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; E-mail: zhenkewu@umich.edu.

[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.

Abstract

In this paper, we propose a general framework for combining evidence of varying quality to estimate underlying binary latent variables in the presence of restrictions imposed to respect the scientific context. The resulting algorithms cluster the multivariate binary data in a manner partly guided by prior knowledge. The primary model assumptions are that 1) subjects belong to classes defined by unobserved binary states, such as the true presence or absence of pathogens in epidemiology, or of antibodies in medicine, or the "ability" to correctly answer test questions in psychology, 2) a binary design matrix Γ specifies relevant features in each class, and 3) measurements are independent given the latent class but can have different error rates. Conditions ensuring parameter identifiability from the likelihood function are discussed and inform the design of a novel posterior inference algorithm that simultaneously estimates the number of clusters, design matrix Γ, and model parameters. In finite samples and dimensions, we propose prior assumptions so that the posterior distribution of the number of clusters and the patterns of latent states tend to concentrate on smaller values and sparser patterns, respectively. The model readily extends to studies where some subjects' latent classes are known or important prior knowledge about differential measurement accuracy is available from external sources. The methods are illustrated with an analysis of protein data to detect clusters representing auto-antibody classes among scleroderma patients.

Keywords: Clustering; Dependent Binary Data; Markov Chain Monte Carlo; Measurement Error; Mixture of Finite Mixture Models; Latent Class Models.

1 Introduction

Let $\mathbf{Y}$ be a $N \times L$ binary data matrix of N observations with L dimensions or features. Such multivariate binary data frequently arise as noisy measurements of presence or absence of a list of *unobservable* or *latent* binary variables $\boldsymbol{\eta}$ called *states*. Suppose we seek to cluster such data subject to the hypothesis that a cluster is likely to be defined by individuals who share a relatively small number of states. That is, there exist subgroups of $\boldsymbol{\eta}$ vectors that take values on a relatively small number of elements in $\{0,1\}^M$ with $M \leq L$; let the subgroups be denoted by $\mathcal{A}$. We propose a method for estimating *scientifically-structured clusters* (SSC). Our method is most useful for a large dimension L with an unknown number of clusters. Structured clustering for multivariate binary data has a number of potential advantages. If the underlying clusters differ from one another only at subsets of features, SSC can more accurately estimate these clusters than standard clustering methods such as latent class analysis and hierarchical clustering. SSC also results in more interpretable clusters.

Consider three examples from medicine, psychology and epidemiology that motivate scientifically-structured clustering. Example 1 is to estimate subgroups of autoimmune disease patients using autoantibody data that have the potential to predict homogenous disease trajectories (e.g., Joseph et al., 2014). The observed binary responses are imperfect indicators of the presence or absence of specific autoantibody combinations detected in patient sera. Inherent limitation of the lab technique used to identify these autoantibodies (immunoprecipitation, IP) and biological and biochemical variability cause discrepancies between the expected presence/absence of each antibody and the observed values from IP assays. In addition, autoantigens (the specific proteins targeted by autoantibodies) frequently exist as multi-protein complexes, which we will refer to as "machines" in this paper (e.g., Rosen and Casciola-Rosen, 2016, also see Section 4.2). The medical goals are to define the "machines" by their component proteins, and infer whether or not each patient has each machine, using the imprecise IP data.

The second example relates to cognitive diagnosis in psychological and educational assessment. The binary outcomes indicate a subject's responses to many diagnostic questions ("items"). The measurements reflect the person's long-term "true" responses to these items, indicating a student's knowledge for correctly answering a test question absent guessing or

other errors. These "true" or "ideal" responses are further assumed to define a smaller number of binary latent skills that indicate the presence or absence of the particular knowledge (called "states" in the psychology literature). For example, teachers assess whether the student possesses basic arithmetic skills (e.g., addition, multiplication); and psychiatrists diagnose whether patients have certain mental disorders based on a subject's survey responses (e.g., Junker and Sijtsma, 2001). Each question or item is designed to measure a particular subset of latent states, where such item-latent-state correspondence may be known, partially known or unknown.

Example 3 is to estimate the causes of childhood pneumonia from a list of more than 30 different species of pathogens including viruses, bacteria and fungi (e.g., O'Brien et al., 2017). The imperfect binary outcomes indicate whether or not each pathogen was detected by the polymerase chain reaction (PCR) or cell culture from two compartments: the nasopharyngeal (NP) cavity and blood. The binary latent states of scientific interest are the true presence or absence of the pathogens in a child's *lung*, the site of infection that can seldom be directly observed in practice. This example differs from Example 1 in that the correspondence between each of the compartment-technology-pathogen diagnostic measurements ("features") and the latent lung infection ("state") is known because each measurement is designed to detect one specific pathogen and hence is expected to have higher positive rates in classes infected by that pathogen. In addition, the two measurements (NP with PCR and blood with cell culture) are known to have different error rates (e.g., Hammitt et al., 2012; Wu et al., 2016).

In each of these examples, the clustering of observations and subject-specific prediction of $\{\boldsymbol{\eta}_i\}$ comprise the scientific targets for inference. Our examples can be distinguished by:

a) whether the latent state variables ($\boldsymbol{\eta}_i$) are constrained or unconstrained to take values from a pre-specified subset $\mathcal{A}$ where classes are defined by distinct values of $\boldsymbol{\eta}_i$,

b) whether it is known, partially known, or unknown about the binary design matrix Γ that specifies for each latent class the set of *relevant* features having the highest positive response probability than other classes; and

c) the form of the conditional distribution of measurements given latent states and the design matrix (Γ) and response probabilities (Λ): $[\boldsymbol{Y}_i \mid \boldsymbol{\eta}_i, \Gamma, \Lambda]$.

This paper discusses a family of latent class models, referred to as restricted latent class models or RLCMs (e.g., Xu and Shang, 2018) specified by the three components listed above. The model formulation includes as special cases: probabilistic boolean matrix decomposition (Rukat et al., 2017), subset clustering models (Hoff, 2005), and partially latent class models (Wu et al., 2016) among others discussed in detail in Section 2.4. The focus is on estimating clusters based on multivariate binary data that exhibit differential errors depending on the true latent class. The design matrix Γ is assumed to be generated from a low-dimensional latent state vector $\boldsymbol{\eta}_i$. However, in many applications, the number of clusters and/or the set of latent states ($\mathcal{A}$) are not known *in a priori* and must be inferred from data.

We discuss large-sample identifiability conditions for RLCM likelihood-based inference to motivate our posterior algorithm design. However, in finite samples, the likelihood function can be relatively flat before asymptotics concentrate the likelihood around the major mode. To improve finite-sample estimation efficiency at the expense of some bias, we specify sparsity-inducing priors that propagate into the posterior distribution to encourage few clusters with sparse latent state patterns.

We begin this paper with a unified survey of restricted latent class models drawing on the previous work of Wu et al. (2016), Wu et al. (2017b), Hoff (2005), Xu and Shang (2018). The second objective is to present novel Markov chain Monte Carlo (MCMC) algorithms for Bayesian RLCMs with discrete component parameters building on the sampling techniques of Jain and Neal (2004), Miller and Harrison (2017) and Chen et al. (2017). Section 2 presents the model formulation including the likelihood, prior distribution and theoretical identifiability results. In Section 3, we present our MCMC algorithm to efficiently estimate posterior distributions for clusters. Section 4.1 compares via simulation the proposed clustering method to three common alternatives. Section 4.2 illustrates the methods with analysis of the autoantibody data for Example 1. The paper concludes with a discussion of model extensions and limitations.

2 Model

Let $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{iL})^\top \in \{0,1\}^L$ represent a L-dimensional multivariate binary response for subject $i = 1, \ldots, N$; Let $\mathbf{Y}$ collect data from all subjects. We assume each observation

is associated with an unobserved or latent state vector $\boldsymbol{\eta}_i \in \mathcal{A}$, where $\mathcal{A} \subset \{0,1\}^M$ is a set of M-dimension binary vectors.

Given a pre-specified dimension of latent states M, we first specify the likelihood $[\boldsymbol{Y}_i \mid \boldsymbol{\eta}_i, \Gamma, \Lambda]$ via restricted latent class models (RLCM) and then, among others, a prior distribution for $H = \{\boldsymbol{\eta}_i, i = 1, \ldots, N\}$ that groups subjects by their binary patterns $\{\boldsymbol{\eta}_i\}$ (Supplementary Material A.1 extends the prior on H to $M = \infty$). Let $\widetilde{K} = |\mathcal{A}|$ represent the number of groups with *non-zero* population prevalence. Although it is no greater than 2^M, $\widetilde{K}$ can be unknown; And when $\widetilde{K}$ is known and $\widetilde{K} < 2^M$, $\mathcal{A}$ can be unknown. The two steps jointly specify a so-called *mixture of finite mixture model* for $\{\boldsymbol{Y}_i\}$ (Miller and Harrison, 2017). In our setting, the salient feature of scientific import is the discrete component parameters $\{\boldsymbol{\eta}_i\} \subset \mathcal{A}$ that requires additional handling in the posterior algorithm (Section 3). Section 2.4 discusses special cases of the RLCM relevant to the motivating examples. By taking N to infinity in the likelihood, Section 2.5 further studies theoretical limits of identifying unknown model parameters.

2.1 Latent Class Model

For the traditional latent class model (LCM) (e.g., Goodman, 1974), we assume that the latent state vectors $\{\boldsymbol{\eta}_i\}$ take values from a set of binary patterns $\mathcal{A} = \{\widetilde{\boldsymbol{\eta}}_k, k = 1, \ldots, \widetilde{K}\}$, where $\widetilde{K} = |\mathcal{A}|$ is the number of *distinct* patterns. Latent classes differ in their latent state patterns. Given an observation's latent states $\boldsymbol{\eta}_i$, we assume the probability of observing a positive response of feature ℓ for subject i is $\mathbb{P}(Y_{i\ell} = 1 \mid \Lambda) = \lambda_{i\ell}$, $\ell = 1, \ldots, L$, where $\Lambda = \{\lambda_{i\ell}\}$ is a $N \times L$ matrix of response probabilities; For Λ and other matrices in this paper, we will use $\Lambda_{\star\ell}$ and $\Lambda_{i\star}$ to denote the ℓ-th column and i-th row, respectively. A more useful, non-saturated model lets the response probability $\lambda_{i\ell}$ depend on the subject's latent state vector $\boldsymbol{\eta}_i$ via $\lambda_{i\ell} = \lambda_\ell(\boldsymbol{\eta}_i)$ where $\lambda_\ell : \mathcal{A} \to [0,1]$. Because $\boldsymbol{\eta}_i$ can be one of $\widetilde{K}$ elements in $\mathcal{A}$, the classes have at most $\widetilde{K}$ distinct response probabilities, referred to as *between-class differential* measurement errors.

The LCM has 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, i.e. $Y_{i\ell} \perp Y_{i\ell'} \mid \boldsymbol{\eta}_i, \lambda_\ell(\cdot)$. Fitting LCM is to attribute, for example, a positive marginal association observed between two dimensions ℓ and ℓ' to their simi-

lar response probabilities that define the latent classes. Taken together, LCMs specify the conditional probability of observing a multivariate binary outcome $\boldsymbol{y} \in \{0,1\}^L$ by

$$\mathbb{P}(\boldsymbol{Y}_i = \boldsymbol{y} \mid \boldsymbol{\eta}_i, \lambda_\ell(\cdot)) = \prod_{\ell=1}^{L} (\lambda_{i\ell})^{y_{i\ell}} (1-\lambda_{i\ell})^{1-y_{i\ell}}, \text{ where } \lambda_{i\ell} = \lambda_\ell(\boldsymbol{\eta}_i). \tag{1}$$

Because $\boldsymbol{\eta}_i$ is not observed, it is integrated out of (1) with respect to its distribution $\mathbb{P}(\boldsymbol{\eta}_i = \boldsymbol{\eta} \mid \boldsymbol{\pi}_{\widetilde{K}}) = \pi_{\boldsymbol{\eta}} > 0$, for $\boldsymbol{\eta} \in \mathcal{A}$. Based on N independent observations, the LCM likelihood takes the form of "mixture of Bernoulli products": $\prod_{i=1}^{N} \sum_{\boldsymbol{\eta} \in \mathcal{A}} \pi_{\boldsymbol{\eta}} \mathbb{P}\left\{\boldsymbol{Y}_i \mid \boldsymbol{\eta}_i = \boldsymbol{\eta}, \lambda_\ell(\boldsymbol{\eta}), \ell = 1, \ldots, L\right\}$.

Given ℓ, traditional LCMs impose no structure upon the response probability vectors except that they differ among classes almost surely: $\lambda_\ell(\boldsymbol{\eta}) \neq \lambda_\ell(\boldsymbol{\eta}')$ for latent classes $\boldsymbol{\eta} \neq \boldsymbol{\eta}'$. Let $\widetilde{Z}_i \in \{1, \ldots, \widetilde{K}\}$ indicate the *unobserved* class assignment for observation i. An equivalent and more familiar formulation $\lambda_{i\ell} = \lambda_\ell(\widetilde{Z}_i)$ results. The LCM approximates any multivariate discrete distribution for sufficiently large $\widetilde{K}$ (Dunson and Xing, 2009, Corollary 1) and, up to class relabeling, is generically identified whenever $L \geq 2\lceil \widetilde{K} \rceil + 1$ (Allman et al., 2009, Corollary 5). Fitted LCM results will show the estimated response probability profiles that differ by class and can be interpreted as population heterogeneity in particular scientific contexts. Estimation of clusters in finite mixture models often makes use of $\{Z_i\}$, for example, by maximizing the plugged-in conditional posterior $\widehat{Z}_i = \arg\max_{k=1,\ldots,\widetilde{K}} \mathbb{P}(Z_i = k \mid \mathbf{Y}, \widehat{\boldsymbol{\pi}}_{\widetilde{K}})$ or a least-square estimate of clusters based on distance from pairwise co-coclustering posterior probabilities $\widehat{\pi}_{i,i'} = \mathbb{P}(Z_i = Z_{i'} \mid \mathbf{Y})$ (Dahl, 2006).

