Biometrics XX, 1–24 DOI: 10.1111/XXXX

XXXX XXX

A Bayesian Approach to Restricted Latent Class Models for Scientifically-Structured Clustering of Multivariate Binary Outcomes

Zhenke Wu[1,*], Livia Casciola-Rosen[2], Antony Rosen[2], and Scott L. Zeger[3]

[1]Department of Biostatistics and Michigan Institute for Data Science,

University of Michigan, Ann Arbor, MI 48109, USA

[2]Division of Rheumatology, Department of Medicine,

Johns Hopkins University School of Medicine, Baltimore, Maryland, 21224, USA

[3]Department of Biostatistics, Johns Hopkins University, Baltimore, MD 21205, USA

email: zhenkewu@umich.edu

Summary: This paper presents a model-based method for clustering multivariate binary observations that incorporates constraints consistent with the scientific context. The approach is motivated by the precision medicine problem of identifying autoimmune disease patient subsets who may require different treatments. We start with a family of restricted latent class models or RLCMs (e.g., Xu and Shang, 2018). However, in the motivating example and many others like it, the unknown number of subsets and the definitions of the latent classes are among the targets of inference. We use a Bayesian approach to RCLMs in order to use informative prior assumptions on the number and definitions of latent classes to be consistent with scientific knowledge so that the posterior distribution tends to concentrate on smaller numbers of clusters and sparser binary patterns. The paper presents a novel posterior inference algorithm to handle discrete mixture parameters. Through simulations under the assumed model and realistic deviations from it, we demonstrate greater interpretability of results and superior finite-sample clustering performance for our method compared to common alternatives. The methods are illustrated with an analysis of protein data to detect clusters representing autoantibody classes among scleroderma patients.

Key words: Autoimmune disease; Clustering; Dependent Binary Data; Latent Class Models; Markov Chain Monte Carlo; Measurement Error; Mixture of Finite Mixture Models; Scleroderma.

This paper has been submitted for consideration for publication in *Biometrics*

1. Introduction

This paper proposes a model-based method for clustering multivariate binary observations while imposing constraints dictated by the scientific context. Suppose $\mathbf{Y}$ is an $N \times L$ binary data matrix of N observations, each with L dimensions or "features". Let $Y_{i\ell}$ be a noisy measurement of $\Gamma_{i\ell}$ that indicates the true presence/absence of feature ℓ for observation i. In the motivating example, $Y_{i\ell}$ and $\Gamma_{i\ell}$ are the observed and actual presence or absence of a protein of molecular weight ℓ in the serum of patient i. The scientific structure is respected by imposing constraints that the $\Gamma_{i\ell}$ can be represented by a smaller number (M) of *unobservable* or *latent* binary indicators $\boldsymbol{\eta}_i = (\eta_{i1}, \ldots, \eta_{iM})^\top$ that represent the true *states* of scientific interest. In the motivating example, multiple proteins form a complex or "machine" that performs a cellular function (e.g., Rosen and Casciola-Rosen, 2016). The immune system responds to abnormal machines rather than to individual proteins. $\boldsymbol{\eta}_{im}$ indicates whether or not the immune system of subject i responded to machine m, that is to all of its proteins. Given these definitions, we define clusters to be comprised of those observations with identical latent states $\boldsymbol{\eta}_i = \boldsymbol{\alpha}$. We assume that $\boldsymbol{\alpha}$ takes values within a finite but unknown subset $\mathcal{A}$ of $\{0,1\}^M$ with $M \leqslant L$.

Figure 1 shows a hypothetical patient whose $\boldsymbol{\eta}_i = (1,0,1)^\top$ indicating that her immune system produced autoantibodies to the proteins (autoantigens) in Machines 1 and 3 but not Machine 2 (middle panel). Subjects with identical $\boldsymbol{\eta}_i$ form a latent class. The $M \times L$ binary matrix Q denotes which proteins constitute each machine: $Q_{m\ell} = 1$ if protein ℓ is a component in machine m. We refer to the rows of Q as "machine profiles". The right panel of Figure 1 shows a simple example of three different machines with non-overlapping protein components. There may exist a protein component ℓ that is not an immunological target and does not contribute to the estimation of clusters. This biological knowledge is represented by $\sum_m Q_{m\ell} = 0$.

Γ represents the actual immune responses that can not be directly observed. We characterize the stochastic discrepancies between the actual $\Gamma_{i\ell}$ and observed presence/absence of autoantibodies $Y_{i\ell}$ using: true positive rates $\boldsymbol{\theta} = \{\theta_\ell = \mathbb{P}(Y_{i\ell} = 1 \mid \Gamma_{i\ell} = 1\})$ and false positive rates $\boldsymbol{\psi} = \{\psi_\ell = \mathbb{P}(Y_{i\ell} = 1 \mid \Gamma_{i\ell} = 0)\}$. In the motivating example, we will assume *a priori* high true and low false positive rates ($\theta_\ell > \psi_\ell$) because the measurement method using immunoprecipitation (IP) is known to be both sensitive and specific (e.g., Orito et al., 2006).

[Figure 1 about here.]

As detailed below, the models proposed to address the protein clustering problem are members of the family of *restricted latent class models* or RLCMs (e.g., Xu and Shang, 2018). In our problem, however, the definitions of machine profiles Q and the number of distinct latent states $|\mathcal{A}|$ are unknown and the dimension L is large relative to M. This corresponds to not knowing either the subsets of proteins that form each cellular machine or the combination of machines that the immune systems can target in the population of patients. The knowledge that the immune system attacks machines of multiple proteins rather than single proteins is why we refer to this approach as *scientifically-structured clustering* (SSC).

SSC for multivariate binary data has a number of potential advantages beyond the motivating example. Most importantly, the resulting clusters conform to the existing scientific context and therefore can be used to address relevant questions. SSC can also estimate clusters more efficiently than standard all-feature clustering methods such as latent class analysis or hierarchical clustering when the true clusters differ from one another at a relatively smaller number of features. The Supplementary Material A1.1 provides other similar examples from psychology (e.g., Junker and Sijtsma, 2001) and epidemiology (e.g., Wu et al., 2016).

In addressing the motivating problem, this paper makes three primary contributions to the

literature on RLCMs. First, in many applications, LCM or RLCM likelihood functions can be multimodal or relatively flat. The use of scientifically-based prior distributions can resolve these ambiguities. The Bayesian RLCM also improves finite-sample estimation efficiency, at the expense of some bias, by 1) inducing sparsity that propagates into the posterior distribution to encourage fewer clusters with sparser latent state patterns, and by 2) shrinking class-specific response probability estimates toward the model-based probabilities imposed by the scientific structure. Second, the paper presents a novel Markov chain Monte Carlo (MCMC) algorithm for Bayesian RLCMs building on the sampling techniques of Jain and Neal (2004), Miller and Harrison (2017) and Chen et al. (2017). The proposed algorithm addresses inferential issues unique to mixture models with discrete component parameters and jointly infers the number of clusters $|\mathcal{A}|$ and the matrix of machine profiles Q in addition to the other model parameters. Third, we connect three other model-based clustering methods for multivariate binary data to the latent class models to better understand their identifiability: probabilistic Boolean matrix decomposition (Rukat et al., 2017), subset clustering models (Hoff, 2005), and partially latent class models (Wu et al., 2016).

The rest of paper is organized as follows: Section 2 presents the model formulation and Section 3 derives its MCMC algorithm for posterior inference. Section 4.1 compares via simulations the proposed clustering method to three common alternatives. Section 4.2 illustrates the methods with an analysis of the autoantibody data from the motivating example. The paper concludes with a discussion of model extensions and limitations.

2. Models

2.1 *Latent Class Models*

First formulated by Lazarsfeld (1950), latent class models (LCMs) have become an important tool for modeling multivariate discrete responses (e.g., Goodman, 1974; Dunson and Xing,

2009) and model-based clustering (e.g., Vermunt and Magidson, 2002). LCMs are examples of latent variable models that assume the observed dependence among multivariate responses is induced by variation among unobserved or "latent" variables. In particular, LCMs attribute the covariance among discrete outcomes to individuals' shared, but unobserved membership in a few latent classes. Let $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{iL})^\top \in \{0,1\}^L$ be the binary response vector for observation $i = 1, \ldots, N$. Let $\widetilde{Z}_i \in \{1, \ldots, \widetilde{K}\}$ indicate the *unobserved* class assignment for observation i. An LCM assumes that subject i has a positive response for feature ℓ with probability: $\mathbb{P}(Y_{i\ell} = 1 \mid \widetilde{Z}_i = k, \lambda_{k\ell}) = \lambda_{k\ell}$, $0 \leqslant \lambda_{k\ell} \leqslant 1$, $k = 1, \ldots, \widetilde{K}$, $\ell = 1, \ldots, L$. Traditional LCMs impose no structure upon the response probabilities for feature ℓ other than that they differ among classes almost surely: $\lambda_{k\ell} \neq \lambda_{k'\ell}$ for latent classes $k \neq k'$, referred to as *between-class differential* measurement errors.

