the central pattern $\boldsymbol{\mu}_j$ for component $j$ $(j = 1, \ldots, K)$, a Metropolis step identical to step 3(a) in the first version of the MCMC sampler for non-missing entries is used.
   - **(b)** The logit of the variability parameter, $\theta_j$ for component $j$ $(j = 1, \ldots, K)$, is updated using a Metropolis step identical to step 3(b) in the first version of the sampler for non-missing entries.
   - **(c)** The $k(i)$'s are updated one at a time by sampling from discrete distributions on $\{1, 2, \ldots, K\}$ following step 3(c) in the first version of the MCMC sampler for non-missing entries.
   - **(d)** Update the values for the missing entries in each vector $\mathbf{x}_i$ conditioned on the mixture component parameters and the indicator variables following formula (41).

## 6.3 | Application to a real dataset

The modified version of the Bayesian mixture model analysis under the assumption that entries are missing at random was applied to the HouseVotes84 dataset [16] which can be found in the R package `mlbench` as well as on the University of California, Irvine Machine Learning Repository. The data concern United States Congressional Voting records in 1984. It includes votes for each of the U.S. House of Representatives Congressmen on the 16 key votes identified by the Congressional Quarterly Almanac (CQA). The CQA lists nine different types of votes: voted for, paired for, and announced for (these three simplified to "y" and are coded as 1), voted against, paired against, and announced against (these three simplified to "n" and are coded as 0), voted present, voted present to avoid conflict of interest, and did not vote or otherwise make a position known (these three simplified to an unknown disposition and are treated as missing). The dataset has $N = 435$ observations and $p = 16$ binary responses. The "Class" variable is the identifier of the political affiliation

("democrat" or "republican") of each congressmen. There are 267 "democrat"' and 168 "republican" representatives.

Two hundred and three of the 435 observations have missing entries. There are 160 distinct 16-dimensional binary vectors in the full dataset without any missing entries. We implemented five runs of the MCMC sampler for $T = 10{,}000$ iterations each with the full dataset, as well as with the dataset split up according to the "Class" variable. The five MCMC chains had different starting values for the mixture component parameters and the indicator variables. The values of $K$ and $\rho$ were specified to be 30 and 20 respectively and $\sigma$ was chosen to be 0.5 in each case. The results presented consider all the 10,000 iterations. The KL divergence measures computed (for the full dataset and the datasets separated by the political affiliations) between the relative frequencies of the complete observations and the respective probabilities estimated from the Bayes mixture model analysis suggest important differences.

The six most common central vectors identified by the model for the full dataset in terms of their frequencies across all iterations (all five MCMC runs combined) are reported in Table 3. Table 4 reports the six most frequent central vectors identified by the model for the datasets separated by the "Class" variable.

The central vectors identified by the model in the two datasets separated by the "Class" variable reveal the difference in opinions between the "democrat" representatives and their "republican" counterparts. The "republican" representatives tend to respond "n" to the third key vote as opposed to the "democrat" congressmen who tend to respond "y." On the other hand the former seem to be in favor of the fourth, fifth, and sixth issues unlike the latter. The same could be said for the 12th, 13th, and 14th key vote. Most democrats are in favor of the seventh, eighth, and ninth issues. Some republicans seem to agree with the democrats on the 16th key vote.

For most of the component central vectors reported, the corresponding variability parameters have chains that seem to be mixing well. It is possible that a single central vector is associated with more than one mixture components with different spread measures. The values of the variability parameter estimates corresponding to the central vectors reported in this analysis indicate that the components formed from these vectors have appreciable contribution towards probabilities of outcomes which are more than one pixel flip count away. For the full dataset, the number of nonempty components identified by the mixture model across the $T = 10{,}000$ iterations tends to range from 17 to 29. The number of nonempty components identified is between 15 and 25 and between 9 and 20 for the datasets with Class = "democrat" and Class = "republican" respectively. For the full dataset, we obtain the matrix of relative frequencies in the MCMC iterates (posterior probabilities) under which two congressmen are put into the same component. The heatmap from the resulting matrix is shown in Figure 1. With high posterior probability the 168 "republican" congressmen tend to belong to the same component as indicated from the lower left block. Similar observations could be made for the remaining 267 "democrat" congressmen as seen from the upper right block. The heatmap shown here corresponds to the fourth run of the MCMC sampler. Heatmaps from other four runs display similar block structure.