2.2 Motivation for Scientifically-Structured Classes

The traditional LCM does not incorporate important prior scientific knowledge about how clusters (classes) structurally differ. In Example 1, autoimmune disease patients may differ in their antibody protein presence or absence patterns at $L = 50$ protein landmarks over a grid of molecular weights. The focus is on estimating groups of patients who differ in their immune responses to unknown machines. We formulate this biological prior knowledge by introducing the following model parameters:

i) An M by L machine matrix Q where $Q_{m\ell} = 1$ indicates presence of landmark autoantigen protein ℓ in machine m; We refer to the rows in Q as "machine profiles". In

addition, feature ℓ with $\sum_m Q_{m\ell} = 0$ indicates landmark autoantigen protein ℓ is not targeted as part of any machine.

ii) A patient-specific vector of length M that represents the presence or absence of M machines ($\boldsymbol{\eta}_i = (\eta_{i1}, \ldots, \eta_{iM})^\top$). For example, In Figure 1, for $M = 3$, a subject with $\boldsymbol{\eta}_i = (1, 0, 1)^\top$ has Machines 1 and 3 (middle panel). The two machines produced her antibody proteins (left panel) subject to further errors. Given Q and $\boldsymbol{\eta}_i$, we can represent the presence or absence of antibody proteins deterministically, for example, by $\Gamma_{i\star} = \boldsymbol{\eta}_i^\top Q$ under a row-orthogonal Q as illustrated in Figure 1. For feature ℓ with $\sum_m Q_{m\ell} = 0$, we have $\Gamma_{i\ell} = 0$ for all subjects.

iii) Positive rate parameters, the true- ($\boldsymbol{\theta} = \{\theta_\ell = \mathbb{P}(Y_{i\ell} = 1 \mid \Gamma_{i\ell} = 1)\}$) and false-positive rates ($\boldsymbol{\psi} = \{\psi_\ell = \mathbb{P}(Y_{i\ell} = 1 \mid \Gamma_{i\ell} = 0)\}$). Two sources of stochastic variations contribute to the discrepancy between the expected presence of autoantibody ($\Gamma_{i\ell} = 1$) and the observed presence ($Y_{i\ell} = 1$) or absence ($Y_{i\ell} = 0$): selective immunological non-response to certain autoantigen proteins in a machine and experimental errors. In a priori, we assume high true- and low false- positive rates ($\theta_\ell > \psi_\ell$) because GEA method is robust for detecting immunoprecipitated antibodies.

In summary, i) and ii) incorporate the prior knowledge that antibody proteins are produced in groups against autoantigen proteins coded by the rows of Q and iii) is the measurement likelihood function that assigns probabilities to observed data accounting for stochastic variations. The other two examples in Section 1 can be parameterized in the same way with known or unknown Q (Section 2.4).

Given subjects with $\Gamma_{i\ell} = 1$ or 0, the response probability $\lambda_{i\ell} = \theta_\ell$ or ψ_ℓ regardless of i's class membership (e.g., true presence of antibody protein in serum no matter which machine it comes). Consequently, unlike traditional LCM, a new model where not all features exhibit difference in response probabilities $\{\lambda_{i\ell}, i = 1, \ldots, N\}$ is needed. Using separate class-specific estimates of $\{\lambda_\ell(1), \ldots, \lambda_\ell(\widetilde{K})\}$ for features without actual between-class differential errors can be imprecise and will result in inferior clustering performance (Figure 2, d). RLCMs provide a general framework for specifying class response probability profiles to respect scientific structures through which we show achieve better clustering performance.

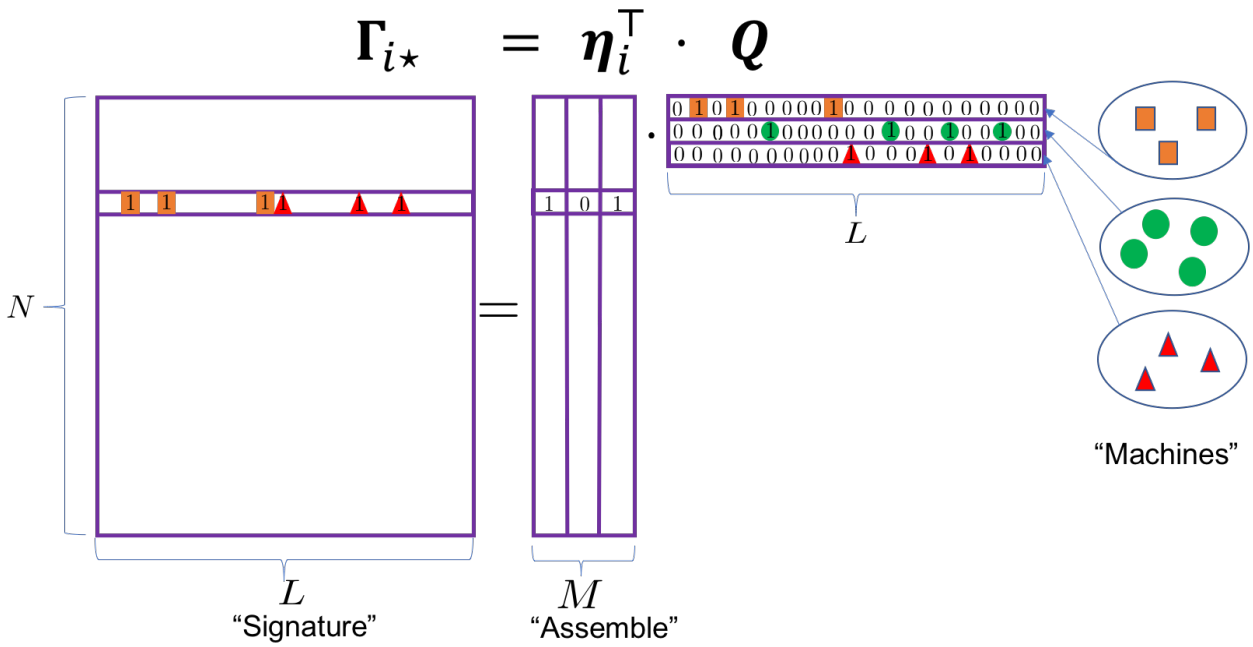

Figure 1: Binary matrix factorization generates composite autoantibody signatures that are further subject to misclassification. The signature $\Gamma_{i\star} = \boldsymbol{\eta}_i^\top \boldsymbol{Q}$ assembles three *orthogonal* machines with 3, 4 and 3 landmark proteins, respectively. The highlighted individual is expected to mount immune responses against antigens in Machines 1 and 3. See texts after model (6).

2.3 Restricted LCMs to Incorporate Scientifically-Structured Classes

RLCMs assume equality among a subset of response probabilities across classes. That is for some ℓ, RLCMs assume $\lambda_{i\ell} = \lambda_{i'\ell}$ for some subjects in distinct latent classes ($\boldsymbol{\eta}_i \neq \boldsymbol{\eta}_{i'}$). The set of RLCM parameters therefore comprises a *Lebesgue measure zero set* in the parameter space for the traditional unconstrained LCM.

The restrictions on response probabilities in RLCMs are specified by introducing a binary design matrix $\Gamma = \{\Gamma_{\boldsymbol{\eta},\ell}\} \in \{0,1\}^{\widetilde{K}\times L}$ with latent classes and dimensions in the rows and columns, respectively. $\Gamma_{\boldsymbol{\eta},\ell} = 1$ represents a positive *ideal* response for which subjects in latent class $\boldsymbol{\eta}$ will have the highest response probability at dimension ℓ; 0 for a negative *ideal* response for which subjects in latent class $\boldsymbol{\eta}$ will have a lower response probability. If $\Gamma_{\boldsymbol{\eta},\ell} = \Gamma_{\boldsymbol{\eta}',\ell} = 1$ for two latent classes $\boldsymbol{\eta}$ and $\boldsymbol{\eta}'$, it is assumed that the ℓ-th dimension is observed with identical positive response probabilities: $\lambda_{i\ell} = \lambda_{i'\ell}$. On the other hand, there can be more than one response probability if $\Gamma_{\boldsymbol{\eta},\ell} = 0$. That is, no equality constraint upon

the response probabilities ($\lambda_{\boldsymbol{\eta},\ell} = \lambda_{\boldsymbol{\eta}',\ell}$) is required for two latent classes with $\Gamma_{\boldsymbol{\eta},\ell} = \Gamma_{\boldsymbol{\eta}',\ell} = 0$.

In this paper, we focus on Q-restricted LCM where the design matrix Γ is determined by the latent state vectors $\boldsymbol{\eta}$ and an M by L binary matrix Q, i.e.,

$$\Gamma_{\boldsymbol{\eta},\ell} = \Gamma(\boldsymbol{\eta}, Q_{\star\ell}), \forall \boldsymbol{\eta} \in \mathcal{A}, \ell = 1, \ldots, L, \tag{2}$$

where the mapping or design matrix $\Gamma(\cdot,\cdot) : \mathcal{A} \times \{0,1\}^M \to \{0,1\}^{\widetilde{K}\times L}$ needs to be specified in the context of the particular scientific study (e.g., $\Gamma_{i\ell} = \boldsymbol{\eta}_i^\top Q_{\star\ell}$ as in Figure 1).

We now introduce scientific structures through the restriction of response probabilities. Let $\mathcal{A}_\ell = \{\boldsymbol{\eta} \in \mathcal{A} : \Gamma_{\boldsymbol{\eta},\ell} = 1\}$ where $\mathcal{A}_\ell$ collects latent classes with the highest response probability for dimension ℓ according to Γ (Gu and Xu, 2018). If $\mathcal{A}_\ell \neq \emptyset$, we restrict the response probabilities at feature ℓ by

$$\max_{\boldsymbol{\eta}\in\mathcal{A}_\ell} \lambda_{\boldsymbol{\eta},\ell} = \min_{\boldsymbol{\eta}\in\mathcal{A}_\ell} \lambda_{\boldsymbol{\eta},\ell} > \lambda_{\boldsymbol{\eta}',\ell}, \ell = 1, \ldots, L, \boldsymbol{\eta}' \notin \mathcal{A}_\ell. \tag{3}$$

Further, there can exist a class $\boldsymbol{\eta} \in \mathcal{A}$ that gives rise to all-zero ideal responses $\Gamma_{\boldsymbol{\eta}\star} = \mathbf{0}_{1\times L}$.

To make the notation clear, in Example 1, $Y_{i\ell}$ represents the observed presence/absence of protein ℓ on the immunoprecipitation gel for patient i, $\boldsymbol{\eta}_i$ indicates this patient's latent class and which protein complexes ("machines") among the rows of Q are present in patient i's class, $Q_{m\star}, m = 1, \ldots, M$, indicates which proteins comprise Machine m, and $\Gamma_{\boldsymbol{\eta},\ell}$ indicates whether or not protein ℓ is present from any machines in latent class $\boldsymbol{\eta}$. The class with no machine has all zeros in its row of Γ. 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. The sensitivity for a given protein is assumed to be the same regardless from which machine(s) it comes; the specificities are allowed to vary across proteins. Finally, the true positive rates are assumed to be larger than the false positive rates.

Finally, we specify the measurement likelihood through parameterization of the response probabilities

$$\lambda_{i\ell} = \lambda_\ell^R(\boldsymbol{\eta}_i; \boldsymbol{\beta}_\ell, Q_{\star\ell}) \in [0,1], \tag{4}$$

where λ_ℓ^R specifies the response probability at feature ℓ with "restriction" (3). The restriction is prescribed by $\Gamma_{\boldsymbol{\eta}_i,\ell}$ which is further determined by $\boldsymbol{\eta}_i$ and $Q_{\star\ell}$ in (2). λ_ℓ^R also depends on

$\boldsymbol{\eta}_i$ and unknown real-value parameters $\boldsymbol{\beta}_\ell$ according to particular parametric models; See model (7) below for an example. In what follows, we use $\Gamma_{i\ell}$ to denote $\Gamma_{\boldsymbol{\eta}_i,\ell}$ unless otherwise noted.

Motivated by our applications, we present an equivalent formulation for λ_ℓ^R that separate true and false positive rates. Let $K_\ell^+ = \#\{\lambda_{i\ell} : \boldsymbol{\eta}_i \in \mathcal{A}_\ell\}$ ($K_\ell^- = \#\{\lambda_{i\ell} : \boldsymbol{\eta}_i \notin \mathcal{A}_\ell\}$) be the number of distinct response probability levels at feature $\ell = 1, \ldots, L$. In RLCMs, we have $K_\ell^+ = 1$ and $K_\ell^- \geq 1$, $\ell = 1, \ldots, L$ (see Table S1 in Supplementary Materials that tabulate the number of distinct response probabilities at dimension ℓ, (K_ℓ^+, K_ℓ^-), for other variants of LCMs). Let θ_l be the maximum response probability at feature ℓ and $\boldsymbol{\psi}_\ell = \{\psi_{l1}, \ldots, \psi_{l,K_\ell^-}\}$ be the rest of response probabilities, respectively. Given $\boldsymbol{\eta}_i = \boldsymbol{\eta} \notin \mathcal{A}_\ell$, let $v_i = v(\boldsymbol{\eta}_i, \ell)$, where $v(\cdot,\cdot)$: $(\boldsymbol{\eta}_i, \ell) \mapsto v$ is the integer-valued function that selects among $\boldsymbol{\psi}_\ell$ her associated response probability ψ_{ℓ,v_i} at feature ℓ. The parameters θ_ℓ and $\boldsymbol{\psi}_\ell$ may be further parameterized by $(\boldsymbol{\beta}_\ell, Q_{\star\ell})$ as in (4). For models with $K_\ell^- = 1$, $v(\cdot) = 1$; Otherwise, $\nu(\cdot,\cdot)$ depends on $\mathcal{A}$ (the set of possible pattern of $\boldsymbol{\eta}_i$), the specific functional form of $\lambda_\ell^R(\cdot)$ and parameter values of $(\boldsymbol{\beta}_\ell, Q_{\star\ell})$ in a RLCM (see the example (7) in Section 2.4; The traditional LCM results by setting $Q = 1_{M\times L}$ and under $K^+ + K^- = \widetilde{K}$ for each ℓ).