LCMs are based upon a *conditional independence* assumption whereby the measurements from distinct dimensions are independent of one another given the latent class and response probabilities in that class so that the conditional probability is given by $\mathbb{P}(\boldsymbol{Y}_i \mid \widetilde{Z}_i, \{\lambda_{k\ell}\}) = \prod_{\ell=1}^{L} (\lambda_{\widetilde{Z}_i\ell})^{Y_{i\ell}} (1 - \lambda_{\widetilde{Z}_i\ell})^{1-Y_{i\ell}}$. Because $\widetilde{Z}_i$ is assumed to be unobserved, it is integrated out with respect to its distribution $\mathbb{P}(\widetilde{Z}_i = k \mid \boldsymbol{\pi}_{\widetilde{K}}) = \pi_k > 0$, where $\boldsymbol{\pi}_{\widetilde{K}} = (\pi_1, \ldots, \pi_{\widetilde{K}})^\top$ are population *mixing weights*. Based on N independent observations, the LCM likelihood takes the form of "mixture of Bernoulli products": $\prod_{i=1}^{N} \sum_k \pi_k \mathbb{P}\left(\boldsymbol{Y}_i \mid \widetilde{Z}_i, \{\lambda_{k\ell}\}\right)$.

Any multivariate discrete data distribution can be approximated arbitrarily closely by an LCM with a sufficiently large $\widetilde{K}$ (Dunson and Xing, 2009, Corollary 1) and, up to class relabeling, is *generically identifiable* whenever $L \geqslant 2\lceil \widetilde{K} \rceil + 1$ (Allman et al., 2009, Corollary 5). LCM estimates quantify how the estimated response probability profiles differ by class. Estimation of clusters in finite mixture models often makes use of class indicator $\{\widetilde{Z}_i\}$, for example, by maximizing the plugged-in conditional posterior $\widehat{Z}_i = \arg\max_k \mathbb{P}(\widetilde{Z}_i = k \mid \mathbf{Y}, \widehat{\boldsymbol{\pi}}_{\widetilde{K}})$ where $\widehat{\boldsymbol{\pi}}_{\widetilde{K}}$ estimates the mixing weights, or in a fully Bayesian framework by a

least-square estimate of clusters based on the distance from pairwise co-clustering posterior probabilities $\widehat{\pi}_{ii'} = \mathbb{P}(\widetilde{Z}_i = \widetilde{Z}_{i'} \mid \mathbf{Y})$ (Dahl, 2006).

To impose scientific structure, we introduce binary latent variables to indicate classes. In our motivating example, these latent states indicate responses to each machine. We assume that subject i's class membership $\widetilde{Z}_i$ is defined by a latent state vector $\boldsymbol{\eta}_i \in \mathcal{A}$, where $\mathcal{A} = \{\boldsymbol{\alpha}_k, k = 1, \ldots, \widetilde{K}\} \subseteq \{0,1\}^M$ is a set of M dimensional binary vectors and $\widetilde{K} = |\mathcal{A}|$ ($\leqslant 2^M$) represents the number of distinct latent state patterns. In most applications, $\widetilde{K}$ is unknown. We further link a subject's response probability $\lambda_{i\ell}$ to the subject's latent states $\boldsymbol{\eta}_i$ via $\lambda_{i\ell} = \lambda_\ell(\boldsymbol{\eta}_i)$, where $\lambda_\ell : \mathcal{A} \to [0,1]$ is an unknown function and as we will illustrate below may depend on other parameters. We recover $\lambda_{i\ell} = \lambda_\ell(\widetilde{Z}_i) = \lambda_{\widetilde{Z}_i\ell}$ in traditional LCMs once replacing $\boldsymbol{\eta}_i$ with the corresponding class membership indicator $\widetilde{Z}_i$. Taken together, the likelihood contributed by subject i conditional on her class membership $\mathbb{P}(\boldsymbol{Y}_i \mid \widetilde{Z}_i = k, \{\lambda_{k\ell}\})$ in Section 2.1 is equivalent to:

$$\mathcal{L}_i(\boldsymbol{\alpha}_k, \lambda_\ell(\boldsymbol{\alpha}_k)) = \mathbb{P}(\boldsymbol{Y}_i \mid \boldsymbol{\eta}_i = \boldsymbol{\alpha}_k, \{\lambda_\ell(\boldsymbol{\eta}_i)\}) = \prod_{\ell=1}^{L} \{\lambda_\ell(\boldsymbol{\alpha}_k)\}^{Y_{i\ell}} \{1 - \lambda_\ell(\boldsymbol{\alpha}_k)\}^{1-Y_{i\ell}}. \quad (1)$$

We integrate (1) with respect to the distribution of latent states $\mathbb{P}(\boldsymbol{\eta}_i = \boldsymbol{\alpha}_k \mid \boldsymbol{\pi}_{\widetilde{K}}) = \pi_k > 0$, for $\boldsymbol{\alpha}_k \in \mathcal{A}$ to obtain the likelihood for N independent observations: $\prod_{i=1}^{N} \sum_{\boldsymbol{\alpha}_k \in \mathcal{A}} \pi_k \mathcal{L}_i(\boldsymbol{\alpha}_k, \lambda_\ell(\boldsymbol{\alpha}_k))$.

2.2 *Model Formulation with Scientific Structures for the Motivating Example*

We specify the model for the motivating example in two steps: 1) impose scientific structure upon the actual presence or absence of proteins ($\{\Gamma_{i\ell}\}$) as a function of latent states $\boldsymbol{\eta}_i$, and 2) parameterize the joint distribution of their noisy measurements $\{Y_{i\ell}\}$. The first step is needed to respect existing biological knowledge in the scientific context and the second step characterizes the measurement process.

$\Gamma_{i\ell}$ indicates whether or not subject i mounted an immune response to protein ℓ. Biological knowledge dictates that protein ℓ is present because subject i responded to one or more machines containing protein ℓ. And it is of scientific interest to estimate and classify patients

based on the machine(s) to which they responded. Let the latent states $\boldsymbol{\eta}_i$ indicate which protein complexes ("machines") are present in patient i's class. We impose the scientific structure $\{\Gamma_{i\ell}\}$ as follows:

$$\Gamma_{i\ell} = \Gamma(\boldsymbol{\eta}_i, Q_{\star\ell}) = 1 - \prod_{m=1}^{M} (1 - \eta_{im})^{Q_{m\ell}}, \ell = 1, \ldots, L, \tag{2}$$

where the m-th machine is represented by the m-th row of a M by L binary matrix Q and the ones in $\{Q_{m\ell}, \ell = 1, \ldots, L\}$ indicate which proteins constitute Machine m, for $m = 1, \ldots, M$ ($Q_{\star\ell}$ and $Q_{m\star}$ denote the ℓ-th column and m-th row, respectively). For example, the class of subjects who did not respond to any machine has all zeros in the corresponding row of Γ, which can be seen from $\Gamma_{i\ell} = \Gamma(\mathbf{0}_{M\times 1}, Q_{\star\ell}) = 1 - \prod_{m=1}^{M}(1-0)^{Q_{m\ell}} = 0$, $\ell = 1, \ldots, L$. As another example, suppose protein ℓ_* is in machine m_* ($Q_{m_*\ell_*} = 1$) to which subject i responded ($\eta_{im_*} = 1$). Protein ℓ_* will actually be present in subject i's serum ($\Gamma_{i\ell_*} = 1$) regardless of whether or not the same protein ℓ_* was targeted as a component in another machine $m \neq m_*$. This can be seen from the irrelevant product term in $\Gamma_{i\ell_*} = 1 - (1 - 1)^1 \prod_{m\neq m_*}(1-\eta_{im})^{Q_{m\ell_*}} = 1$. Finally, (2) simplifies to $\boldsymbol{\eta}_i^\top Q_{\star\ell}$ if machines have non-overlapping components represented by orthogonal rows in Q ($Q_{m\star}^\top Q_{m'\star} = 0$ for any $m \neq m'$) as in Figure 1.

$Y_{i\ell}$ represents the observed presence/absence of protein ℓ on the immunoprecipitation gel for patient i. The probability of observing a protein given it is present is its true positive rate or sensitivity. The probability of observing the protein given it is absent is its false positive rate or one minus its specificity. We write this parameterization of response probabilities as

$$\lambda_{i\ell} = \lambda_\ell(\boldsymbol{\eta}_i; Q_{\star\ell}, \boldsymbol{\theta}, \boldsymbol{\psi}) = \theta_\ell^{\Gamma_{i\ell}} (\psi_\ell)^{1-\Gamma_{i\ell}}, \ell = 1, \ldots, L, \tag{3}$$

where $\boldsymbol{\theta} = \{\theta_\ell\}$ and $\boldsymbol{\psi} = \{\psi_\ell\}$ are the true positive rates assumed to be larger than the false positive rates. The sensitivity for a given protein is assumed to be the same regardless from which machine(s) it comes. Importantly, both the sensitivities and specificities are allowed to vary across proteins.