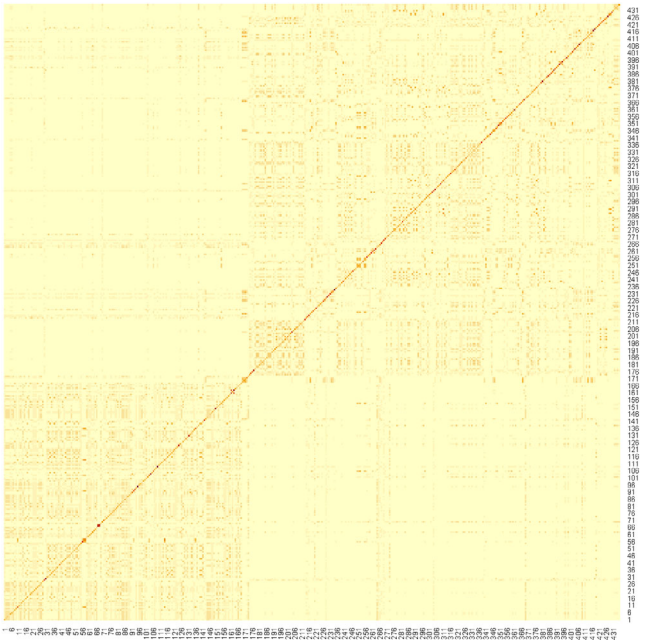

**FIGURE 1** Heatmap of posterior probabilities under which two congressmen are put into the same component. Darker shades indicate probabilities closer to 1

## 7 | CONCLUSIONS

We have proposed a symmetric distribution on multivariate vectors of 0's and 1's. It is based on the Euclidean distance between two $p$-dimensional binary vectors. This distribution has a simple form and its parameters are easily interpretable.

Subsequently, we have described the construction of a Dirichlet process mixture model within the Bayesian nonparametrics framework where the components of the mixture take the above distributional form. Using the Dirichlet prior on the component weights (or equivalently, the mixture proportions) has the attractive property that enables us to integrate out the weights and sample the latent indicators marginally using the CRP representation of the Dirichlet process. An alternative sampler is also

**TABLE 3** Six most frequent central vectors identified by the mixture model in the full HouseVotes84 dataset reported in descending order of frequencies

| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 0 | 1 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 1 | 0 | 1 |
| 0 | 0 | 1 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 1 | 1 |
| 0 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 0 | 1 |
| 1 | 1 | 1 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 1 | 0 | 1 |
| 0 | 0 | 0 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 0 |
| 0 | 0 | 1 | 0 | 0 | 1 | 1 | 1 | 1 | 0 | 1 | 0 | 0 | 1 | 1 | 1 |

**TABLE 4** Six most frequent central vectors identified for the two datasets defined by the "Class" variable reported in descending order of frequencies