In this paper, because we focus on models with the structure in (2), we can equivalently represent the response probability parameters λ_ℓ^R in (4) by

$$\lambda_\ell^R(\boldsymbol{\eta}_i; \boldsymbol{\beta}_\ell, Q_{\star\ell}) = \{\theta_\ell\}^{\Gamma_{\boldsymbol{\eta}_i,\ell}} \cdot \left\{\boldsymbol{\psi}_{\ell,v(\boldsymbol{\eta}_i,\ell)}\right\}^{1-\Gamma_{\boldsymbol{\eta}_i,\ell}} \in [0,1], \tag{5}$$

where $\boldsymbol{\beta}_\ell = \{\boldsymbol{\theta} = \{\theta_\ell\}, \Psi = \{\boldsymbol{\psi}_\ell\}\}$ with constraints $\theta_\ell > \psi_{\ell,v}, \forall v = 1, \ldots, K_\ell^-$. RLCMs therefore let observations with latent state patterns in $\mathcal{A}_\ell$ take identical and the highest probability than other classes in $\mathcal{A}_\ell^c$. Other classes in $\mathcal{A}_\ell^c$ respond with lower probabilities at dimension ℓ. We now discuss some examples.

2.4 Examples of RLCMs in the Literature

Special cases of restricted LCMs result when $K_\ell^+ = K_\ell^- = 1$. For example, a class of models assumes the response probabilities

$$\lambda_{i\ell} = \theta_\ell^{\Gamma_{i\ell}} (\psi_\ell)^{1-\Gamma_{i\ell}}, \qquad \Gamma_{i\ell} = 1 - \prod_{m=1}^{M} (1-\eta_{im})^{Q_{m\ell}}. \tag{6}$$

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 partially latent class models in disease epidemiology (Wu et al., 2016, PLCM) or Deterministic In and Noisy Or (DINO) in cognitive diagnostic models (e.g., Templin and Henson, 2006, DINO) that needs just one required state ($\{m : Q_{m\ell} = 1\}$) for a positive ideal response $\Gamma_{i\ell} = 1$. Imposing constant and symmetric error rates $\theta_\ell = \psi_\ell$, $\ell = 1, \ldots, L$, the one-layer model of Rukat et al. (2017) results. The model can also be viewed as Boolean matrix factorization (BMF, Miettinen et al., 2008) by noting that $\Gamma_{i\ell} = \vee_{m=1}^{M} \eta_{im} Q_{m\ell}$ 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. BMF further reduces to nonnegative matrix factorization (e.g., Lee and Seung, 1999) $\Gamma = HQ$ where $H = \{\eta_{im}\}$ if Q has orthogonal rows. See Supplementary Materials A.2 for a connection to subset clustering in Hoff (2005). A second two-parameter example results by assuming $\Gamma_{i\ell} = \prod_{m=1}^{M} (\eta_{im})^{Q_{m\ell}}$ (e.g., Junker and Sijtsma, 2001). This model, referred to as Deterministic In and Noise And (DINA) gate model in the cognitive diagnostic literature, assumes a conjunctive (noncompensatory) relationship among latent states $m = 1, \ldots, M$. That is, it is necessary to possess *all* the attributes (states) indicated by non-zero elements in $Q_{\star\ell}$ to be capable of providing a positive ideal response $\Gamma_{i\ell} = 1$. The model also imposes the assumption that possessing additional unnecessary attributes does not compensate for the lack of the necessary ones. These two-parameter models are equivalent upon defining $\eta^*_{im} = 1 - \eta_{im}$, $\Gamma^*_{i\ell} = 1 - \Gamma_{i\ell}$, $\psi^*_\ell = 1 - \psi_\ell$ and $\psi^*_\ell = 1 - \theta_\ell$ (Chen et al., 2015). There are several other examples in this category as discussed by Xu (2017).

Two-parameter models assume that "$\Gamma_{\boldsymbol{\eta},\ell} = \Gamma_{\boldsymbol{\eta}',\ell} = 0$ implies identical response probabilities $\lambda_{\boldsymbol{\eta},\ell} = \lambda_{\boldsymbol{\eta}',\ell} = \psi_\ell$", regardless of the distinct patterns $\boldsymbol{\eta} \neq \boldsymbol{\eta}'$. In practice, deviation from such assumptions occurs if $\boldsymbol{\eta}$ has more nonzero elements than $\boldsymbol{\eta}'$ and alters the response probabilities, i.e., $K^-_\ell > 1$. Multi-parameter models where $K^-_\ell > K^+_\ell = 1$, popular in multidimensional item response theory, is readily specified for example by assuming an all-effect model: $\lambda_{i\ell} = \lambda^R_\ell(\boldsymbol{\eta}_i; \boldsymbol{\beta}_\ell, Q_{\star\ell}) = \text{expit}\left\{\boldsymbol{\beta}^\top_\ell \boldsymbol{h}(\boldsymbol{\eta}_i, Q_{\star\ell})\right\} =$

$$\text{expit}\left\{\beta_{\ell 0} + \sum_{m=1}^{M} \beta_{\ell m}(Q_{m\ell}\eta_{im}) + \sum_{m<m'} \beta_{\ell mm'}(Q_{m\ell}\eta_{im})(Q_{m'\ell}\eta_{im'}) + \ldots + \beta_{\ell 12\ldots M} \prod_{m}(Q_{m\ell}\eta_{im})\right\} \quad (7)$$

that includes higher order interactions among latent states required by an item (Henson et al., 2009); Here $\text{expit}(x) = \frac{\exp(x)}{1+\exp(x)}$. When $\prod_{m=m_1,\ldots,m_s} Q_{m\ell} = 0$, this saturated model

needs no $\beta_{\ell,m_1\ldots m_s}$ term. Setting second or higher order terms to zero, an additive main-effect model results. The effects of latent states need not be additive. For example, $\mathsf{log}(\lambda_{i\ell}) = \beta_{\ell 0} + \sum_{m=1}^{M} \beta_{\ell m} Q_{m\ell} \eta_{im}$ specifies a multiplicative model that penalizes the absence of an required latent state m if $Q_{m\ell} = 1$.

Table S1 in Supplementary Materials summarizes these and other variants of LCMs by specifications of the latent state space, design matrix, and measurement processs.

2.5 Identifiability

There are two sources of indeterminancy in restricted LCMs: invariance of the likelihood function to permutation of the ordering of the latent states and over-parameterized models. The permutation invariance manifests itself as a multimodal posterior distribution. Where Q is unknown, we address the permutation invariance by labeling the latent states, one dimension at a time, by the non-zero patterns of the corresponding rows in an estimated Q. We address the over-parameterization by introducing prior distributions that encourage *in a priori* few clusters hence a small number of parameters via mixture of finite mixture models (Miller and Harrison, 2017). It helps to show identifiability results or lack thereof to motivate such sparsity-inducing priors.

Given $\widetilde{K}$ and M, identifiability conditions characterize the theoretical limits of recovering the unknown model parameters (Q, Λ, $\boldsymbol{\pi}_{\widetilde{K}}$) from the likelihood for all or a subset of the parameter space. We first discuss the identifiability of Q because it is needed for interpreting latent states (see Section 4.2) and for estimating both H and $\boldsymbol{\pi}_{\widetilde{K}}$. Based on the likelihood $[\boldsymbol{Y}_i \mid \boldsymbol{\pi}_{\widetilde{K}}, \Lambda, \Gamma = \Gamma(Q)]$ with a given Q and a saturated $\mathcal{A}$ (or "full diversity": $\pi_{\boldsymbol{\eta}} > 0, \forall \boldsymbol{\eta} \in \mathcal{A} = \{0,1\}^M$), Xu (2017) studied sufficient conditions for *strict* identifiability of Λ and $\boldsymbol{\pi}_{\widetilde{K}}$ over the entire parameter space in RLCMs. Under weaker conditions upon the design matrix Γ (instead of Q) and possibly non-saturated $\mathcal{A}$, Gu and Xu (2018) established conditions that guarantee *partial* identifiability for general RLCMs which means the likelihood function is flat over a subset of the parameter space. When Q-matrix is completely unknown, it is possible to identify $\{\boldsymbol{\pi}_{\widetilde{K}}, \Lambda, Q\}$ just using likelihood $[\boldsymbol{Y}_i \mid \boldsymbol{\pi}_{\widetilde{K}}, \Lambda, \Gamma = \Gamma(Q)]$. In particular, Chen et al. (2015) provided sufficient conditions for the special cases of DINA and DINO models (see Section 2.4); Xu and Shang (2018) further generalized them to general RLCM: (Q, Λ, $\boldsymbol{\pi}_{\widetilde{K}}$) are strictly identifiable (up to row reordering of Q) in RLCMs with saturated $\mathcal{A}$

if the following two conditions hold:

C1) The true Q can be written as a block matrix $Q = [I_M; I_M; \widetilde{Q}]$ after necessary column and row reordering, where $\widetilde{Q}$ is a $M \times (L - 2M)$ binary matrix and

C2) $(\Lambda_{\boldsymbol{\eta},\ell}, \ell > 2M)^\top \neq (\Lambda_{\boldsymbol{\eta}',\ell}, \ell > 2M)^\top$ for any $\boldsymbol{\eta} \neq \boldsymbol{\eta}'$ and $\boldsymbol{\eta} \succeq \boldsymbol{\eta}'$,

where $\boldsymbol{a} \succeq \boldsymbol{b}$ for $\boldsymbol{a} = \{a_j\}$ and $\boldsymbol{b} = \{b_j\}$ if and only if $a_j \geq b_j$ holds element-wise.

Because condition (C2) depends on Q, Λ and row and column permutations, the number of operations to check (C2) increases exponentially with M, $\mathcal{O}((L - 2M)2^M M)$, for a saturated $\mathcal{A}$ with 2^M patterns of latent state vectors. We instead use condition (C3) that just depends on Q and that is invariant to row or column permutations:

C3) Each latent state is associated to at least three items, $\sum_{\ell=1}^{L} Q_{m\ell} \geq 3$ for all m.

Condition (C3) enables convenient restrictions in MCMC sampling and takes just $\mathcal{O}(LM)$ operations to check. For special cases of RLCM, the DINA and DINO models (Section 2.4) with a saturated $\mathcal{A}$, Conditions (C1) and (C3) suffice to identify (Q, Λ, $\boldsymbol{\pi}_{\widetilde{K}}$) (Theorem 2.3, Chen et al., 2015).

Posterior algorithms typically restrict MCMC sampling of non-identified parameters by identifiability conditions to prevent aggregation of posterior probability mass from multiple modes. For example, in factor analysis of multivariate continuous data, one can restrict the loading matrices in lower triangular forms (e.g., Geweke and Zhou, 1996). Alternatively, one may first perform MCMC sampling with weak and simple-to-check constraints without fully ensuring identifiability and just check afterwards whether the parameters are conditionally identifiable. One then performs necessary deterministic transformations on parameters that may only be identified up to equivalent classes to pick coherent and economical representatives, for example, by relabeling sampled mixture components at each iteration or varimax rotations of factor loading matrices in classical Gaussian factor analysis (e.g., Ročková and George, 2016).

We initialize the sampling chain from the set defined by simple identifiability conditions (C1) and (C3) and only check afterwards at each iteration whether the parameters are conditionally identifiable according to conditions (C1) and (C2) that are stronger and

computationally more expensive. The relabeling of the latent states is done by inspecting the non-zero patterns in the rows of Q (Step 7, Supplementary Material C.1).

In applications where Q is unknown with $M < L/2$, we focus on the set of Q-matrices that satisfy both (C1) and (C3):

$$\mathcal{Q} = \{Q \in \{0,1\}^{M \times L} : Q = P_1 Q^{\dagger} P_2,\ Q^{\dagger} = [I_M; I_M; \widetilde{Q}],\ \widetilde{Q}\mathbf{1}_{L-2M} \succeq \mathbf{1}_{L-2M}\}, \tag{8}$$

where P_1 and P_2 are M- and L-dimensional permutation matrices for rows and columns, respectively. The constraint $\mathcal{Q}$ also greatly facilitates posterior sampling by focusing on a small subset of binary matrices. In fact, among all M by L binary matrices, the fraction of $Q \in \mathcal{Q}$ is at most $\frac{\binom{L}{2M}\left[2^{(L-2M)M}\right]}{2^{L \cdot M}}$ and quickly decay as the number of machines M increases. In some applications it may also simplify posterior inference by exploiting further assumptions upon Q for example partially known Q or non-overlapping (i.e., orthogonal) rows of Q. See Supplementary Materials A.3 and A.4 for other identifiability considerations that motivate our posterior algorithms.

2.6 Priors

Given M, we specify the prior for $H = \{\boldsymbol{\eta}_i\}$ with cluster structure among N subjects in five steps: 1) Generate the vector of probabilities of a subject i belonging to each of K clusters $\boldsymbol{\pi}_K = (\pi_1, \ldots, \pi_K)^{\top}$ where K is possibly unknown and sampled from its prior $p_K(\cdot)$; 2) Partition observations by indicators $Z_i \overset{i.i.d}{\sim} \mathsf{Categorical}(\boldsymbol{\pi}_K)$; Suppose we obtain T distinct $\{Z_i\}$ values; 3) Draw the vector of marginal probabilities of each latent state being active $\boldsymbol{p} = \{p_m\}$; 4) Draw from $[\boldsymbol{\eta}_j^* \mid \boldsymbol{p}, M]$, for clusters labeled $j = 1, \ldots, T$, where "*" indicates cluster-specific quantities; 5) Combine $\{\boldsymbol{\eta}_j^*\}$ and $\{Z_i\}$ to obtain subject-specific latent states $\boldsymbol{\eta}_i = \boldsymbol{\eta}_{Z_i}^*$, $i = 1, \ldots, N$.

2.6.1 Prior for Partitioning Observations

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 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, $\mathcal{C} = \{C_j : |C_j| > 0\}$ 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, the number of components for the *population*. Further let $C \in \mathcal{C}$ denote one of the clusters in partition $\mathcal{C}$; let j be the index associated with cluster C_j, for $j \in \{1, \ldots, T\}$. 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 $\boldsymbol{\eta}_i$ be the latent state vector for subject $i = 1, \ldots, N$, and $\boldsymbol{\eta}_j^*$ be the latent state vectors for cluster $j = 1, \ldots, T$.