REMARK 1: Model (1,2,3) is related to some existing models proposed in cognitive diagnosis and epidemiology. In general, consider N subjects each responding to L items where $Q_{m\ell} = 1$ means item ℓ requires positive latent state m, otherwise $Q_{m\ell} = 0$. This model, referred to as a *partially latent class model (PLCM)* in disease epidemiology (Wu et al., 2016, PLCM) or *Deterministic In and Noisy Or (DINO) model* in cognitive science (e.g., Templin and Henson, 2006, DINO), needs just one required state ($\{m : Q_{m\ell} = 1\}$) for a positive error-free response $\Gamma_{i\ell} = 1$. Imposing constant and symmetric error rates $\theta_\ell = \psi_\ell$, $\ell = 1, \ldots, L$, gives the *one-layer model* of Rukat et al. (2017). The model can also be viewed as Boolean matrix factorization (BMF, Miettinen et al., 2008) because model (2) is equivalent to $\Gamma_{i\ell} = \vee_{m=1}^{M} \eta_{im} Q_{m\ell}$ where the logical "OR" operator " $\vee$ " outputs one if any argument equals one. The rows in Q are basis patterns for compactly encoding the L dimensional $\Gamma_{i\star}$ vector by $M(\ll L)$ bits in $\boldsymbol{\eta}_i$. BMF further reduces to nonnegative matrix factorization (e.g., Lee and Seung, 1999) $\Gamma = HQ$ where $H = \{\eta_{im}\}$ if Q has orthogonal rows (Figure 1).

Supplementary Material A1.2 presents a general technical formulation of RLCMs that include (2) as a special case. Three specifications define a RLCM: the latent state space ($\mathcal{A}$), design matrix (Γ), and measurement likelihood. Table S1 in the Supplementary Materials summarizes these and other variants of LCMs. Supplementary Materials A1.3 and A1.4 provide more examples and connections to another model-based clustering method for multivariate binary observations (Hoff, 2005). Finally, our connection to general RLCMs makes existing identifiability results available to evaluating the theoretical limit of recovering parameters with an unbounded sample size, which is discussed Supplementary Material A1.5.

2.3 *Priors*

Given M, a Bayesian approach must specify the prior distributions for: the latent states $H = \{\boldsymbol{\eta}_i\}$ that *a priori* cluster subjects; the measurement error parameters $\boldsymbol{\theta}$ and $\boldsymbol{\psi}$; and the

Q matrix if unknown. Since in this application, the number of classes $\mathcal{A}$ is a scientific focus, we treat it as unknown and seek to infer it from the data and prior distribution. Following Miller and Harrison (2017), we specify a prior distribution for $\mathcal{A}$. In our application, identifying a parsimonious set of machines is desirable so we use a prior like the geometric distribution that assigns larger prior probabilities to fewer classes. We also need a prior distribution for the assignment of individuals to clusters. We achieve this by randomly drawing cluster indicators given M. The next step is to draw the latent states in each group of subjects. Since they are also unknown, we specify a prior distribution to encourage sparser binary patterns. We specify the prior distributions for the rest of parameters in Supplementary Material A1.10 and the joint distribution of data $\mathbf{Y} = \{\boldsymbol{Y}_i\}$, the true and false positive rates, Q matrix, and latent state vectors $H = \{\boldsymbol{\eta}_i\}$ in Supplementary Material A1.11.

In the following, we provide the details of the prior specifications that enable inference of the unknown number of classes, the clusters and the unknown latent states for each cluster.

Prior for clustering observations with an unknown number of classes. Though used interchangeably by many authors, we first make a distinction between a "component" that represents one of the true mixture components in the specification of a mixture model (referred to as "classes" in LCMs) and a "cluster" that represents one element in any partition of observations. Let K be the number of mixture components in the *population* and T the number of clusters in the *sample* (Miller and Harrison, 2017).

To establish notation, let $Z_i \in \{1, 2, \ldots, K\}$ be the subject-specific component indicators, $E_z = \{i : Z_i = z\}$ the set of subjects in component j, $\mathcal{C} = \{C_j : |C_j| > 0, j = 1, \ldots, T\}$ the partition of N subjects induced by $\boldsymbol{Z} = \{Z_i, i = 1, \ldots, N\}$; Note the partition $\mathcal{C}$ is invariant to component relabeling. Let $T = |\mathcal{C}|$ be the number of clusters formed by the N *subjects*; it may differ from K ($T \leqslant K$), the number of components for the *population*. Let $\mathcal{C}_{-i} = \{C_j \setminus \{i\} : |C_j \setminus \{i\}| > 0\}$ be the partition of subjects excluding subject i. For

simplicity, let $\mathbf{Y}_C = \{\boldsymbol{Y}_i, i \in C\}$ be the collection of data in a cluster $C \in \mathcal{C}$. Finally, let $\boldsymbol{\eta}_i (\in \{0,1\}^M)$ be the latent state vector for subject $i = 1, \ldots, N$, and $\boldsymbol{\eta}_j^* (\in \{0,1\}^M)$ be the latent state vectors for cluster $j = 1, \ldots, T$.

We specify a prior distribution of partition $\mathcal{C}$ induced by the following three steps that produce samples of cluster indicators $\boldsymbol{Z}$: 1) draw the number of components $K \sim p_K$ where p_K is a probability mass function over positive integers $\{1, 2, \ldots\}$, 2) draw mixing weights $\boldsymbol{\pi}_K \sim \mathsf{Dirichlet}(\gamma, \ldots, \gamma)$ where $\gamma > 0$ is the hyperparameter for symmetric Dirichlet distribution, 3) draw the cluster indicators $Z_i \sim \mathsf{Categorical}\{\boldsymbol{\pi}_K = (\pi_1, \ldots, \pi_K)\}, i = 1, \ldots, N$. Note that though $\widetilde{K} \leqslant 2^M$, K is not upper bounded (unless constrained through the support of p_K). It can be shown that partition $\mathcal{C}$ is *a priori* distributed according to $p(\mathcal{C} \mid \gamma, p_K) = V_N(T) \prod_{C \in \mathcal{C}} \gamma^{(|C|)}$, where $V_N(T) = \sum_{k=1}^{\infty} \frac{k_{(T)}}{(\gamma k)^{(N)}} p_K(k)$, $T = |\mathcal{C}|$ is the number of blocks/partitions for N subjects and by convention $k^{(n)} = k \cdot (k+1) \cdots (k+n-1)$, $k_{(n)} = k \cdot (k-1) \cdots (k-n+1)$, and $k^{(0)} = k_{(0)} = 1$, $k_{(n)} = 0$ if $k < n$ (Miller and Harrison, 2017).

Prior for the cluster-specific latent states $\boldsymbol{\eta}_j^*$. *Pre-specified latent state space $\mathcal{A}$.* In some applications, the latent state space may be known: $\mathcal{A} = \{\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_{\widetilde{K}}\}$. We enumerate the elements in $\mathcal{A}$ by setting cluster-specific latent states to be $\boldsymbol{\eta}_j^* = \boldsymbol{\alpha}_j$, $j = 1, \ldots, \widetilde{K}$. We specify a simple categorical distribution for these distinct latent states with probability parameters $\boldsymbol{\pi}_{\widetilde{K}} = (\pi_1, \ldots, \pi_{\widetilde{K}})$, referred to as mixing weights in finite mixture models. For example, Wu et al. (2016) analyzed data from a childhood pneumonia etiology study to estimate the mixing weights which represent the population fractions of cases caused by different pathogen infections in the lung. They specified $\mathcal{A} = \{\boldsymbol{e}_1, \ldots, \boldsymbol{e}_M, \mathbf{0}_M\}$ among pneumonia *cases* to represent the latent states of lung infection caused by pathogen $1, 2, \ldots, M$ or none-of-the-above and $\boldsymbol{\eta}_i = \mathbf{0}_M$ among observed *controls*. The pre-specification is therefore appealing especially when the scientific interest lies in the estimation of mixing weights $\boldsymbol{\pi}_{\widetilde{K}}$,

one for each pattern of latent states. Absent the uncertainty in $\mathcal{A}$, simpler posterior sampling algorithms result.

However, in other applications, pre-specifying $\mathcal{A} \subsetneqq \{0,1\}^M$ ignores its uncertainty in estimating clusters. For example, $\mathcal{A}$ is unknown because of lack of strong prior knowledge about the machines to which the patients respond. Analysts may conservatively specify $\mathcal{A} = \{0,1\}^M$, fit the model and keep the most important patterns. However, latent states $\boldsymbol{\eta}_i = \boldsymbol{\alpha}_{Z_i}$ then take value from a space that grows exponentially in size with M (e.g., $M = 30$ in Wu et al. (2016)). One can fit a model like (2) to infer $\pi_k, k = 1, \ldots, \widetilde{K}(= 2^M)$. However, many marginal posterior distributions of mixing weights $[\pi_k \mid \boldsymbol{Y}]$, $k = 1, \ldots, 2^M$, may concentrate near zero, but not exactly zero. Important elements in $\mathcal{A}$ are commonly selected by *ad hoc* post-processing of the posterior samples of $\{\pi_k\}$, for example, by requiring the posterior probability of exceeding a low threshold τ that is deemed meaningful in the application (e.g., greater than 0.05).