| Class = "democrat" | Class = "republican" |
|---|---|
| (0, 0, 1, 0, 0, 1, 1, 1, 1, 1, 0, 0, 0, 1, 0, 1) | (0, 0, 0, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 0, 1) |
| (1, 1, 1, 0, 1, 1, 0, 0, 0, 0, 1, 0, 1, 1, 0, 1) | (0, 1, 0, 1, 1, 1, 0, 0, 0, 0, 0, 1, 1, 1, 0, 1) |
| (0, 0, 1, 0, 0, 1, 1, 1, 1, 0, 1, 0, 0, 1, 1, 1) | (0, 1, 0, 1, 1, 1, 0, 0, 0, 0, 0, 1, 1, 1, 0, 0) |
| (1, 1, 1, 0, 0, 0, 1, 1, 1, 0, 0, 0, 0, 0, 1, 1) | (0, 0, 0, 1, 1, 1, 0, 0, 0, 1, 0, 1, 1, 1, 0, 1) |
| (1, 0, 1, 0, 0, 0, 1, 1, 1, 1, 0, 0, 0, 0, 1, 1) | (0, 1, 0, 1, 1, 1, 0, 0, 0, 1, 0, 1, 1, 1, 0, 1) |
| (1, 1, 1, 0, 0, 0, 1, 1, 1, 0, 1, 0, 0, 0, 1, 1) | (0, 1, 0, 1, 1, 1, 0, 0, 0, 1, 0, 1, 1, 1, 0, 0) |

presented that includes the mixing proportions and results in a considerably faster MCMC algorithm. The results for the two versions of the sampler are similar with respect to the most frequent central vectors identified, the mixing behavior of iterates of the variability parameters, the estimated probabilities of vectors, and the KL divergences. The Dirichlet process mixture model forms a suitable modeling strategy when the "true" number of components is unknown and does not require the user to specify a "correct" number of components. Since the number of distinct binary vectors for any dimension $p$ is limited to $2^p$ we effectively use the mixture model with a finite number of components $K$. For practical purposes $K$ is chosen to be large and the number of nonempty components of the mixture is controlled using the parameter $\rho$. Inference is made from the posterior samples generated using the proposed Metropolis-within-Gibbs sampling techniques. The implementation of the model is illustrated through simulation studies and applications to two real datasets. The method shows good performance in estimating the relative frequencies (or, "true" probabilities from the generating mixture) for the vectors observed in the datasets. Although the model does not produce a single "best" set of mixture component parameters, the central vectors generating the simulated datasets are accurately identified in terms of the frequencies of vectors observed across the MCMC iterations. We have also extended the mixture model to allow for missing entries in the data vectors. As seen from the application on a real dataset, the method seems to identify potentially interpretable central vectors and allow assessment of similarities between incomplete observations based on posterior probabilities of arising from the same mixture component.

The proposed modeling approach draws a clear parallel to (the absolutely standard) mixture modeling with multivariate normal components in the case of continuous distributions. The fact that multivariate normal mixture modeling itself is closely related to multivariate kernel density estimation makes this approach a sort of "binary vector 'kernel' pmf estimation" methodology. Besides being as flexible (in terms of modeling of an arbitrary distribution for binary vectors) as other modeling approaches such as restricted Boltzmann machines (RBMs) for example, our proposed methodology remains far more interpretable, relatively parsimonious (in comparison to the RBMs), and does not suffer from serious limitations [9, 10]. It also provides rational bases for clustering training cases on the basis of posterior probabilities of coming from the same component, a real strength of the conceptualization (especially in where data vectors can be incomplete).

## ACKNOWLEDGMENT

Open access funding enabled and organized by Projekt DEAL.

## CONFLICT OF INTEREST

The authors declare no potential conflict of interests.

## DATA AVAILABILITY STATEMENT

The datasets and codes that support the findings of this study are publicly available in the GitHub repository 'Bayes-Mixture-Analysis-for-Binary-Vectors' at https://github.com/abhicc/Bayes-Mixture-Analysis-for-Binary-Vectors.git.