We assume the indicators $\boldsymbol{Z}$ are drawn as follows:

$$\text{Number of components}: \quad K \sim p_K, \tag{9}$$

$$\text{Mixing weights}: \quad \boldsymbol{\pi}_K \sim \text{Dirichlet}(\gamma, \ldots, \gamma), \tag{10}$$

$$\text{Cluster indicators}: \quad Z_i \sim \text{Categorical}\{\boldsymbol{\pi}_K = (\pi_1, \ldots, \pi_K)\}, i = 1, \ldots, N, \tag{11}$$

where p_K is a probability mass function over non-zero integers $\{1, 2, \ldots\}$ and $\gamma > 0$ is the hyperparameter for symmetric K-dimensional Dirichlet distribution. Note that though $\widetilde{K} \leq 2^M$, K is not upper bounded (unless constrained through the support of p_K). The prior of partition $\mathcal{C}$ induced by (9-11) is $p(\mathcal{C} \mid \gamma, p_K(\cdot)) = V_N(T) \prod_{C \in \mathcal{C}} \gamma^{(|C|)}$, 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).

2.6.2 Prior for H^*

Given $\{Z_i\}$, we draw the latent state vector $\boldsymbol{\eta}_j^* \in \{0, 1\}^M$ for which $Z_i = j$ indicates, referred to as "component-specific parameters" in mixture models. We discuss priors for these *discrete* component parameters according as $\mathcal{A}$ is known or not.

<u>*Pre-specified* $\mathcal{A}$.</u> In applications such as Example 3, pre-specifying $\mathcal{A}$ is appealing when the scientific interest lies in itemized characterization of the population fractions for each element of $\mathcal{A}$. Given $\mathcal{A}$, the cluster membership indicators $\{Z_i\}$ take value from $\{1, \ldots, T\}$

where $T = \widetilde{K}$. Existing approaches then assign to each cluster one of $\{\boldsymbol{\eta}_1^*, \ldots, \boldsymbol{\eta}_T^*\}$ by enumerating the distinct *known* elements in $\mathcal{A}$. For example, see Chen et al. (2015) for $\mathcal{A} = \{0,1\}^M$, $\widetilde{K} = 2^M$. Wu et al. (2016) analyzed data from Example 3 and specified $\mathcal{A} = \{\boldsymbol{e}_1, \ldots, \boldsymbol{e}_M, \mathbf{0}_M\}$ among pneumonia *cases* that represents latent states as the lung infection caused by pathogen $1, 2, \ldots, M$ or none-of-the-above and $\boldsymbol{\eta}_i = \mathbf{0}_M$ among observed *controls*. Absent the uncertainty in $\mathcal{A}$, simpler posterior sampling algorithms result.

In practice, to avoid misleading estimates based on a pre-specified $\mathcal{A}$ subject to potential misspecification, analysts may conservatively specify $\mathcal{A} = \{0,1\}^M$. However, $\boldsymbol{\eta}_i = \boldsymbol{\eta}_{Z_i}^*$ then take its value from a space that grows exponentially with M (e.g., $M = 30$ in Example 3). Consequently, upon fitting the model for inferring $\pi_k, k = 1, \ldots, \widetilde{K}(= 2^M)$, although many elements in $\mathcal{A}$ may receive low posterior probabilities, none is exactly zero. Important elements in $\mathcal{A}$ are commonly selected by *ad hoc* thresholding. In addition, pre-specifying $\mathcal{A} \subsetneqq \{0,1\}^M$ does not address the question of what are the distinct latent state patterns $\widetilde{\boldsymbol{\eta}}_j^*$ in the data.

<u>*Unknown* $\mathcal{A}$</u>. Absent knowledge of $\mathcal{A}$, we draw *in a priori* the component-specific parameters $H^* = \{\eta_{jm}^*\}$ in two steps for regularizing $\boldsymbol{\eta}_j^*$ towards sparsity:

$$\text{probability of an active state}: \quad p_m \mid \alpha_1, \alpha_2 \sim \text{Beta}(\alpha_1\alpha_2/M, \alpha_2), \tag{12}$$

$$\text{latent states}: \quad \eta_{jm}^* \mid p_m \sim \text{Bernoulli}(p_m), j = 1, \ldots, T, \tag{13}$$

for $m = 1, \ldots, M$. Note that it is possible that $\boldsymbol{\eta}_j^* = \boldsymbol{\eta}_{j'}^*$ for some $j, j' = 1, \ldots, T$ where equality holds element-wise. For example, $\boldsymbol{\eta}_{Z_i}^*$ may equal $\boldsymbol{\eta}_{Z_i'}^*$ even if $Z_i \neq Z_{i'}$. Because we are interested in estimating distinct $\boldsymbol{\eta}_j^*$'s that represent distinct values of scientific latent constructs, we will merge such clusters j and j' into one, referred to as a "scientific cluster"; We denote it by $\widetilde{\mathcal{C}}$. We also denote the unique values in $H^* = \{\boldsymbol{\eta}_j^*, j = 1, \ldots, T\}$ by $\widetilde{H}^* = \{\widetilde{\boldsymbol{\eta}}_j^*, j = 1, \ldots, \widetilde{T}\}$. Supplementary Material A.5 and A.6 further remarks on the induced priors on the partitions $\mathcal{C}$ and $\widetilde{\mathcal{C}}$.

Remark 1. *The K introduced in the prior specification is to make it not upper bounded and therefore differs from $\widetilde{K}$. The latter represents the number of* ***distinct*** *latent state vectors in the population and must be no greater than 2^M. $\widetilde{\boldsymbol{\eta}}_k, k = 1, \ldots, \widetilde{K}$ represent the set of true distinct latent state vectors in the population; while $\boldsymbol{\eta}_j^*, j = 1, \ldots, T$ ($T \leq K$) represent the realized latent state vectors that are* ***possibly duplicated*** *in the data generating*

process (13) or the posterior sampling. With unconstrained K, we are able to build on the algorithm of Miller and Harrison (2017) that does not bound the number of mixture components. The resulting algorithm works for general mixture of finite mixture models with ***discrete*** *component distributions (Section 3).*

By Beta-Bernoulli conjugacy, we integrate $[H^* \mid \boldsymbol{p}][\boldsymbol{p} \mid \alpha_1, \alpha_2]$ over $\boldsymbol{p}$ to obtain the marginal prior:

$$pr(H^*) = \prod_{m=1}^{M} \frac{(\alpha_1\alpha_2/M)\Gamma(s_m + \alpha_1\alpha_2/M)\Gamma(T - s_m + \alpha_2)}{\Gamma(T + \alpha_2 + \alpha_1/M)}, \tag{14}$$

where $\Gamma(\bullet)$ is the Gamma function and $s_m = \sum_{m=1}^{T} \eta^*_{jm}$, $j = 1, \ldots, M$. Holding α_2 constant, the average number of positives among $\boldsymbol{\eta}^*_j$ decreases with α_1; Holding α_1 constant, the latent state vectors, $\boldsymbol{\eta}^*_j$ and $\boldsymbol{\eta}^*_{j'}$, $j \neq j'$, become increasingly similar as α_2 decreases. In fact, the probability of two subjects with distinct cluster indicators Z_i and $Z_{i'}$ have identical m-th latent state, $\mathbb{P}[\eta^*_{im} = \eta^*_{i'm} \mid Z_i = j, Z_{i'} = j', j \neq j', \alpha_1, \alpha_2] = \mathbb{E}\{p_m^2 + (1 - p_m)^2 \mid \alpha_1, \alpha_2\} = 1 - 2\frac{\alpha_1}{\alpha_1+M}\left(1 - \frac{\alpha_1\alpha_2+M}{\alpha_1\alpha_2+\alpha_2 M+M}\right)$ approaches one when α_2 goes to zero. 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 $p_m = 0.5$ to specify uniform distribution over all possible patterns over $\mathcal{A} = \{0, 1\}^M$.

2.6.3 Priors for Other Model Parameters

We focus on the situation where Q is completely unknown. Let Q be uniformly distributed over the constrained space in $\{0, 1\}^{M\times L}$ defined by (8). In applications where Q is not fully identifiable and/or encouraged to be different among its rows in finite samples, we specify sparsity priors for each column of Q to encourage proteins to be specific to a small number of machines (see Supplementary Material A.6).

We specify the priors for response probabilities $\Lambda = \{\lambda_{i\ell}\}$ in (5) to satisfy the monotonic constraints in (3) as follows

$$\psi_{\ell,v} \sim \mathsf{Beta}(N_\psi a_\psi, N_\psi(1 - a_\psi)), v = 1, \ldots, K^-_\ell, \text{ constrained to } \Delta = \left\{\{\boldsymbol{\psi}_\ell\} : \psi_{\ell,1} < \ldots < \psi_{\ell,K^-_\ell}\right\},$$

$$\theta_1, \ldots, \theta_L \sim \mathsf{Beta}(N_\theta a_\theta, N_\theta(1 - a_\theta))\, \mathbb{I}\{(\max_{1\leq v\leq K^-_\ell} \psi_{\ell,v}, 1)\}, a_\psi \sim \mathsf{Beta}(a_0, b_0), \text{ and } a_\theta \sim \mathsf{Beta}(a'_0, b'_0),$$

for $\ell = 1, \ldots, L$, where $K_\ell^- \geq 1$ is the number of response probability parameters for latent classes $\boldsymbol{\eta}$ with $\Gamma_{\boldsymbol{\eta},\ell} = 0$ defined in (2) and the truncation of θ_ℓ follows from the definition of RLCM (3). With $(a_\theta,\, a_\psi)$ unknown, the hierarchical priors on $\boldsymbol{\theta}$ and $\{\boldsymbol{\psi}_v\}$ propagate into the posterior and have the effect of shrinking the parameters towards a population value by sharing information across dimensions; $(N_\theta,\, N_\psi)$ can further be sampled in the posterior algorithm or fixed. When multi-parameter RLCMs specify particular parametric forms of the response probability for feature ℓ (e.g., in (7)), other sets of priors on the parameters may be readily incorporated into posterior sampling by modifying Step 4 in Supplementary Material C.1. Finally, we specify prior for hyperparameter α_1 in (12). One may specify a prior conjugate to $[H^* \mid \alpha_1]$ by $\alpha_1 \overset{d}{\sim} \mathsf{Gamma}(e_0, f_0)$ (shape and inverse scale parameterization with mean e_0/f_0 and variance e_0/f_0^2). Posterior sampling for non-conjugate prior for α_1 can also be carried out by sampling over a dense grid upon bounded reparameterization (see Step 5 in Supplementary Material C.1).

Taken together, the likelihood and priors give the joint distribution of data $\mathbf{Y} = \{\boldsymbol{Y}_i\}$, the true and false positive rates $\boldsymbol{\theta}$ and $\boldsymbol{\Psi}$, Q matrix, and latent state vectors $H = \{\boldsymbol{\eta}_i\}$ (see Supplementary Material A.8).

3 Posterior Inference

We design posterior sampling algorithms to address three questions, 1) how many scientific clusters ($\widetilde{T}$) in the sample (data); 2) what are the latent state vectors $\{\widetilde{\boldsymbol{\eta}}_j^*, j = 1, \ldots, \widetilde{T}\}$ in the sample; and 3) what are the subjects' latent states $\boldsymbol{\eta}_i$ and the scientific clusters $\widetilde{\mathcal{C}}$.

Given Q, $\boldsymbol{\theta}$ and Ψ, RLCM as a mixture model has *discrete* component-specific parameters $\boldsymbol{\eta}_i \in \mathcal{A}$. This is to be contrasted with mixture models with a continuous base measure from which component parameters are drawn to differ from one another with probability one. Therefore, when sampled conditional on other parameters, the discrete component parameters $\{\boldsymbol{\eta}_j^*, j = 1, \ldots, T\}$ may be duplicated. Because we are interested in estimating scientific clusters with distinct latent states, we *post-process* the posterior samples by merging clusters in $\mathcal{C}$ associated with identical $\boldsymbol{\eta}_j^*$ at each MCMC iteration. Given M, no more than 2^M distinct latent state vectors $\widetilde{\boldsymbol{\eta}}_j^*$ results after merging. More generally, for inference based on mixture of finite mixture (MFM) models with *discrete* component parameters, (9) uses

a prior over all non-negative integers to remove the otherwise hard constraint $K = \widetilde{K} \leq 2^M$ (would be so if we force distinct latent states in the prior) and greatly simplify the design of posterior algorithms (see Remark 1).

We use Markov chain Monte Carlo (MCMC) algorithm for posterior inference which by design simulate samples that approximate the joint posterior distribution of unknown parameters and latent variables: $(\boldsymbol{Z}, H^*, Q, \boldsymbol{\theta}, \Psi, \alpha_1)$. See Supplementary Material C.1 for more details of the sampling algorithms and convergence checks. We discuss information from data that updates the clusters $\mathcal{C}$.