Unknown latent state space $\mathcal{A}$. Absent knowledge of $\mathcal{A}$, we specify the prior distribution for the component-specific parameters $H^* = \{\eta^*_{jm}\}$ so that we regularize $\boldsymbol{\eta}^*_j$ towards sparser binary patterns:

$$\text{hyperprior for probability of positive state m}: \quad p_m \mid \alpha_1, \alpha_2 \sim \text{Beta}(\alpha_1\alpha_2/M, \alpha_2), \tag{4}$$

$$\text{prior for latent states}: \quad \eta^*_{jm} \mid p_m \sim \text{Bernoulli}(p_m), j = 1, \ldots, T, \tag{5}$$

for $m = 1, \ldots, M$. The two-step prior induces a marginal prior $[H^* \mid \alpha_1, \alpha_2]$ upon integrating over $\{p_m\}$ (see Supplementary Material A1.6). Supplementary Material A1.7 extends the prior on H^* to $M = \infty$ and connects it to Indian Buffet Process (Ghahramani and Griffiths, 2006). In what follows, α_2 is set to 1 which offers good clustering results in simulations and data analyses. Finally in applications where no pooling across j is needed, one can set $p_m = 0.5$ in (5) to specify a uniform distribution over all possible patterns over $\mathcal{A} = \{0,1\}^M$.

REMARK 2: (4) and (5) may generate a random draw of identical $\boldsymbol{\eta}^*_j = \boldsymbol{\eta}^*_{j'}$ for some $j \neq$

$j' = 1, \ldots, T$ where the equality holds element-wise. Because we are interested in estimating distinct $\boldsymbol{\eta}_j^*$'s that represent distinct latent states in particular scientific contexts, we will merge such clusters j and j' into one, referred to as a "scientific cluster"; We denote the resultant merged clusters by $\widetilde{\mathcal{C}}$. We also denote the *unique* values in $H^* = \{\boldsymbol{\eta}_j^*, j = 1, \ldots, T\}$ by $\widetilde{H}^* = \{\widetilde{\boldsymbol{\eta}}_j^*, j = 1, \ldots, \widetilde{T}\}$ where by definition $\widetilde{T} \leqslant T \leqslant K$. Supplementary Materials A1.8 and A1.9 further remark on the induced priors on the partitions $\mathcal{C}$ and $\widetilde{\mathcal{C}}$.

3. Posterior Inference

We develop inferential procedures to address the following three questions: 1) how many scientific clusters ($\widetilde{T}$) in the sample (data); 2) what are the unique latent states $\{\widetilde{\boldsymbol{\eta}}_j^*, j = 1, \ldots, \widetilde{T}\}$ in the sample; and 3) what are the subjects' latent states $\boldsymbol{\eta}_i$ and the scientific clusters $\widetilde{\mathcal{C}}$.

Given Q and $\boldsymbol{\theta}$, $\boldsymbol{\psi}$, model (1-3) as a mixture model has *discrete* component-specific parameters $\boldsymbol{\eta}_i \in \mathcal{A} \subseteq \{0, 1\}^M$. This is to be contrasted with mixture models with a continuous distribution from which component parameters are drawn and differ from one another with probability one. When sampled conditional on other parameters, the discrete mixture component parameters $\{\boldsymbol{\eta}_j^*, j = 1, \ldots, T\}$ may be duplicated. At each MCMC iteration, we *post-process* the posterior samples by merging clusters in $\mathcal{C}$ associated with identical $\boldsymbol{\eta}_j^*$ to obtain scientific clusters with distinct latent states. Given M, no more than 2^M distinct latent state patterns $\widetilde{\boldsymbol{\eta}}_j^*$ results after merging. More generally, for inference based on mixture of finite mixture (MFM) models with *discrete* component parameters, p_K is a prior for K (not $\widetilde{K}$) over all non-negative integers and offers technical convenience of removing the otherwise hard constraint $K = \widetilde{K} \leqslant 2^M$ (would be so if we force distinct latent states in the prior). This greatly simplifies the design of posterior algorithms.

We use Markov chain Monte Carlo (MCMC) algorithm for posterior inference which by design upon convergence simulates samples that approximate the joint posterior distribution

of any functions of unknown parameters and latent variables (Gelfand and Smith, 1990): $(\boldsymbol{Z}, H^*, Q, \boldsymbol{\theta}, \boldsymbol{\psi}, \alpha_1)$. The posterior algorithms also sets an upper bound $M^\dagger$ for M and at each iteration may produce less than $M^\dagger$ effective machines. All model estimations are performed by an `R` package "`rewind`", which is freely available at https://github.com/zhenkewu/rewind. Given our focus on estimating clusters, we choose to directly sample $\mathcal{C}$ from its posterior without the need for considering component labels or empty components. See Supplementary Material A2 for more details of the sampling algorithms and convergence checks as well as a discussion about information from data that updates the clusters $\mathcal{C}$. Supplementary Material A3 presents posterior summaries of co-clustering and latent states.

4. Results

We illustrate the utility of RLCM on both simulated and real data where Q is unknown. First, we assess the performance of RLCM on cluster estimation under simulation scenarios corresponding to varying levels of measurement error, dimension, sparsity level of each machine, sample size and mixing weight. Using data simulated under the assumed RLCM and realistic deviations from it, the proposed Bayesian analyses performs clustering as well as or better than common alternative binary-data clustering methods (including two likelihood-based methods that include the RLCM simulation truths as special cases). We first analyze a single randomly generated data set to highlight differences among the methods. We then evaluate frequentist performance of Bayesian RLCM in cluster estimation and contrast with the alternatives. Finaly, protein data from scleroderma patients are analyzed.

4.1 *Simulated Examples to Study Model Performance*

Simulation 1: More accurate clustering through feature selection in scientifically structured classes. We set $N = 50$, $L = 100$ and $M = 3$. We randomly generate a matrix Q (M by L) where each row has on average $s = 20\%$ non-zero elements: $Q_{m\ell} \overset{i.i.d}{\sim} \mathsf{Bernoulli}(0.2), \ell =$

$1, \ldots, L$. In the rare event where a randoml $Q \notin \mathcal{Q}$ (identifiability constraint (S7) in Supplementary Material A1.5), we randomly permute pairs of elements in $Q_{m\star}$ until $Q \in \mathcal{Q}$. We draw latent states for each observation independently according to $\boldsymbol{\eta}_i \overset{d}{\sim} \mathsf{Categorical}\,(\mathcal{A}; \boldsymbol{\pi}_0 = \boldsymbol{\pi}_b)$ where $\boldsymbol{\pi}_0 = \{\mathbb{P}(\boldsymbol{\eta}_i = (0,0,0),(1,0,0),(0,1,0),(1,1,0),(0,0,1),(1,0,1),(0,1,1),(1,1,1))\}$, and $\boldsymbol{\pi}_b = (1/6, 1/6, 1/6, 1/6, 1/12, 1/12, 1/12, 1/12)$. Because we focus on model (2), we assume the response probabilities shift between two levels $\theta_\ell = 0.8$ and $\psi_\ell = 0.15$. The distinct subsets of features where shifts occur define eight classes $\widetilde{K} = 8 = (2^M)$, which upon enumeration by observation gives an N by L design matrix Γ. Figure 2 shows the resulting data $\mathbf{Y}$, the design matrix Γ, as well as the clusters obtained using complete-linkage, Hamming distance hierarchical clustering (HC), standard eight-class Bayesian latent class analysis (LCA, e.g., Garrett and Zeger (2000)), subset clustering analysis (Hoff, 2005) and our Bayesian RLCM with unknown number of clusters fitted with truncation level $M^\dagger = 5$. In this setting, HC is sensitive to noise and tends to split a true cluster or group observations from different true clusters. Unlike the others, the Bayesian RLCM automatically selects and filter subsets of features that distinguish eight classes (through scientific structures in (2)) hence has superior clustering performance producing clusters that agrees quite well with the truth. This relative advantage of Bayesian RLCM persists under data replications (see Simulation 2).

In contrast to traditional all-feature clustering methods, through the inference of all-zero columns of design matrix $\Gamma_{\star\ell} = \mathbf{0}$, Bayesian RLCM removes irrelevant features hence reduces the impact of noise at less important features and in the current setting has better clustering performance (see Supplementary Material A4 for additional simulations on this point).