## ORCID

*Abhishek Chakraborty* https://orcid.org/0000-0002-8460-6696

## REFERENCES

1. D. J. Aldous, *Exchangeability and related topics*, in *École d'Été St Flour 1983*, Lecture Notes in Math, Vol 1117, Springer-Verlag, Berlin, Germany, 1985, 1–198.
2. C. E. Antoniak, *Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems*, Ann. Stat. 2(6) (1974), 1152–1174.
3. S. N. Bennett and J. Jordan, *A typology of teaching styles in primary schools*, Br. J. Educ. Psychol. 45(1) (1975), 20–28.
4. D. Blackwell and J. B. MacQueen, *Ferguson distributions via Pólya urn schemes*, Ann. Stat. 1(2) (1973), 353–355.
5. B. S. Everitt, *Finite mixture distributions*, Wiley StatsRef: Statistics Reference Online, New York, NY, 2014.
6. T. S. Ferguson, *A Bayesian analysis of some nonparametric problems*, Ann. Stat. 1 (1973), 209–230.
7. D. Görür and C. E. Rasmussen, *Dirichlet process Gaussian mixture models: Choice of the base distribution*, J. Comput. Sci. Technol. 25(4) (2010), 653–664.
8. H. Jeffreys, *An invariant form for the prior probability in estimation problems*, Proc. R. Soc. Lond. A Math. Phys. Sci. 186(1007) (1946), 453–461.
9. A. Kaplan, D. J. Nordman, and S. B. Vardeman, *Properties and Bayesian fitting of restricted Boltzmann machines*, Stat. Anal. Data Mining 12(1) (2019), 23–38.
10. A. Kaplan, D. J. Nordman, and S. B. Vardeman, *On the S-instability and degeneracy of discrete deep learning models*, Inform Inference 9(3) (2020), 627–655.
11. S. Kullback, *Information theory and statistics*, Wiley, New York, 1959.
12. S. Kullback and R. A. Leibler, *On information and sufficiency*, Ann. Math. Stat. 22(1) (1951), 79–86.
13. R. M. Neal, *Bayesian mixture modeling*, in *Maximum Entropy and Bayesian Methods*, Springer, Dordrecht, Netherlands, 1992, 197–211.
14. J. Pitman, *Combinatorial stochastic processes*, Ecole d'Eté de Probabilités de Saint-Flour XXXII-2002, Springer, 2006.
15. C. E. Rasmussen, *The infinite Gaussian mixture model*, in *Advances in Neural Information Processing Systems*, MIT Press, Cambridge, 2000, 554–560.
16. J. C. Schlimmer, Concept acquisition through representational adjustment, Ph.D. Thesis, Univ. California, Irvine, 1987.
17. A. C. Tamhane, D. Qiu, and B. E. Ankenman, *A parametric mixture model for clustering multivariate binary data*, Stat. Anal. Data Mining 3(1) (2010), 3–19.
18. Y. W. Teh, *Dirichlet process*, in *Encyclopedia of Machine Learning*, Springer, Boston, MA, 2010, 280–287.
19. D. M. Titterington, A. F. M. Smith, and U. E. Makov, *Statistical analysis of finite mixture distributions*, Wiley, New York, 1985.

## AUTHOR BIOGRAPHIES

**Abhishek Chakraborty** is an Assistant Professor of Statistics in the Department of Mathematics, Statistics, and Computer Science at Lawrence University, Wisconsin. He earned his PhD in Statistics from Iowa State University. His research interests include development of statistical methodologies for analysis of complex datasets, statistical learning, data mining, predictive modeling, application of Bayesian methodologies, statistics pedagogy, and history of statistics.

**Stephen B. Vardeman** is a University Professor of Statistics and Industrial Engineering at Iowa State University. He is a Fellow of the American Statistical Association, an Elected Member of the International Statistical Institute, was LAS Kingland Data Analytics Faculty Fellow 2017 through 2019, and was Editor of Technometrics from 1993 through 1995. His current professional interests include Statistical Machine Learning, Business and Engineering Analytics, Engineering and Natural Science applications of Statistics, Statistics and Metrology, Directional Data Analysis, Industrial applications, Statistical Education, and the development of new Statistical Theory and Methods.

## SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

**How to cite this article:** A. Chakraborty, and S. B. Vardeman, *Modeling and inference for mixtures of simple symmetric exponential families of p-dimensional distributions for vectors with binary coordinates*, Stat Anal Data Min: The ASA Data Sci Journal. (2021), 1–14. https://doi.org/10.1002/sam.11528