Gibbs updates of the partitions. 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. A key step is to sample $\mathcal{C}$ based on an urn process that begins with one cluster comprised of all subjects (or a warm start informed by crude initial clusters) and re-assigns each subject to an old or new cluster (Miller and Harrison, 2017). In sampling $\{Z_i\}$ one subject at a time, the full conditional distribution $[Z_i \mid \boldsymbol{Z}_{-i}, \mathbf{Y}, \boldsymbol{\theta}, \boldsymbol{\Psi}, Q, \boldsymbol{p}]$ given cluster assignments for the rest $\boldsymbol{Z}_{-i} = \{Z_{i'}, i' \neq i\}$, other model parameters and data is proportional to the product of the conditional prior $pr(Z_i \mid \boldsymbol{Z}_{-i}, \gamma)$ and the complete data likelihood integrated over latent states $[\mathbf{Y} \mid \boldsymbol{Z}, \boldsymbol{\theta}, \Psi, Q, \boldsymbol{p}]$ (equivalent to conditional upon partition $\mathcal{C}$ ignoring the labels). Because of exchangeability among subjects, we view subject i as the last observation to be updated during a Gibbs step which assigns subject i to an existing cluster $C \in \mathcal{C}_{-i}$ or a new cluster on its own with probabilities:

$$\mathbb{P}(Z_i = j \mid -) \propto \begin{cases} (|C| + \gamma) \cdot \frac{g(C \cup \{i\})}{g(C)}, & \text{if } C \in \mathcal{C}_{-i}, j = 1, \ldots, |\mathcal{C}_{-i}|, \text{ or} \\ \gamma \frac{V_N(t+1)}{V_N(t)} \cdot g(C), & \text{if } C = \{i\}, j = |\mathcal{C}_{-i}| + 1, \end{cases} \tag{15}$$

where $g(C) = g(C; \boldsymbol{\theta}, \Psi, Q, \boldsymbol{p}) = \prod_{\ell=1}^{L} pr(\{Y_{i\ell} : i \in C\} \mid \boldsymbol{\theta}, \Psi, Q, \boldsymbol{p})$ is the marginal likelihood for data in cluster C (see (S4) in Supplementary Material B for an illustration using model (6)). If adding subject i to any existing cluster fits poorly with data $\mathbf{Y}_C$, i.e., knowing $\mathbf{Y}_C$ tells little about $\boldsymbol{Y}_i$, low marginal likelihood ratio $\frac{g(C \cup \{i\})}{g(C) g(\{i\})}$ will result for any $C \in \mathcal{C}_{-i}$. The Gibbs update will favor forming a cluster of its own $\{i\}$.

Posterior summaries. We summarize the posterior distribution of partitions $[\mathcal{C} \mid \mathbf{Y}]$ by computing the empirical frequencies $\widehat{\pi}_{ii'}$ for every pair of subjects being clustered together, re-

ferred to as the posterior co-clustering probabilities $\pi_{ii'} = \mathbb{P}(Z_i = Z_{i'} \mid \mathbf{Y})$, for subjects i, $i' = 1, \ldots, N$. We compute a simple least square (LS) clustering $\widehat{\mathcal{C}}^{(LS)}$ on the basis of the squared distance from the posterior co-clustering probabilities, $\arg\min_b \sum_{i,i'} \left\{ \delta(Z_i^{(b)}, Z_{i'}^{(b)}) - \widehat{\pi}_{ii'} \right\}^2$, where $\delta(a, a') = 1$ if $a = a'$ and zero otherwise (Dahl, 2006).

RLCM has the salient feature of subject-specific discrete latent states $\boldsymbol{\eta}_i$. However, the interpretation of $\boldsymbol{\eta}_i$ depends on Q which is of scientific interest on its own in many applications. Based on the posterior samples obtained from a model with an unknown Q, we select the iteration(s) b^* with the mininum loss, $\min_{b^*} \|Q^{(b^*)\top} Q^{(b^*)} - \frac{1}{B} \sum_{b=1}^{B} Q^{(b)\top} Q^{(b)}\|_F$ where $\|A\|_F = \sqrt{\sum a_{ij}^2}$ is the matrix Frobenius norm. $Q^\top Q$ is a L by L matrix invariant to relabeling of latent states. The (ℓ, ℓ')-th element of $Q^\top Q$ represents the number of activated states at feature ℓ when $\ell = \ell'$ and the number of co-activated states at feature pair (ℓ, ℓ') when $\ell \neq \ell'$. Minimization of the least squares criterion therefore selects an iteration closest to the posterior means of all the co-activation counts. Turning to the inference of $\boldsymbol{\eta}_i$, although in the original MCMC chain the subset of the $H^{*(b)}$ and $\boldsymbol{Z}^{(b)}$ samples drawn along with $Q^{(b^*)}$ usefully approximate $[\boldsymbol{\eta}_i = \boldsymbol{\eta}^*_{Z_i}, i = 1, \ldots, N \mid Q = Q^{(b*)}, \mathbf{Y}]$, inferences of their functions enjoy reduced Monte Carlo errors through refitting a model with $Q = Q^{(b*)}$ that generate more posterior samples. Section 4.2 further illustrates these use of the posterior summaries through detailed analyses of data from Example 1.

4 Results

We illustrate the utility of RLCM on both simulated and real data. We focus on scenarios where Q is unknown. First, we assess the performance of RLCM on estimating clusters under simulation scenarios corresponding to distinct levels of measurement errors, feature dimensions, sparsity levels of each machine, sample sizes, and population fractions of latent state patterns. Here the goal is to show that the proposed Bayesian RLCM performs clustering as well as or better than common alternative binary-data clustering methods. We first analyze a single randomly generated data set to highlight the differences among the methods. We then investigate the frequentist property of Bayesian RLCM in cluster estimation and compare it to other methods through repeated application of each method to replication data sets. Finally, data from Example 1 is analyzed, focusing on the posterior inferences of

clusters, cluster-specific latent states and the estimated Q-matrix.

4.1 Simulated Examples to Study Model Performance

Simulation 1: More accurate clustering through feature selection in scientifically structured classes. $N = 50$ independent observations are generated from an $L = 100$ dimension multivariate binary distribution with $M = 3$ machines. Here we randomly generated an M by L matrix Q where each row has on average $s = 20\%$ non-zero elements. That is, $Q_{m\ell} \overset{i.i.d}{\sim} \mathsf{Bernouli}(0.2), \ell = 1, \ldots, L$; In the rare event where a randomly generated $Q \notin \mathcal{Q}$ (identifiability constraint (8)), 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 $\boldsymbol{\eta}_i \overset{d}{\sim} \mathsf{Categorical}\left(\boldsymbol{\pi}_0 = (1/6, 1/6, 1/6, 1/6, 1/12, 1/12, 1/12, 1/12)\right)$ where

$$\boldsymbol{\pi}_0 = \{\mathbb{P}(\boldsymbol{\eta}_i = (0,0,0), (1,0,0), (0,1,0), (1,1,0), (0,0,1), (1,0,1), (0,1,1), (1,1,1))\}.$$

Here we focus on the two-parameter model ((6), DINO) which will be applied to Example 1 in Section 4.2. 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. 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 $M^\dagger = 5$ can be seen in Figure 2. Specifically, for Bayesian LCA, RLCM and subset clustering (Hoff, 2005), we plot the posterior co-clustering probability matrix $\{\widehat{\pi}_{i,i'}\}$ for N observations; For HC, we indicate co-clustering by filled cells. The true clusters are separated (dashed grids) and ordered according to the truth. Filled blocks on the main diagonal indicate perfect recovery of the true clusters. In this setting, HC is sensitive to noise and tends to split a true cluster (blank cells within the main diagonal blocks) or group observations from different true clusters (blue cells in the off-diagonal blocks). Unlike the Bayesian LCA and the subset clustering, the Bayesian RLCM automatically selects and filter subsets of features that distinguish eight classes (through scientific structures in (6)) hence has superior clustering performance producing clusters that agrees quite well with the truth. This advantage of

Bayesian RLCM relative to alternatives is maintained under data replications (see Simulation 2).

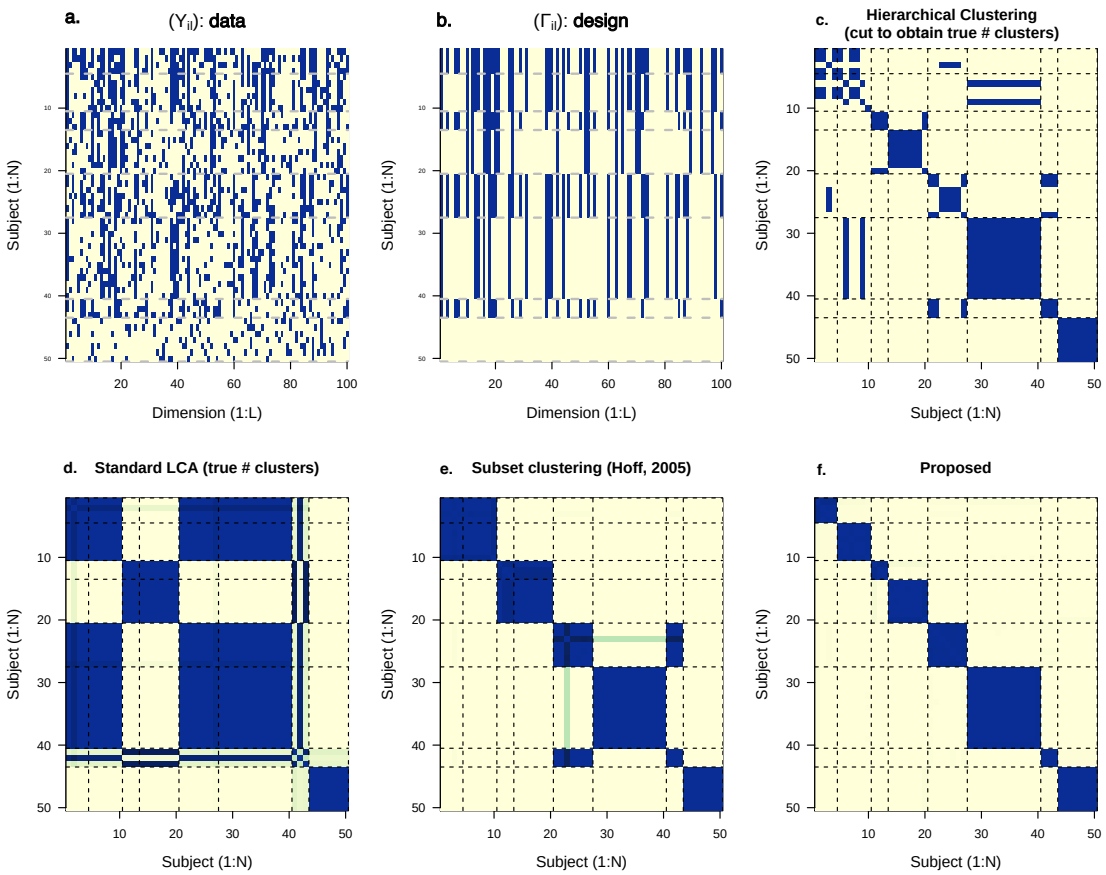

Figure 2: In the 100-dimension multivariate binary data example, the eight classes differ with respect to subsets of measured features. Bayesian restricted latent class analysis accounts for measurement errors, selects the relevant feature subsets and filters the subsets by a low-dimensional model (6) and therefore yields superior clustering results.

Compared to traditional all-feature methods under large dimensions, through the inference of all-zero columns of Q ($\{\ell : \Gamma_{\boldsymbol{\eta},\ell} = 0, \forall \boldsymbol{\eta} \in \mathcal{A}\}$), 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 E for additional simulated examples on this point).

Simulation 2: Assess clustering performance under various parameter settings. We simulated $R = 60$ replication data sets for each of $1,920$ combinations of (#features, sample size, true positive rate, false positive rate, population fractions, 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 Examples 1-3. 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 $\mathsf{aRI}(\mathcal{C}, \mathcal{C}') = \frac{\sum_{r,c}\binom{n_{rc}}{2} - [\sum_r \binom{n_{r\cdot}}{2} \sum_c \binom{n_{\cdot c}}{2}]/\binom{N}{2}}{0.5[\sum_r \binom{n_{r\cdot}}{2} + \sum_c \binom{n_{\cdot c}}{2}] - [\sum_r \binom{n_{r\cdot}}{2} \sum_c \binom{n_{\cdot c}}{2}]/\binom{N}{2}}$, 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 $\mathcal{C}'$, $\sum_{r,c}\binom{n_{rc}}{2} (\leq 0.5\left[\sum_r \binom{n_{r\cdot}}{2} + \sum_c \binom{n_{\cdot c}}{2}\right])$ 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 between the two clustering methods.

First we apply Bayesian RLCM to each replication data set and focus on studying its performance in recovering the true clusters (boxes with solid lines in Figure S3). The clustering performance varies by the sparsity level (s) in each machine, level of measurement errors (θ_ℓ, ψ_ℓ), population fractions of latent classes $\{\pi_{\boldsymbol{\eta}}, \boldsymbol{\eta} \in \mathcal{A}\}$ and sample sizes (N). 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 S3 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 $s = 20\%$. More generally, 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 S3). In the context of Example 1, 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).

We obtain more accurate cluster estimates under larger discrepancies between θ_ℓ and ψ_ℓ. For θ_0 fixed at 0.8 or 0.9, the mean aRI averaged over replications is higher under $\psi_0 = 0.05$ than $\psi_0 = 0.15$ over all combinations of the rest of parameters. Under the non-uniform population fraction $\boldsymbol{\pi}_0 = \boldsymbol{\pi}_b$, the clustering performance by Bayesian RLCM is similar or slightly worse than under a uniformly distributed population ($\boldsymbol{\pi}_a$). Finally, we

observe mixed relative performances at distinct sample sizes as a result of two competing factors: more precise estimation of measurement error parameters under large sample sizes that improve clustering and a larger space of clusterings under a larger N.

Figure S3 also shows better clustering performance of Bayesian RLCM (boxes with solid lines) relative to the three common alternatives (boxes with dotted lines). The Bayesian RLCM on average most accurately recovers the clusters compared to other methods. Bayesian RLCM produces the highest aRIs compared to others which are in many settings perfect (close to one). For example, 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 (the leftmost group of four boxplots in Column 1, Figure S2 of Supplementary Materials $\psi_0 = 0.05$, $s = 10\%$, $\boldsymbol{\pi}_0 = \boldsymbol{\pi}_a$); The relative advantage of Bayesian RLCM and HC narrows under a higher false positive rate ($\psi_0 = 0.15$) as shown by the smaller aRI ratios 1.23, 1.62, 1.49, 1.16 (the leftmost group of four boxplots in Column Two, Figure S2).

We remark on the performance of other three methods. Over all parameter settings investigated here, the traditional LCA performed the worst in the recovery of true clusters (aRI < 0.68). The likelihood function of subset clustering is a special case of 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 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.