Simulation 2: Assess clustering performance under various parameter settings. We simulated $R = 60$ replication data for each of $1,920$ combinations of (#features, sample size, true positive rate, false positive rate, mixing weights, sparsity level of the rows of Q):

$(L, N, \theta_0, \psi_0, \boldsymbol{\pi}_0, s) \in \{50, 100, 200, 400\} \otimes \{50, 100, 200\} \otimes \{0.8, 0.9\} \otimes \{0.05, 0.15\} \otimes \{\boldsymbol{\pi}_a = (\frac{1}{8}, \ldots, \frac{1}{8}), \boldsymbol{\pi}_b = (\frac{1}{6}, \ldots, \frac{1}{6}, \frac{1}{12}, \ldots, \frac{1}{12})\} \otimes \{10\%, 20\%\}$. The parameter values are designed to mimic what would be expected in the motivating example. We use adjusted Rand index (aRI, Hubert and Arabie, 1985) to assess the agreement between two clusterings, e.g,. the estimated and the true clusters. aRI is defined by $\mathsf{aRI}(\mathcal{C}, \mathcal{C}') = \frac{\sum_{r,c}\binom{n_{rc}}{2} - \left[\sum_r \binom{n_{r\cdot}}{2} \sum_c \binom{n_{\cdot c}}{2}\right]/\binom{N}{2}}{0.5\left[\sum_r \binom{n_{r\cdot}}{2} + \sum_c \binom{n_{\cdot c}}{2}\right] - \left[\sum_r \binom{n_{r\cdot}}{2} \sum_c \binom{n_{\cdot c}}{2}\right]/\binom{N}{2}}$, where n_{rc} represents the number of observations placed in the rth cluster of the first partition $\mathcal{C}$ and in the cth cluster of the second partition $\mathcal{C}'$, $\sum_{r,c}\binom{n_{rc}}{2} (\leqslant 0.5\left[\sum_r \binom{n_{r\cdot}}{2} + \sum_c \binom{n_{\cdot c}}{2}\right])$ is the number of observation pairs placed in the same cluster in both partitions and $\sum_r \binom{n_{r\cdot}}{2}$ and $\sum_c \binom{n_{\cdot c}}{2}$ calculates the number of pairs placed in the same cluster for the first and the same cluster for second partition, respectively. aRI is bounded between -1 and 1 and corrects for chance agreement. It equals one for identical clusterings and is on average zero for two random partitions; larger values indicate better agreements.

The performance of Bayesian RLCM of recovering the true clusters varies by the sparsity level (s) in each machine, level of measurement errors (θ_ℓ, ψ_ℓ), mixing weights and sample sizes (N) (the leftmost boxes in groups of four in Figure 3). Firstly, clustering performance improves by increasing the sparsity level in each machine from $s = 10\%$ to 20% (compare the 1st and 3rd, 2nd and 4th RLCM boxplots with solid lines in each panel of Figure 3). In the context of our motivating example, given a fixed number of protein landmarks L, patients will be more accurately clustered if each machine comprises more component proteins. This observation is also consistent with simulation studies conducted in the special case of $Q = I_L$ (Hoff, 2005, Table 1). For a given s, a larger L means a larger number of relevant features per machine and leads to better cluster recovery. In Figure S2 of Supplementary Materials (Figure 3 here shows its 8 subplots), increasing L from 50 to 400 (from the top to the bottom row), the mean aRI (averaged over replications) increases, e.g., in the first column, from 0.7 to 0.98 at the sparsity level $s = 10\%$, 0.88 to 0.99 under $s = 20\%$. Secondly, more

accurate clustering results under larger discrepancies between θ_ℓ and ψ_ℓ. The aRI averaged over replications is higher under $\psi_0 = 0.05$ than $\psi_0 = 0.15$ over all combinations of other parameters. Thirdly, under non-uniform mixing weights $\boldsymbol{\pi}_0 = \boldsymbol{\pi}_b$, Bayesian RLCM performs similarly or slightly worse than under uniform mixing weights ($\boldsymbol{\pi}_a$). Finally we observe mixed relative performances at distinct sample sizes as a result of two competing factors as the sample size increases: more precise estimation of measurement error parameters that improve clustering and a larger space of clusterings.

The Bayesian RLCM on average most accurately recovers the clusters compared to three common alternatives. In Figure 3, Bayesian RLCM produces the highest aRIs (boxes with solid lines) compared to others (boxes with dotted lines) and are in many settings perfect. For example, under false positive rate ($\psi_0 = 0.05$) the ratio of the mean aRIs (averaged over replications) for Bayesian RLCM relative to subset clustering is 2.06, 2.04, 1.88, 1.71 for the sample-size-to-dimension ratios $N/P = 1, 0.5, 0.25, 0.125$, respectively. As another example, under a higher $\psi_0 = 0.15$, the relative advantage of Bayesian RLCM to HC narrows as shown by the smaller aRI ratios 1.23, 1.62, 1.49, 1.16.

[Figure 2 about here.]

We remark on the performance of the other three methods. Over all parameter settings investigated here, the traditional LCA performed worst in the recovery of true clusters (aRI < 0.68). The advantage of RLCM comes from the regularization of estimated response probability profiles towards a scientific structure that improves finite-sample clustering performance. We also obtain better clustering performance of RLCMs compared to LCM for data simulated from LCMs with realistic deviations from RLCMs (not shown here). The likelihood function of subset clustering is a special case of the RLCM that assumes a non-parsimonious $Q = I_L$ and therefore loses power for detecting clusters compared to RLCM that estimates a structured Q with multiple non-zero elements in its rows. HC is fast and recovers the

true clusters reasonably well (ranked second or first among the four methods for more than two thirds of the parameter settings here; See Figure S3 in Supplementary Materials). The performance of HC is particularly good under a low level of measurement errors ($\psi_0 = 0.05$) and a large number of relevant features per machine and sometimes performs much better than traditional LCA and subset clustering (e.g., $L = 200$, $N = 50$, $\theta_\ell = 0.8$, $\psi_\ell = 0.05$ in Figure S2, Supplementary Materials). The HC studied here requires a pre-specified number of clusters to cut the dendrogram at an appropriate level and produces clusters that require separate methods for uncertainty assessment (e.g., Suzuki and Shimodaira, 2006). The proposed Bayesian RLCM, in contrast, enjoys superior clustering performance and provides direct internal assessment of the uncertainty of clusters and measurement error parameters through the posterior distribution.

[Figure 3 about here.]

4.2 *Analysis of GEA Data*

GEA Data, Preprocessing and Informative Priors. The goal is to estimate autoimmune disease patient clusters via reconstructing components of protein complexes. Autoantibodies are the immune system's response to specific cellular protein complexes or "machines". We seek to identify components of the machines and to quantify the variations in their occurrence among individuals. The binary responses $\boldsymbol{Y}_i$ indicate the observed presence of autoantibodies at equi-spaced molecular weight landmarks as produced via a preprocessing method (Wu et al., 2019) implemented using publicly available software `R` package "`spotgear`" (https://github.com/zhenkewu/spotgear). We ran 4 GEA gels, each loaded with immunoprecipitations (IPs) performed using sera from 19 different patients, and one reference lane. All sera were from scleroderma patients with cancer, and were all negative for the three most common autoantibodies found in scleroderma (anti-RNA polymerase III, anti-topoisomerase I, and anti-centromere). The IPs were loaded in random order on each gel;

the reference sample is comprised of known molecules of defined sizes (molecular weights) and was always loaded in the first lane. The left panel in Figure 4 shows for each sample lane (labeled in the left margin; excluding the reference lanes) the binary responses indicating the observed presence or absence of autoantibodies at $L = 50$ landmarks.

Patients differ in their antibody protein presence or absence patterns at the protein landmarks. Eleven out of $L = 50$ aligned landmarks are absent among the patients tested. The rest of the landmarks are observed with prevalences between 1.3% and 94.7%. We apply two-parameter RLCM (2) with unknown $M(< L/2 = 50)$ and Q, $\boldsymbol{\theta}$, $\boldsymbol{\psi}$. The GEA technologies are known to be highly specific and sensitive for nearly all proteins studied in this assay so we specify the priors for the true and false positive rates by $\mathsf{Beta}(a_{\theta\ell}, b_{\theta\ell})$ and $\mathsf{Beta}(a_{\psi\ell}, b_{\psi\ell})$, $\ell = 1, \ldots, L$ respectively. We set $a_{\theta\ell} = 9$, $b_{\theta\ell} = 1$, $a_{\psi\ell} = 1$, $b_{\psi\ell} = 99$ and conducted sensitivity analyses varying these hyperparameter values. Because proteins of distinct weights may have systematically different response probabilities, we choose not to share measurement error rates across dimension. In our analysis, we sampled many Q values across the iterations of the MCMC. Because the interpretation of $\boldsymbol{\eta}_i$ depends on the row patterns in Q, we condition on the least square clustering ($\widehat{\mathcal{C}}^{(LS)}$) and refit the model to obtain the least square Q (Section 3). We also draw posterior samples of α_1 for inference.