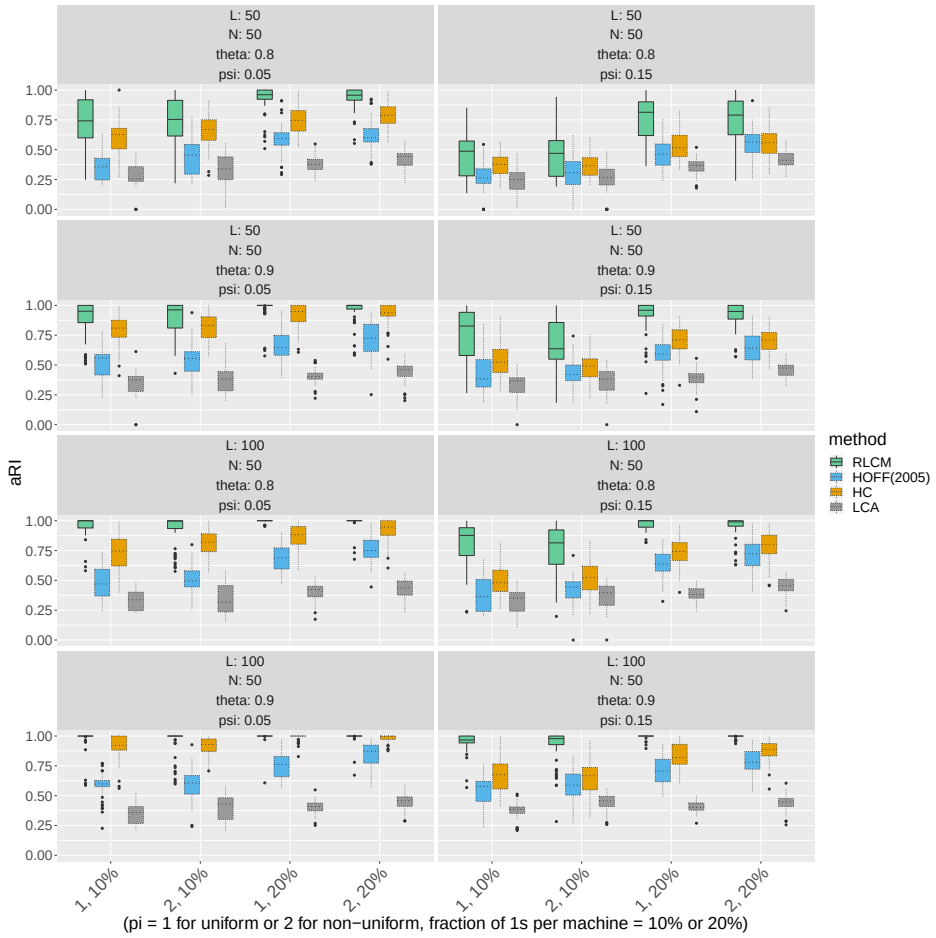

Figure 3: Based on $R = 60$ replications for each parameter setting, Bayesian RLCM (boxplots with solid lines) most accurately recovers the true clusters compared to subset clustering (Hoff, 2005) hierarchical clustering (HC) and traditional Bayesian latent class analysis (LCA) (from the left to the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

4.2 Analysis of GEA Data

4.2.1 GEA Data, Preprocessing and Informative Priors

Example 1 is about estimating 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., 2017a) implemented using publicly

available software R package "spotgear" (https://github.com/zhenkewu/spotgear). We ran 4 GEA gels, each loaded with 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 (6) 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 measurement errors, we choose not to share measurement error rates across dimension in this analysis. In our analysis, we sampled many Q across iterations of 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). The prior of H^* (Section 2.6.2) prevents overfitting by encouraging a small number of *active* latent states ($\{m : \sum_i \eta_{im} \neq 0\}$) for small α_1 which in this analysis we draw its posterior samples 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. They proposed that these proteins 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 $\widetilde{T}^{(b)} \leq |\mathcal{C}^{(0)}| + N - \sum_{j=1}^{|\mathcal{C}^{(0)}|} C_j^{(0)}$, 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.

4.2.2 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 marginal 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).

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 marginal posterior probabilities of having each of the estimated machines. Among 76 subjects analyzed, 23 of them have greater than 95% marginal posterior

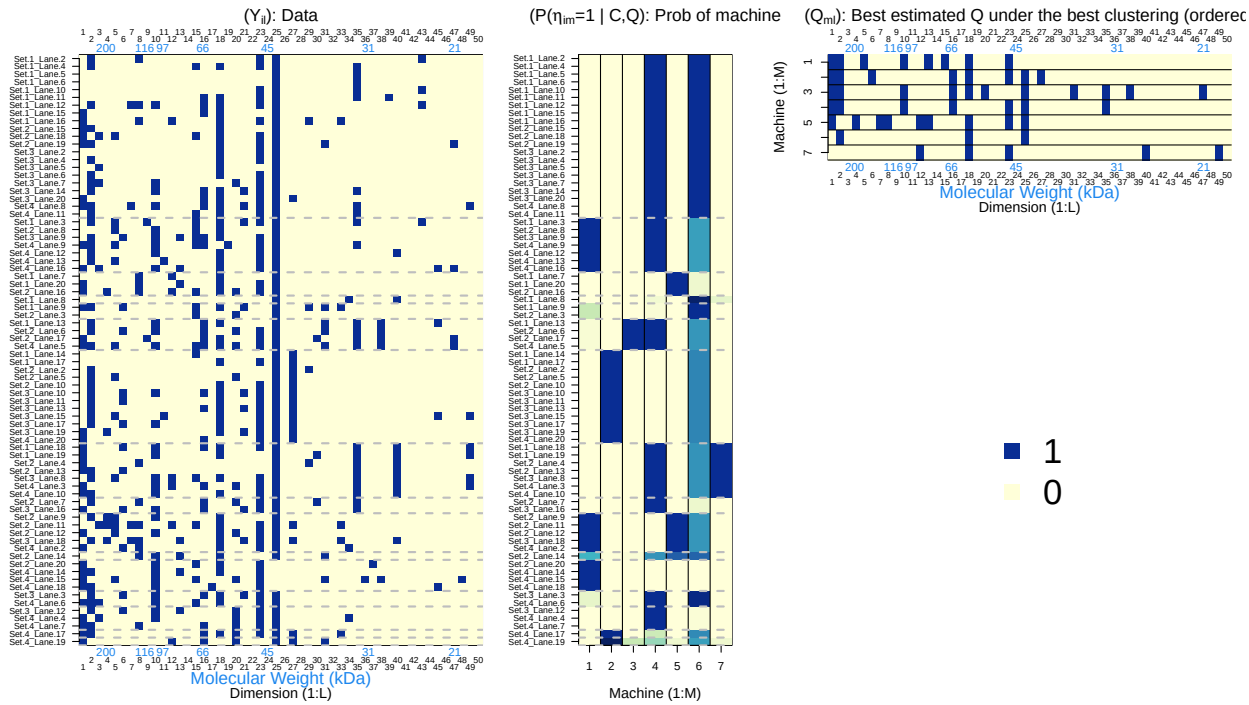

Figure 4: Results for GEA data in Example 1. *Left*) Aligned data matrix for band presence or absence; row for 76 serum lanes, reordered into optimal estimated clusters (not merged) $\widehat{\mathcal{C}}^{(LS)}$ separated by gray horizontal lines "—; columns for $L = 50$ protein landmarks. A blue vertical 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 "|".

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 also fitted a Bayesian RLCM without the partial clusters $\mathcal{C}^{(0)}$ identified in prior work by the scientists. We estimated lower true positive rates so that it is more likely to observe negative protein landmarks within clusters partially identified by having a machine with a protein at that landmark. This makes the findings more difficult to interpret. As discussed in the simulation studies, clustering performance of Bayesian RLCM is poorer under lower sparsity levels $s = 10\%$. As our scientific team recruits and tests more serum samples from

their scleroderma patient cohort, samples with novel antibodies will improve inference about the measurement error parameters. This highlights the importance of using available prior knowledge about the measurement technologies in inferring latent states in finite samples (e.g., Wu et al., 2016). Figure S5 in Supplementary Materials compares for each landmark the prior and posterior distributions of the true and false positive rates. The discrepancies observed at many landmarks suggest the learning of measurement error parameters from the data. Other landmarks have similar prior and posterior distributions as a result of nearly flat likelihood function or absence of protein at that landmark so learning based only on likelihood is impossible.

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.

There are potential improvements in our analysis. The posterior predictive probabilities (PPP) of observing a more extreme log odds ratio in future data $\mathbb{P}(\mathsf{LOR}_{1,2}(\mathbf{Y}^{\mathsf{rep}}) < \mathsf{LOR}_{1,2}(\mathbf{Y}) \mid \mathbf{Y})$ are between 0.004 and 0.024. Most of these misfits of marginal log odds ratio occurred for landmark pairs with an observed marginal two-way table with small cell counts. Because the Bayesian RLCM treats the zeros as random, if these zero cells correspond to impossible combinations of proteins, or structural zeros, it may overestimate the probability for these cells; See Manrique-Vallier and Reiter (2014) for a truncated extension of traditional latent class models that can be adapted to address the structural zero issue. On the other hand, the neighboring Landmarks 1 and 2 have an observed log odds ratio of -1.17 (s.e. 0.48) with PPP 0.011. The two landmarks compete for being aligned with an observed band during pre-processing (Wu et al., 2017a) hence creating negative dependence even within a latent class. Deviation from local independence can be further accounted

for by explicitly modeling local dependence structure, discussed elsewhere, e.g., by nesting subclasses within each class (e.g., Wu et al., 2017b).

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 discussed Bayesian restricted latent class model for analyzing multivariate binary data in the presence of between-class differential errors. The focus has been on the clustering of observations with unknown number of clusters, uncertainty assessment of the clustering and the prediction of individual latent states. The proposed method is motivated by clustering autoimmune disease patients based on their antibody presence or absence in sera where it is scientifically meaningful to *restrict* the values of response probabilities among latent classes. 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. The Bayesian RLCM performs what we have called scientifically-structured clustering. It automatically selects subset of features for each latent class and filters them through a low dimensional model to improve our ability to accurately estimate clusters. Though the present paper focused on demonstrating the method through an example in medicine, the developed method and algorithms apply to many problems including Example 2 and 3 (Section 1).

RLCMs decompose the variation among multivariate binary responses into structure that reflects prior scientific knowledge and stochastic variation without a known explanation. In Example 1, 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.

This paper has focused on developing and applying RLCMs and algorithms to identify clusters and estimate subject-specific latent states. However, applied to public health research (e.g., pneumonia etiology research in Example 3), RLCM analyses more often focus on population quantities such as $\Pi = \{\pi_{\boldsymbol{\eta}}, \sum_j \pi_{\boldsymbol{\eta}} = 1, \boldsymbol{\pi}_{\boldsymbol{\eta}} \geq 0, \boldsymbol{\eta} \in \{0,1\}^M\}$ an M-way contingency table characterizing the population frequencies of the latent state vector $\boldsymbol{\eta}_i$. Further research into flexible and parsimonious parameterization of Π and its regression formulation in RLCMs are warranted. For example, quadratic exponential family (Zhao and Prentice, 1990) with negative second-order natural parameters assigns higher probabilities for $\boldsymbol{\eta}$ comprised of few ones or use another level of latent Gaussian variables to induce flexible dependence among $\boldsymbol{\eta}_i$ (e.g., Xu and Craig, 2009).

We are currently studying a few potentially useful model extensions. First, nested partially LCMs (Wu et al., 2017b) incorporate local dependence and multiple sensitivity parameters ($K^+ > 1$) that would improve the utility of Bayesian RLCMs as well. Second, because the algorithm involves iterating over subjects to find clusters in (15), 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).

Software Availability

All model estimations are performed by an `R` package "`rewind`", which is freely available at https://github.com/zhenkewu/rewind.

Supplementary Materials

The supplementary materials contain referenced figures, a table, remarks, 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 R01 AR073208, P30 AR070254 and P30 CA-046592 (ZW, Cancer Center Support Grant (CCSG) Development Funds from University of Michigan Comprehensive Cancer Center (UMCCC)). We also thank Gongjun Xu, Peter Hoff and Jian Kang for their insightful comments.

References

Albert, P. S., McShane, L. M., and Shih, J. H. (2001). Latent class modeling approaches for assessing diagnostic error without a gold standard: with applications to p53 immunohistochemical assays in bladder tumors. *Biometrics*, 57(2):610–619.

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.

Anderson, T. W. (1954). On estimation of parameters in latent structure analysis. *Psychometrika*, 19(1):1–10.

Brooks, S. and Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, 7(4):434–455.

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

Chen, Y., Liu, J., Xu, G., and Ying, Z. (2015). Statistical analysis of q-matrix based diagnostic classification models. *Journal of the American Statistical Association*, 110(510):850–866.

Chiu, C.-Y., Douglas, J. A., and Li, X. (2009). Cluster analysis for cognitive diagnosis: Theory and applications. *Psychometrika*, 74(4):633.

Cuthill, E. and McKee, J. (1969). Reducing the bandwidth of sparse symmetric matrices. In *Proceedings of the 1969 24th national conference*, pages 157–172. ACM.

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

De La Torre, J. (2011). The generalized dina model framework. *Psychometrika*, 76(2):179–199.

Dunson, D. and Xing, C. (2009). Nonparametric bayes modeling of multivariate categorical data. *Journal of the American Statistical Association*, 104(487):1042–1051.

Erosheva, E. A., Fienberg, S. E., and Joutard, C. (2007). Describing disability through individual-level mixture models for multivariate binary data. *The annals of applied statistics*, 1(2):346.

Fox, E. B., Hughes, M. C., Sudderth, E. B., Jordan, M. I., et al. (2014). Joint modeling of multiple time series via the beta process with application to motion capture segmentation. *The Annals of Applied Statistics*, 8(3):1281–1313.

Garrett, E. and Zeger, S. (2000). Latent class model diagnosis. *Biometrics*, 56(4):1055–1067.

Gelman, A., Meng, X.-L., and Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica*, 6(4):733–760.

Geweke, J. and Zhou, G. (1996). Measuring the pricing error of the arbitrage pricing theory. *The review of financial studies*, 9(2):557–587.

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.

Gilks, W. R. and Wild, P. (1992). Adaptive rejection sampling for gibbs sampling. *Applied Statistics*, pages 337–348.

Goodman, L. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika*, 61(2):215–231.

Green, P. J. (1995). Reversible jump markov chain monte carlo computation and bayesian model determination. *Biometrika*, 82(4):711–732.