In this application, the scientists had previously identified and independently verified through additional protein chemistry the importance of a small subset of protein bands in determining clusters among a subset of subjects. They proposed that these subjects should be grouped together. We therefore fitted the Bayesian RLCM without further splitting these partial clusters $\mathcal{C}^{(0)}$ so that the number of scientific clusters visited by the MCMC chain has an upper bound $\widetilde{T}^{(b)} \leqslant |\mathcal{C}^{(0)}| + N - \sum_{j=1}^{|\mathcal{C}^{(0)}|} C_j^{(0)}$, where $C_j^{(0)}$ counts the number of observations in the initial cluster j. We fitted models and compared the results under multiple working truncation levels $M^\dagger = 8, 9, \ldots, 15$ and obtained identical clustering results.

GEA Results. Figure 4 shows: the observations grouped by the RLCM-estimated clusters (not merged) $\widehat{\mathcal{C}}^{(LS)}$ (left), the estimated Q-matrix $\widehat{Q}(\widehat{\mathcal{C}}^{(LS)})$ (right), and the conditional posterior probabilities of the machines $\mathbb{P}(\eta_{im} = 1 \mid \widehat{\mathcal{C}}^{(LS)}, \widehat{Q}(\widehat{\mathcal{C}}^{(LS)}), \mathbf{Y})$ (middle).

The matrix Q is estimated from the observed marginal associations (positive or negative) among the protein landmarks. Landmark protein pairs observed with *positive* association tend to be placed in the same estimated machine. For example, Landmarks 4, 7 and 8 appear together in Machine 5. Subjects either have all three landmarks or none at all, which induces strong positive pairwise associations among these landmarks. Indeed, the estimated log odds ratio (LOR) is 3.13 (standard error 1.16) for Landmark 4 versus 7, 2.21 (s.e., 0.98) for Landmark 4 versus 8, and 2.92 (s.e. 1.2) for Landmark 7 versus 8. The observed *negative* marginal associations between two landmarks suggest existence of machines with discordant landmarks. For example, Landmarks 10 and 27 are rarely estimated to be present or absent together in a subject as a result of 1) estimated machines with discordant landmarks and 2) subject-specific machine assignments. First, the model estimated that Landmark 10 (in Machine Set A: 1, 3 and 4) belongs to machines not having Landmark 27 (it is in Machine Set B: 2). Second, with high posterior probabilities, most observations have machines from one of, not both Set A and B hence creating discordance (high posterior probability $\mathbb{P}(\Gamma_{i,10} \neq \Gamma_{i,27} \mid \mathbf{Y})$). In the presence of observation errors, strong negative marginal association results (observed LOR for Landmark 10 versus 27: -1.98, s.e. 0.8).

[Figure 4 about here.]

Our algorithm also directly infers the number of scientific clusters in the data given an initial partial clustering $\mathcal{C}^{(0)}$. The marginal posterior of the number of scientific clusters $\widetilde{T}$ can be approximated by empirical samples of $\{\widetilde{T}^{(b)}\}$ which result in a posterior median of 12 (95% credible interval: $(8, 16)$; Figure S4 in Supplementary Materials). The advantage of Bayesian RLCM is the posterior inference about both the clusters and the distinct latent state

variables $\boldsymbol{\eta}_i$ interpreted based on the inferred Q matrix. The middle panel of Figure 4 shows that clusters differ in their posterior probabilities of having each of the estimated machines. Among 76 subjects analyzed, 23 of them have greater than 95% posterior probabilities of having both Machine 4 and 6. A group of seven observations are enriched with Machine 4 and 7 which as expected from the raw band patterns have distinctive combination of Landmarks 35, 40 and 49 (33, 27 and 18 kDa bands, respectively). Such inference about $\boldsymbol{\eta}_i$ is not available to us based on hierarchical clustering or traditional latent class models.

We performed posterior predictive checking to assess model fit (Gelman et al., 1996). At each MCMC iteration, given the posterior sample of model parameters (without conditioning on the best clustering $\widehat{C}^{(LS)}$ or the best $\widehat{Q}$), we simulated a data set of the same size as the original set. For each replicated data set, we compute the marginal means and marginal pairwise log odds ratios (0.5 adjustment for zero counts). Across all replications, we compute the 95% posterior predictive confidence intervals (PPCI) defined by the 2.5% and 97.5% quantiles of the PPD. All the observed marginal means are covered by their respective PPCIs; The 95% PPCIs cover all but 24 of $\binom{L}{2} = 1,225$ landmark pairs of observed pairwise log odds ratios (see Figure S6 and S7 in Supplementary Materials). The proposed model adequately fits the GEA data. Supplementary Materials A5 provides additional results, model interpretations for model fits without a partial cluster $\mathcal{C}^{(0)}$ as well as potential improvements.

5. Discussion

Modern scientific technologies give rise to measurements of varying precision and accuracy that are better targeted at the underlying state variables than ever before. In this paper we have focused on finite-sample Bayesian inference of restricted latent class model for analyzing multivariate binary data in the presence of between-class differential errors. The primary advantage of such models lies in their expressive characterization of the between-

class differential errors structured to respect specific scientific context about the biological and measurement processes. Using simulations and real data analysis, we studied the clustering of observations with an unknown number of clusters, uncertainty assessment of the clustering and the prediction of individual latent states. We develop and apply a novel Markov chain Monte Carlo (MCMC) algorithm for Bayesian RLCMs. The proposed method addresses inferential issues unique to mixture models with discrete component parameters and jointly infers the number of clusters, the design matrix Γ and other model parameters. We have compared the proposed method with variants of latent class models through their specifications in Table S1 in Supplementary Materials and illustrated its advantage through simulations relative to three commonly used binary-data clustering. Finally, viewed from regularization perspective, in the scleroderma example, the inferential procedure automatically selects subsets of features for each latent class and filters them through a low-dimensional model that shrinks class-specific response probability estimates toward one that represents the scientific structure and improves our ability to accurately estimate clusters. We have also implemented an extension to settings where some subjects' latent classes are known or important prior knowledge about differential measurement accuracy is available from external sources.

RLCMs decompose the variation among multivariate binary responses into structure that reflects prior scientific knowledge and stochastic variation without a known explanation. In our motivating example, it is certainly likely that there is some variability related to the vagaries of the measurement assay. However, it is also highly likely that there are systematic biological and biochemical processes not included in the structural part because they are unknown to us today. RLCM analyses can be a useful tool in the effort to uncover the unknown structure. One approach would be to show that the latent classes are diagnostic of specific diseases. Another is that we might uncover a novel mechanism by defining distinct

patterns of the same autoantigen machine in patients with the same disease or potentially in patients with different diseases that target the same machines. Though the present paper focused on an example in medicine, the developed method and algorithms apply to many problems in psychology and epidemiology (see Supplementary Material A1.1).

We are currently studying a few potentially-useful extensions. First, nested partially LCMs (Wu et al., 2017) incorporate local dependence and multiple sensitivity parameters that would improve the utility of Bayesian RLCMs. Second, because the algorithm involves iterating over subjects to find clusters, the computational time increases with the number of subjects N. Divide-Cluster-Combine schemes that estimate clusters in subsamples which are then combined may improve the computational speed at the expense of the approximation introduced by the multi-stage clustering (Ni et al., 2018). Finally, in applications where the clustering of multivariate binary data comprises an important component of a hierarchical Bayesian model with multiple components, the posterior uncertainty in clustering propagates into other parts of the model and can be integrated into posterior inference of other model parameters (e.g., Jacob et al., 2017).

Supplementary Materials

The supplementary materials contain referenced remarks, figures, a table and further technical details, e.g., on identifiability and sampling algorithms, as well as additional simulations and extended data analysis results.

Acknowledgment

The research is supported in part by a gift from the Jerome L. Greene Foundation and by the Patient-Centered Outcomes Research Institute (PCORI) Award (ME-1408-20318), National Institutes of Health (NIH) grants R01AR073208, P30AR070254 and P30CA046592 (ZW: National Cancer Institute Cancer Center Support Grant Development Funds from Rogel

Cancer Center). We also thank Gongjun Xu, Peter Hoff and Jian Kang for their insightful comments.

References

Allman, E. S., Matias, C., and Rhodes, J. A. (2009). Identifiability of parameters in latent structure models with many observed variables. *The Annals of Statistics* pages 3099–3132.

Chen, Y., Culpepper, S. A., Chen, Y., and Douglas, J. (2017). Bayesian estimation of the dina q matrix. *Psychometrika* .

Dahl, D. B. (2006). Model-based clustering for expression data via a dirichlet process mixture model.

Dunson, D. and Xing, C. (2009). Nonparametric bayes modeling of multivariate categorical data. *Journal of the American Statistical Association* **104,** 1042–1051.

Garrett, E. and Zeger, S. (2000). Latent class model diagnosis. *Biometrics* **56,** 1055–1067.