Gu, Y. and Xu, G. (2018). Partial Identifiability of Restricted Latent Class Models. *ArXiv e-prints*.

Gustafson, P. (2009). What are the limits of posterior distributions arising from nonidentified models, and why should we care? *Journal of the American Statistical Association*, 104(488):1682–1695.

Hammitt, L., Murdoch, D., Scott, J., Driscoll, A., Karron, R., Levine, O., O′Brien, K., et al. (2012). Specimen collection for the diagnosis of pediatric pneumonia. *Clinical Infectious Diseases*, 54(suppl 2):S132–S139.

Hartigan, J. A. (1990). Partition models. *Communications in statistics-Theory and methods*, 19(8):2745–2756.

Henson, R. A., Templin, J. L., and Willse, J. T. (2009). Defining a family of cognitive diagnosis models using log-linear models with latent variables. *Psychometrika*, 74(2):191.

Hoff, P. D. (2005). Subset clustering of binary sequences, with an application to genomic abnormality data. *Biometrics*, 61(4):1027–1036.

Hubert, L. and Arabie, P. (1985). Comparing partitions. *Journal of classification*, 2(1):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(1):158–182.

Joseph, C. G., Darrah, E., Shah, A. A., Skora, A. D., Casciola-Rosen, L. A., Wigley, F. M., Boin, F., Fava, A., Thoburn, C., Kinde, I., et al. (2014). Association of the autoimmune disease scleroderma with an immunologic response to cancer. *Science*, 343(6167):152–157.

Junker, B. W. and Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement*, 25(3):258–272.

Kadane, J. (1974). The role of identification in Bayesian theory. *Studies in Bayesian Econometrics and Statistics*, pages 175–191.

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(6755):788–791.

Leighton, J. P., Gierl, M. J., and Hunka, S. M. (2004). The attribute hierarchy method for cognitive assessment: A variation on tatsuoka's rule-space approach. *Journal of educational measurement*, 41(3):205–237.

Manrique-Vallier, D. and Reiter, J. P. (2014). Bayesian estimation of discrete multivariate latent structure models with structural zeros. *Journal of Computational and Graphical Statistics*, 23(4):1061–1079.

McCullagh, P., Yang, J., et al. (2008). How many clusters? *Bayesian Analysis*, 3(1):101–120.

Miettinen, P., Mielikäinen, T., Gionis, A., Das, G., and Mannila, H. (2008). The discrete basis problem. *IEEE Transactions on Knowledge and Data Engineering*, 20(10):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*.

Nobile, A. and Fearnside, A. T. (2007). Bayesian finite mixtures with an unknown number of components: The allocation sampler. *Statistics and Computing*, 17(2):147–162.

O'Brien, K. L., Baggett, H. C., Brooks, W. A., Feikin, D. R., Hammitt, L. L., Howie, S. R., Deloria Knoll, M., Kotloff, K. L., Levine, O. S., Madhi, S. A., et al. (2017). Introduction to the epidemiologic considerations, analytic methods, and foundational results from the pneumonia etiology research for child health study. *Clinical infectious diseases*, 64(suppl_3):S179–S184.

Pepe, M. S. and Janes, H. (2006). Insights into latent class analysis of diagnostic test performance. *Biostatistics*, 8(2):474–484.

Pitman, J. (1995). Exchangeable and partially exchangeable random partitions. *Probability theory and related fields*, 102(2):145–158.

Ročková, V. and George, E. I. (2016). Fast bayesian factor analysis via automatic rotations to sparsity. *Journal of the American Statistical Association*, 111(516):1608–1622.

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(12):1540–1542.

Teh, Y. W., Grür, D., and Ghahramani, Z. (2007). Stick-breaking construction for the indian buffet process. In *Artificial Intelligence and Statistics*, pages 556–563.

Templin, J. L. and Henson, R. A. (2006). Measurement of psychological disorders using cognitive diagnosis models. *Psychological methods*, 11(3):287.

Wu, Z., Casciola-Rosen, L., Shah, A. A., Rosen, A., and Zeger, S. L. (2017a). Estimating autoantibody signatures to detect autoimmune disease patient subsets. *Biostatistics*, page kxx061.

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(1):97–114.

Wu, Z., Deloria-Knoll, M., and Zeger, S. L. (2017b). Nested partially latent class models for dependent binary data; estimating disease etiology. *Biostatistics*, 18(2):200.

Xu, G. (2017). Identifiability of restricted latent class models with binary responses. *The Annals of Statistics*, 45(2):675–707.

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

Xu, H. and Craig, B. A. (2009). A probit latent class model with general correlation structures for evaluating accuracy of diagnostic tests. *Biometrics*, 65(4):1145–1155.

Zhao, L. P. and Prentice, R. L. (1990). Correlated binary regression using a quadratic exponential model. *Biometrika*, 77(3):642–648.

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

The supplementary materials contain referenced remarks, figures and a table in Main Paper, and further technical details, e.g., on identifiability and sampling algorithms, as well as additional simulations and extended data analysis results. In particular, Section A contains remarks, Section B illustrates the calculation of marginal likelihood central to the posterior sampling of clusters ((15) in Main Paper), Section C details the posterior algorithms for pre-specified M (Section C.1) and infinite M (Section C.2), respectively. Section D briefly summarizes useful theoretical identifiability conditions for RLCMs based on Gu and Xu (2018). Section E illustrates through simulations the benefit of removing irrelevant features. Finally, Section F collects a table for variants of LCMs as well as figures for model results on the data analysis in Main Paper.

A Remarks

A.1 On Extending Prior of H to $M = \infty$

In Main Paper, we have focused on models with a finite number of latent states with $M = M^\dagger$ typically set to a number that is large enough for the particular applications. In the MCMC sampling (Supplementary Material C.1), not all of the "working" $M^\dagger$ states will be used by the observations. The active number of states is usually strictly smaller than $M^\dagger$ based on simulations. We extend to infinite M to obtain a prior for H^* under infinite dimension of latent state vectors ($\boldsymbol{\eta}_i$). We take M in (14) in Main Paper to infinity and obtain infinite-column prior for H (through a prior on H^* in Section 2.6.2 in Main Paper); This construction defines the infinite Indian Buffet process (Ghahramani and Griffiths, 2006). Supplementary Material C.2 provides posterior sampling algorithms for dealing with an infinite number of latent states by a novel slice sampler without the need of truncation (Teh et al., 2007).

A.2 RLCM Connection to Hoff (2005)

Setting $Q = I_{L\times L}$ and $\boldsymbol{\eta}_i \in \mathcal{A} = \{0,1\}^L$ (i.e., $M = L$) gives "mixture of Bernouli products" with each latent class (defined by $\boldsymbol{\eta}_i$) having *relevant* features at possibly overlapping subsets of features $\mathcal{L}_{\boldsymbol{\eta}} = \{\ell : \Gamma_{\boldsymbol{\eta},\ell} = 1\}$, $\boldsymbol{\eta} \in \mathcal{A}$ (Hoff, 2005). Hoff (2005) assumes the positive response probability $\lambda_{i\ell} = \{\theta_{\ell,v}\}^{\Gamma_{i\ell}} (\psi_\ell)^{1-\Gamma_{i\ell}}$, where $\Gamma_{i\ell} = \eta_{i\ell}$ given $Q = I_{L\times L}$ and the multiple true positive rates $\{\theta_{\ell,v}\}$ are greater than a single false positive rate ψ_ℓ, for $\ell = 1,\ldots,L$. This model can be written into a RLCM form with $K^+ = 1$ and $K^- \geq 1$ by reparametrization: $\Gamma^*_{i\ell} = 1 - \Gamma_{i\ell}$, $\psi^*_{\ell,v} = 1 - \theta_{\ell,v}$ and $\theta^*_\ell = 1 - \psi_\ell$ and relabeling of the outcomes $Y^*_{i\ell} = 1 - Y_{i\ell}$. Indeed, the positive response probability under relabeling and reparameterization is $\lambda^*_{i\ell} = \mathbb{P}(Y^*_{i\ell} = 1 \mid -) = 1 - \mathbb{P}(Y_{i\ell} = 1 \mid -) = 1 - \lambda_{i\ell} = \left\{\psi^*_{\ell,v}\right\}^{1-\Gamma^*_{i\ell}} (\theta^*_\ell)^{\Gamma^*_{i\ell}}$.

A.3 Additional Identifiability Considerations for Designing Posterior Algorithms

We now turn to inferring subject-specific latent state vectors $H = \{\boldsymbol{\eta}_i\}$ based on *complete-data* likelihood $[\{\boldsymbol{Y}_i\} \mid H, \Lambda, Q]$. Even given Q, conditions for identifying H exist but may fall short of ensuring consistent estimation of H because the number of unknowns in H diverges as the sample size increases. For example, it requires extra conditions that the number of measurements L increases with the sample size (e.g., Chiu et al., 2009). In finite samples and dimensions, we address this issue in a Bayesian framework by *in a priori* encouraging H to be of low complexity, i.e., few clusters of distinct and sparse latent state vectors $\{\boldsymbol{\eta}_i\}$, which combined with data likelihood will by design tend to concentrate the posterior at such low-complexity H.

In addition, when the latent space $\mathcal{A} \subsetneqq \{0,1\}^M$, general identifiability theory for Q depends on the identifiability of Γ, the structure of which then determines the set of Qs that are identifiable from the observed data distribution. Some RLCMs motivate our posterior algorithm design. For example, in two-parameter RLCMs, if two latent states are either always present or absent at the same time ("partners"), it is impossible for the likelihood alone to distinguish it from a model that combines the two latent states. In our posterior algorithm, we therefore merge such "partner" latent states if present at some iterations and the corresponding rows in Q (Step 3, Supplementary Material C.1). As another example,

two latent states can form a hierarchical structure, that is, one latent state cannot be present unless the other is. Suppose the second latent state require the first latent state, then Q_{2*} values at $\{\ell : Q_{1\ell} = 1\}$ can be zero or one without altering the model likelihood. The sparsity priors on H and the rows of Q constraining $\sum_\ell Q_{m\ell}$ therefore concentrate the posterior distributions of H and Q towards low-dimensional latent states and a smaller number of rows in Q (Section 2.6.2 in Main Paper).

A.4 Prior information about Λ.

In applications where prior information about a subset of response probabilities Λ is available, it is essential to integrate the informative priors into model estimation if strict or generic identifiabilities do not hold (e.g., Gustafson, 2009; Wu et al., 2016). The sufficient conditions (C1) and (C2) in Main Paper ensure identifiability of Q with completely unknown $(\Lambda, \boldsymbol{\pi}_{\widetilde{K}})$. Otherwise, absent likelihood-based identifiability of Q and other parameters, prior information about Λ alleviates the non-identifiability issue by concentrating the posterior at parameter values that better explain the observed data in light of the informative priors. In general non-identified models, the uncertainty in the prior will propagate into the posterior and will not vanish even as the sample size approaches infinity (e.g., Kadane, 1974).

A.5 Prior for Partition $\mathcal{C}$

The prior distribution $p(\mathcal{C} \mid \gamma, p_K(\cdot))$ is an exchangeable partition probability function (EPPF, Pitman, 1995), because it only symmetrically depends on the sizes of each block of the partition $\{|\mathcal{C}_j| : \mathcal{C}_j \in \mathcal{C}\}$. Miller and Harrison (2017, Theorem 4.1) also derives an *urn process* for generating partitions $\mathcal{C}_1, \mathcal{C}_2, \ldots$, such that the probability mass function for $\mathcal{C}_N$ is given by $p(\mathcal{C} \mid \gamma, p_K(\cdot)) = V_N(T) \prod_{C \in \mathcal{C}} \gamma^{(|C|)}$,; we will use this urn process for Gibbs updates of $\{Z_i\}$ one subject at a time in (17) in Main Paper. Note that the mapping from $\boldsymbol{Z}$ to $\mathcal{C}$ is many-to-one with each $\mathcal{C}$ corresponding to $\binom{K}{T} T!$ distinct $\boldsymbol{Z}$ that differ by relabeling. Starting from a prior for partition $\mathcal{C}$ then followed by drawing component-specific parameters from their prior distributions is particularly fruitful in product partition models (e.g., Hartigan, 1990).

A.6 On Merging Clusters with Identical Discrete Latent States

At each MCMC iteration, two observations falling in distinct clusters ($Z_i \neq Z_{i'}$) might have identical latent states, i.e., $\boldsymbol{\eta}^*_{Z_i} = \boldsymbol{\eta}^*_{Z'_i}$ where the equality holds elementwise. At each iteration, we use unique multivariate binary vectors among all subjects $H = \{\boldsymbol{\eta}_i = \boldsymbol{\eta}^*_{Z_i}, i = 1, \ldots, N\}$ to define "scientific clusters" $\widetilde{\mathcal{C}}$ through merging clusters associated with identical latent states. That is,

$$\widetilde{\mathcal{C}} = \left\{ \{i : \boldsymbol{\eta}_i = \widetilde{\boldsymbol{\eta}}^*_j\}, j = 1, \ldots, \widetilde{T} \right\}$$

where $\{\widetilde{\boldsymbol{\eta}}^*_j, j = 1, \ldots, \widetilde{T}\}$ collects $\widetilde{T}(\leq T)$ unique patterns among $\{\boldsymbol{\eta}^*_j, j = 1, \ldots, T\}$. Let $\mathcal{M} : \{\boldsymbol{\eta}^*_{Z_i}, i = 1, \ldots, N\} \mapsto \widetilde{\mathcal{C}}$ represent this merge operation, i.e., $\widetilde{\mathcal{C}} = \mathcal{M}(\{\boldsymbol{\eta}^*_j\}, \{Z_i\})$.