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

Gelman, A., Meng, X.-L., and Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica* **6,** 733–760.

Ghahramani, Z. and Griffiths, T. L. (2006). Infinite latent feature models and the indian buffet process. In *Advances in neural information processing systems*, pages 475–482.

Goodman, L. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika* **61,** 215–231.

Hoff, P. D. (2005). Subset clustering of binary sequences, with an application to genomic abnormality data. *Biometrics* **61,** 1027–1036.

Hubert, L. and Arabie, P. (1985). Comparing partitions. *Journal of classification* **2,** 193–218.

Jacob, P. E., Murray, L. M., Holmes, C. C., and Robert, C. P. (2017). Better together? statistical learning in models made of modules. *arXiv preprint arXiv:1708.08719* .

Jain, S. and Neal, R. M. (2004). A split-merge markov chain monte carlo procedure for the dirichlet process mixture model. *Journal of Computational and Graphical Statistics* **13,** 158–182.

Junker, B. W. and Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement* **25,** 258–272.

Lazarsfeld, P. F. (1950). *The logical and mathematical foundations of latent structure analysis*, volume IV, chapter The American Soldier: Studies in Social Psychology in World War II, pages 362–412. Princeton, NJ: Princeton University Press.

Lee, D. D. and Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature* **401,** 788–791.

Miettinen, P., Mielikäinen, T., Gionis, A., Das, G., and Mannila, H. (2008). The discrete basis problem. *IEEE Transactions on Knowledge and Data Engineering* **20,** 1348–1362.

Miller, J. W. and Harrison, M. T. (2017). Mixture models with a prior on the number of components. *Journal of the American Statistical Association* pages 1–17.

Ni, Y., Müller, P., Diesendruck, M., Williamson, S., Zhu, Y., and Ji, Y. (2018). Scalable Bayesian Nonparametric Clustering and Classification. *ArXiv e-prints* .

Orito, H., Kaji, K., Komura, K., Takehara, K., Fujimoto, M., Hasegawa, M., Kondo, M., Matsushita, T., Hamaguchi, Y., Saito, Y., Ogawa, F., Yanaba, K., Itoh, M., Seishima, M., Sato, S., and Asano, Y. (2006). Identification of a novel autoantibody reactive with 155 and 140kDa nuclear proteins in patients with dermatomyositis: an association with malignancy. *Rheumatology* **46,** 25–28.

Rosen, A. and Casciola-Rosen, L. (2016). Autoantigens as partners in initiation and

propagation of autoimmune rheumatic diseases. *Annual review of immunology* **34,** 395–420.

Rukat, T., Holmes, C. C., Titsias, M. K., and Yau, C. (2017). Bayesian boolean matrix factorisation. In *International Conference on Machine Learning*, pages 2969–2978.

Suzuki, R. and Shimodaira, H. (2006). Pvclust: an R package for assessing the uncertainty in hierarchical clustering. *Bioinformatics* **22,** 1540–1542.

Templin, J. L. and Henson, R. A. (2006). Measurement of psychological disorders using cognitive diagnosis models. *Psychological methods* **11,** 287.

Vermunt, J. K. and Magidson, J. (2002). Latent class cluster analysis. *Applied latent class analysis* **11,** 89–106.

Wu, Z., Casciola-Rosen, L., Shah, A. A., Rosen, A., and Zeger, S. L. (2019). Estimating autoantibody signatures to detect autoimmune disease patient subsets. *Biostatistics* **20,** 30–47.

Wu, Z., Deloria-Knoll, M., Hammitt, L. L., and Zeger, S. L. (2016). Partially latent class models for case–control studies of childhood pneumonia aetiology. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **65,** 97–114.

Wu, Z., Deloria-Knoll, M., and Zeger, S. L. (2017). Nested partially latent class models for dependent binary data; estimating disease etiology. *Biostatistics* **18,** 200.

Xu, G. and Shang, Z. (2018). Identifying latent structures in restricted latent class models. *Journal of the American Statistical Association* **0,** 1–12.

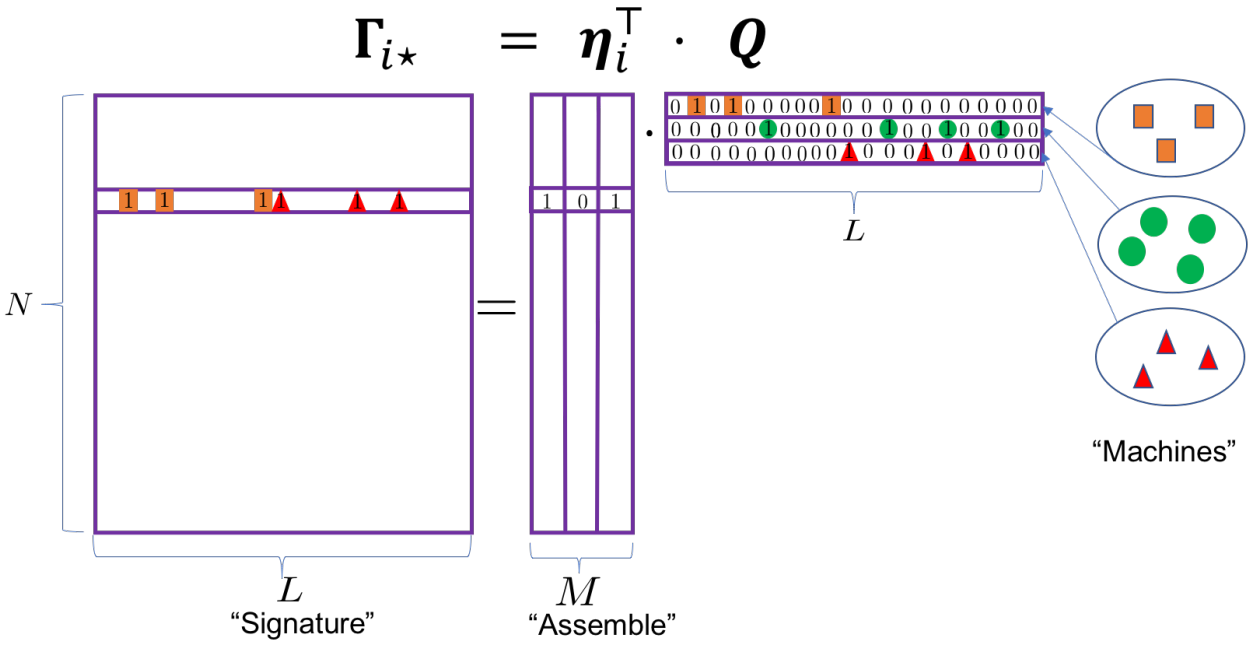

Figure 1: Binary matrix factorization generates autoantibodies that are further subject to misclassification. The hypothetical individual has latent states $\boldsymbol{\eta}_i = (1, 0, 1)^\top$ and is expected to mount immune responses against six antigens in Machines 1 and 3. The expected antibodies $\Gamma_{i\star} = \boldsymbol{\eta}_i^\top Q$ are produced against three orthogonal machines with 3, 4 and 3 landmark proteins, respectively.

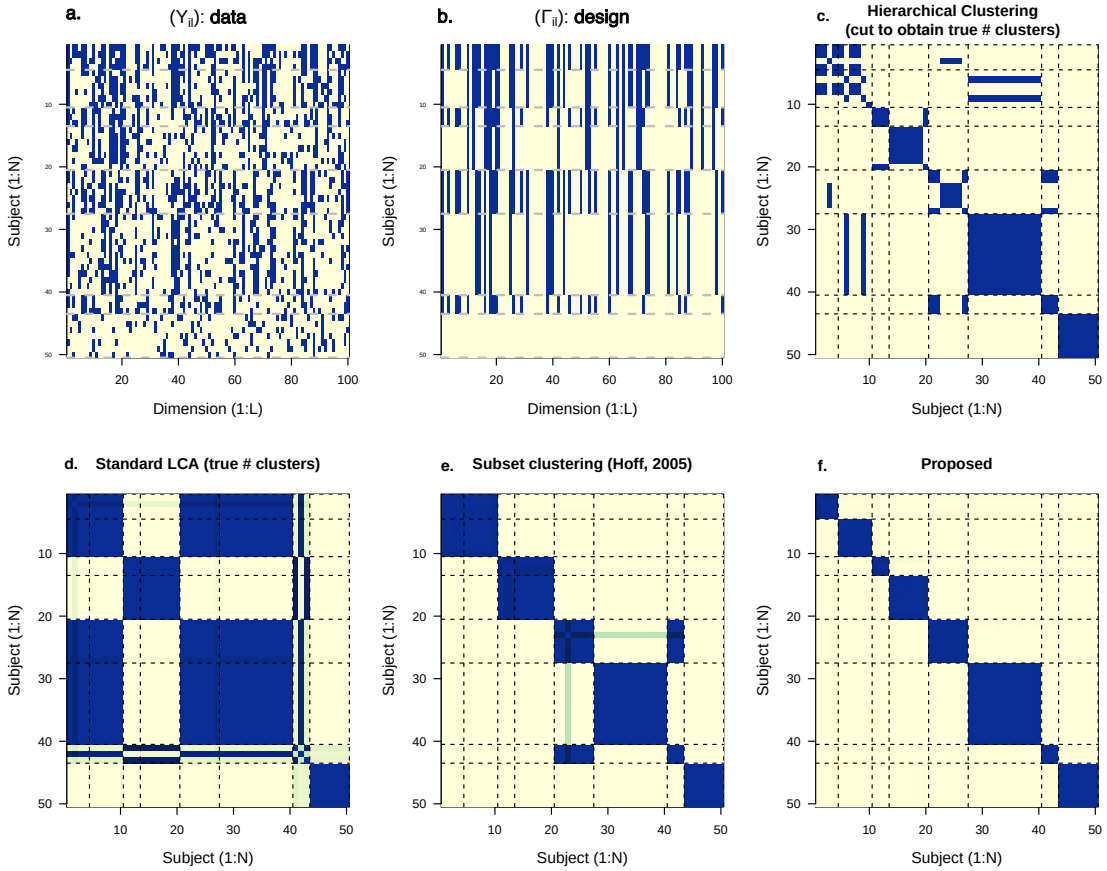

Figure 2: In the 100-dimension multivariate binary data example (a), the eight classes differ with respect to subsets of measured features (b). In (c) HC, we indicate co-clustering by filled cells. The true clusters are separated (dashed grids) and ordered according to the truth; (d, e, f): For Bayesian LCA, RLCM and subset clustering (Hoff, 2005), we plot the posterior co-clustering probability matrix $\{\widehat{\pi}_{i,i'}\}$ for N observations. Filled blocks on the main diagonal only indicate perfect recovery of the true clusters; Blank cells within the main diagonal blocks indicate true cluster being split and blue cells in the off-diagonal blocks indicate two observations being incorrectly co-clusted. Bayesian restricted latent class analysis accounts for measurement errors, selects the relevant feature subsets and filters the subsets by a low-dimensional model (2) and therefore yields superior clustering results.

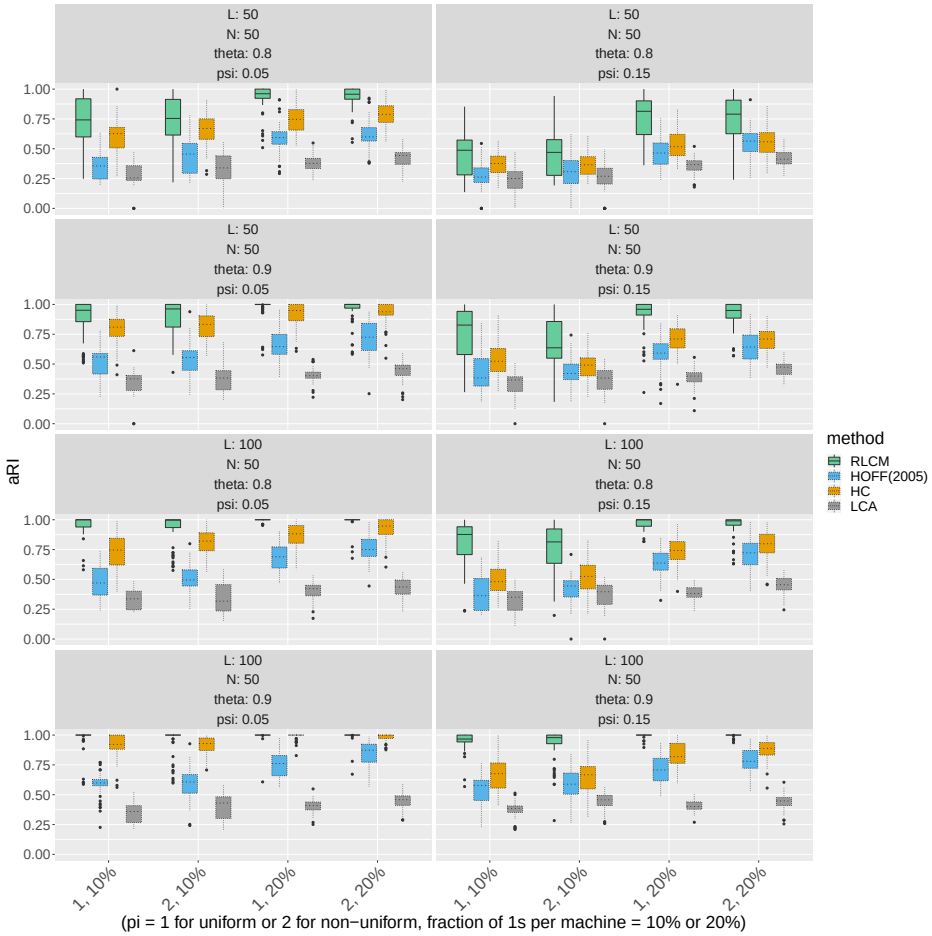

Figure 3: Based on $R = 60$ replications for each parameter setting, from the left to the right in each group of four boxplots, Bayesian RLCM (boxplots with solid lines) most accurately recovers the true clusters compared to subset clustering (Hoff, 2005) hierarchical clustering (HC) and traditional Bayesian latent class analysis (LCA). See Figure S2 in Suppmentary Materials for an expanded version over more parameter settings.

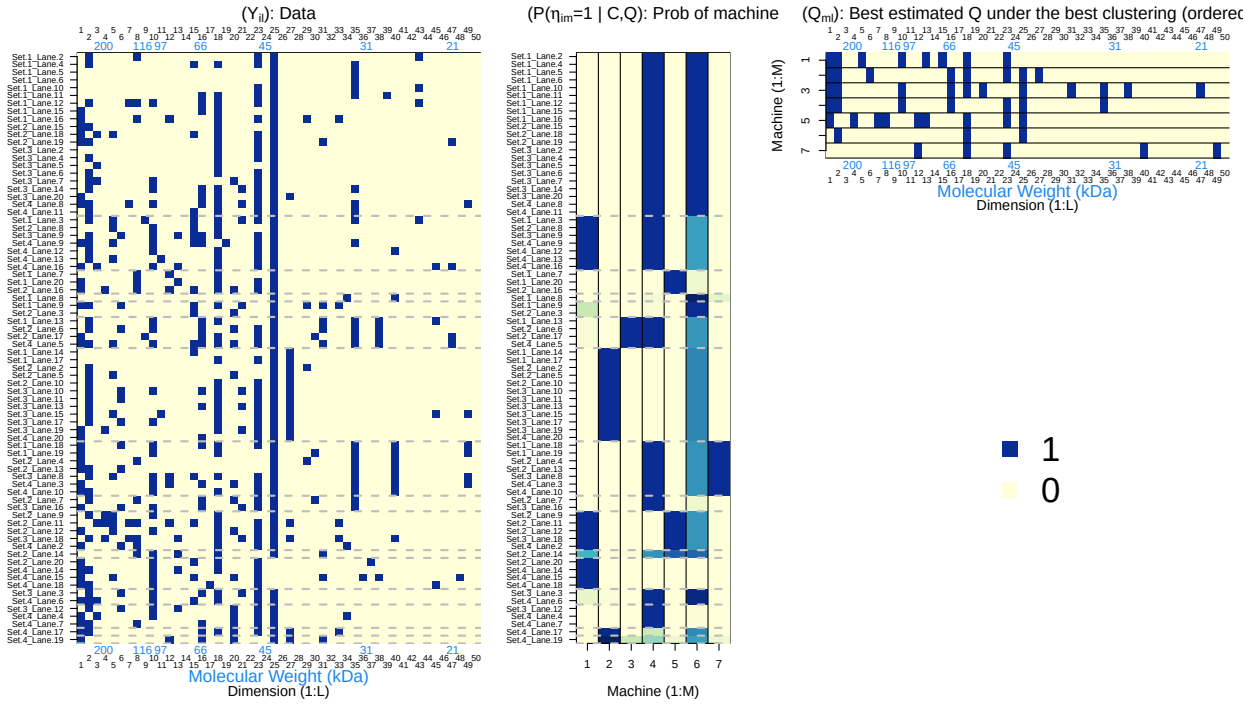

Figure 4: Results for GEA data. *Left*: Aligned data matrix for band presence or absence; row for 76 serum lanes, reordered into optimal estimated clusters (not merged) $\widehat{\mathcal{C}}^{(LS)}$ separated by gray horizontal lines "—"; columns for $L = 50$ protein landmarks. A blue vertical line "|" indicates a band; *Middle*: lane-machine matrix for the probability of a lane (serum sample) having a particular machine. The blue cells correspond to high probability of having a machine in that column. Smaller probabilities are shown in lighter blue;. *Right*: The estimated machine profiles. Here seven estimated machines are shown, each with component proteins shown by a blue bar "|".