As detailed in Section 3 in Main Paper, we first build on Gibbs updates (15) and split-merge updates (e.g., Jain and Neal, 2004) to efficiently sample $\{Z_i\}$ from its posterior distribution. Given $\{Z_i\}$, we then update $H^* = \{\boldsymbol{\eta}^*_j\}$ and merge clusters $\mathcal{C}$ to obtain $\widetilde{\mathcal{C}}$ via the mapping $\mathcal{M}$. Define partial ordering " $\preceq$ " over partitions $\mathcal{C}_1 \preceq \mathcal{C}_2$ if for any $C_1 \in \mathcal{C}_1$, one can find a $C_2 \in \mathcal{C}_2$ satisfying $C_1 \subseteq C_2$. We have $\mathcal{C} \preceq \widetilde{\mathcal{C}}$, i.e., $\widetilde{\mathcal{C}}$ is coarser than $\mathcal{C}$. Our procedure for obtaining clusters $\widetilde{\mathcal{C}}$ differs from mixture models where distinct Z_i values with probability one correspond to distinct component parameters sampled from a continuous base measure (e.g., Miller and Harrison, 2017, Proof of Theorem 4.2). $\widetilde{\mathcal{C}} = \mathcal{C}$ is implicitly assumed in Hoff (2005) under a Dirichlet process mixture model.

We specify priors on K that represents the distinct values that $\{Z_i\}$ can take and a prior on $H^* = \{\boldsymbol{\eta}^*_j, j = 1, \ldots, T\}$, which together induce a prior for $\widetilde{\mathcal{C}}$ via

$$p(\widetilde{\mathcal{C}} \mid \alpha_1, \gamma) = \sum_{\mathcal{C}:\mathcal{C}\preceq\widetilde{\mathcal{C}}} p(\widetilde{\mathcal{C}} \mid \mathcal{C}, \alpha) \cdot p(\mathcal{C} \mid \gamma) \tag{S2}$$

$$= \sum_{\mathcal{C}:\mathcal{C}\preceq\widetilde{\mathcal{C}}} \binom{2^M}{\widetilde{T}} (\widetilde{T})! \left\{ \int p(H^* \mid \mathcal{S}, \boldsymbol{p}) p(\boldsymbol{p} \mid \alpha_1) \mathrm{d}\boldsymbol{p} \right\} \cdot p(\mathcal{S} \mid \gamma) \cdot T!, \tag{S3}$$

where $\mathcal{S} = \{S_1, \ldots, S_T\}$ is a ordered partition of N subjects, obtained by randomly ordering parts or blocks of $\mathcal{C}$ uniformly over $T!$ possible choices and $p(\mathcal{S} \mid \gamma) \cdot T! = p(\mathcal{C} \mid \gamma)$.

The prior for the number of components K serves to regularize the number of clusters $T = |\mathcal{C}|$ among observed subjects (see Miller and Harrison (2017, Equation 3.6)). Because $\widetilde{\mathcal{C}}$ is coarser than $\mathcal{C}$, a exponentially decaying prior on K then encourages a small number

of scientific clusters $\widetilde{\mathcal{C}}$ among N subjects which results in using fewer component specific parameters to fit finite samples and improves estimation of unknown H^* and Q.

A.7 On Prior for Q

In applications where Q is not fully identifiable or encouraged to be different among its rows, we specify sparsity priors for each column of Q to encourage proteins to be specific to a small number of machines. That is, $\mathbb{P}(Q_{m\ell} \mid \{Q_{m',\ell}, m' \neq m\}, \zeta) = 1/\left\{1 + \exp\left\{-\zeta \sum_{1 \leq m' < m'' \leq M^*} Q_{m'\ell} Q_{m''\ell}\right\}\right\}$, where ζ is the canonical parameter characterizing the strength and direction of interactions among m. We either fix ζ to be a negative number, or specify a hyperprior for ζ; In this paper, we fix $\zeta = 0$.

A.8 Joint Distribution

The joint distribution of data $\mathbf{Y} = \{\boldsymbol{Y}_i\}$, true and false positive rates $\boldsymbol{\theta}$ and $\boldsymbol{\Psi}$, Q matrix, and latent state vectors $H = \{\boldsymbol{\eta}_i\}$, denoted by $pr(\boldsymbol{Y}, H = H(H^*, \boldsymbol{Z}), Q, \boldsymbol{\theta}, \boldsymbol{\Psi})$, is

$$\begin{aligned}
&\left\{\prod_{i=1}^{N}\prod_{\ell=1}^{L}\left[\Gamma_{\boldsymbol{\eta}_i,\ell}\theta_\ell^{Y_{i\ell}}(1-\theta_\ell)^{1-Y_{i\ell}} + (1-\Gamma_{\boldsymbol{\eta}_i,\ell})\psi_{\ell,v_i}^{Y_{i\ell}}(1-\psi_{\ell,v_i})^{1-Y_{i\ell}}\right]\right\} \\
&\times \prod_{\ell=1}^{L}\left[\mathsf{TruncatedBeta}(\theta_\ell; a_\theta, b_\theta, (\max_{1\leq v\leq K_\ell^-}\psi_{\ell v}, 1))\prod_{v}\mathsf{Beta}(\psi_{\ell v}; a_\psi, b_\psi)\mathbf{1}\{\boldsymbol{\psi}_\ell \in \Delta\}\right]\cdot \\
&\times f(\alpha_1)\cdot \mathsf{IBP}_M(H^*; \alpha_1, K)\cdot \mathbb{P}(\mathcal{C}; \gamma, p_K(\cdot)), \qquad \text{(S4)}
\end{aligned}$$

where $f(\alpha_1)$ is the density function of the hyperprior of truncated IBP (to at most M columns) parameter α_1 and $\mathbb{P}(\mathcal{C}; \gamma, p_K(\cdot))$ is the prior in the space of partitions of observations.

A.9 On Posterior Summary Given a Pre-specified Q

In applications where Q is known (Example 3), we infer for each subject the probability of having a latent state pattern $\boldsymbol{\eta}$, $\mathbb{P}(\boldsymbol{\eta}_i = \boldsymbol{\eta} \mid \mathbf{Y})$, as estimated by the relative frequency of the event $\boldsymbol{\eta}_i = \boldsymbol{\eta}$ across MCMC iterations: $\frac{1}{B}\sum_{b=1}^{B}\mathbf{1}\{\boldsymbol{\eta}_i^{(b)} = \boldsymbol{\eta}\}, \forall \boldsymbol{\eta} \in \mathcal{A}$ where b indexes the stored MCMC samples obtained in Supplementary Material C.1. Similarly, the posterior

distribution for the total number of positive latent states $\mathbb{P}(\sum_{m=1}^{M} \eta_{im} = z \mid \mathbf{Y})$ is estimated by the empirical frequencies $\frac{1}{B}\sum_{b=1}^{B} \mathbf{1}\{\sum_{m=1}^{M} \boldsymbol{\eta}_{im}^{(b)} = z\}$, $z = 0, \ldots, M$, which in Example 3 represents the number of pathogens infecting the lung of a pneumonia child. To characterize the differential importance of each latent state among clusters, we also compute the posterior probability for m-th state being positive $\mathbb{P}(\boldsymbol{\eta}_{(j)m}^{*} = 1 \mid \{Y_i\})$, $j = 1, \ldots, J'$, for J' largest clusters across MCMC iteration. Note that given Q, no merging or relabeling is required as in Step 3 and 7 in Supplementary Material C.1. The number of scientific clusters $\widetilde{K}$ can also be summarized by its empirical frequencies based on posterior samples.

B Marginal Likelihood $g(C)$

To illustrate the calculation of marginal likelihood $g(C)$, we focus on two-parameter DINO model; see Remark S2 for extensions to general restricted LCMs. Given assignment of subjects to clusters $\mathcal{C}$, the model likelihood in a cluster $C_j \in \mathcal{C}$ is

$$pr\left(\{\boldsymbol{Y}_i, i \in C_j\} \mid \boldsymbol{\eta}_j^{*}, \Theta, \Psi, Q\right) = \prod_{\ell:\xi_{j\ell}=0} \psi_\ell^{n_{j\ell 1}} \left(1-\psi_\ell\right)^{n_{j\ell 0}} \cdot \prod_{\ell:\xi_{j\ell}=1} \theta_\ell^{n_{j\ell 1}} \left(1-\theta_\ell\right)^{n_{j\ell 0}}, \tag{S5}$$

where $n_{j\ell 1} = \sum_{i:Z_i=j} Y_{i\ell}$ and $n_{j\ell 0} = \sum_{i:Z_i=j}(1-Y_{i\ell})$ are the number of positive and negative responses at dimension ℓ for subjects in cluster C_j, and $\xi_{j\ell} = \Gamma_{\boldsymbol{\eta}_j^{*},\ell} = 1 - \prod_{m=1}^{M}(1-\eta_{jm}^{*})^{Q_{m\ell}}$ indicates the true status for $\ell = 1, \ldots, L$ and the product over ℓ is due to conditional independence given a cluster. We obtain the marginal likelihood $g(C)$ for cluster C_j by integrating out latent states $\boldsymbol{\eta}_j^{*}$ in (S5):

$$g(C) = \sum_{\boldsymbol{\eta}\in\{0,1\}^M} pr\left(\{\boldsymbol{Y}_i, i \in C_j\} \mid \boldsymbol{\eta}, \Theta, \Psi, Q\right) \mathbb{P}(\boldsymbol{\eta}_j^{*} = \boldsymbol{\eta} \mid \boldsymbol{p}), \tag{S6}$$

where $\mathbb{P}(\boldsymbol{\eta}_j^{*} = \boldsymbol{\eta} \mid \boldsymbol{p}) = \prod_{m=1}^{M} p_m^{\eta_m}(1-p_m)^{1-\eta_m}$. Note that $g(C)$ factorizes with respect to ℓ when $M = L$ and $Q = I_{L\times L}$ that leads to $\xi_{j\ell} = \eta_{j\ell}^{*}$.

Remark S1. *Computational considerations. One of the computational costs results from the summation under a large M in (S6), or "add" operation over $\boldsymbol{\eta} \in \{0,1\}^M$. The factorization with respect to ℓ allows the summations to be done for each ℓ separately and therefore reduces the number of "add" operations from $\mathcal{O}(2^M)$ to $\mathcal{O}(M)$ (Hoff, 2005, Equation (8)). More*

generally, $g(C)$ also factorizes with respect to blocks that partition $\{1,\ldots,M\}$, $\{\mathcal{M}_u, u = 1,\ldots,U\}$ with $\cup\mathcal{M}_u = \{1,\ldots,M\}$ when the corresponding row blocks of Q are orthogonal ($\check{Q}_u = \vee_{m\in\mathcal{M}_u} Q_{m\star}$, $u = 1,\ldots,U$ are orthogonal), resulting in reduced "add" operations $\mathcal{O}(2^{\max_u|\mathcal{M}_u|}L)$. Given Q, we use Reverse Cuthill-McKee (RCM) algorithm (Cuthill and McKee, 1969) for the M by M matrix $QQ^\top$ to simultaneously rearrange its rows and columns to obtain this block structure.

Remark S2. *To generalize (S5) from two-parameter models to general restricted LCMs, simply replace the first product with $\prod_{\ell:\Gamma_{\boldsymbol{\eta}_j^*,\ell}=0}\left(\psi_{\ell,v(\boldsymbol{\eta}_i^*,\ell)}\right)^{n_{j\ell 1}}\left(1-\psi_{\ell,v(\boldsymbol{\eta}_i^*,\ell)}\right)^{n_{j\ell 0}}$.*

C Details of Posterior Algorithm

C.1 Pre-specified Latent State Dimension $M < \infty$

When the number of components K is unknown, one class of techniques updates component-specific parameters along with K. For example, the reversible-jump MCMC (Green, 1995, RJ-MCMC) works by an update to K along with proposed updates to the model parameters which together are then accepted or rejected. However, designing good proposals for high-dimensional component parameters can be non-trivial. Alternative approaches include direct sampling of K(e.g., Nobile and Fearnside, 2007; McCullagh et al., 2008). Here we build on the algorithm of Miller and Harrison (2017) for sampling clusters with discrete component parameters $\boldsymbol{\eta}_j^*$. We focus on model (6) in Main Paper to illustrate the posterior algorithm.

1. Initialization. Initialize all model parameters from prior distributions. When a $Q_{m\star}$ is initialized to have redundant ones under high true positive rates, the likelihood of a sparse observation $\boldsymbol{Y}_i$ is much lower under $\eta_{im} = 0$ than under $\eta_{im} = 1$. Consequently, the sampling chain will visit $\eta_{im} = 0$, i.e., inactive latent state m, with high probability. To better initialize active latent states, we therefore use a more stringent data-driven initialization for $Q_{\star\ell}$ by $Q_{m\ell} \overset{d}{\sim} \mathsf{Bernoulli}(p), m = 1,\ldots,M$, only if many observations are positive at dimension ℓ: $N^{-1}\sum_i Y_{i\ell} > \tau_1$, where p and τ_1 can be prespecified. In our simulations and data analysis, we set $p = 0.1$ and $\tau_1 = 0.3$.

2. Split-merge update clusters $\mathcal{C}$.

The one-subject-at-a-time, Gibbs-type update is typically slow in exploring a large space of clusterings. In fact, the number of ways to partition N subjects is B_N, referred to as the Bell number and can be computed through the iterative formula $B_{N+1} = \sum_{n=0}^{N} \binom{N}{n} B_n$ with $B_0 = B_1 = 1$ resulting in $B_{50} > 2^{157}$. We remedy this by adding split-merge updates designed for conjugate models (Jain and Neal, 2004) that alter the cluster memberships for many subjects at once.

Because the Gibbs update (15) in Main Paper assigns clusters one subject at a time and updates clusters in a local fashion resulting in potential slow mixing of the sampling chain for $\mathcal{C}$, we use global updates to create or remove clusters for multiple subjects at a time that are likely to be accepted according to a Metropolis-Hastings ratio. We adapt an existing recipe designed for models with priors conjugate to the component-specific parameters (Jain and Neal, 2004), which uses split-merge updates to make global changes to cluster configuration followed by further refinement of clusters via Gibbs update one subject at a time. Given $\boldsymbol{\theta}$, $\boldsymbol{\psi}$, Q and $\mathbf{Y}$, a single split-merge update comprises the following steps:

1a) Randomly choose two observations i and j from N subjects; Let S be the indices of subjects either belonging to C_{Z_i} or C_{Z_j}.

1b) Perform $r = 5$ steps of *intermediate* Gibbs scan (17) restricted to observations in the same clusters as i or j. That is, use (17) to update observation $k \in S \setminus \{i, j\}$ with the constraint that $Z_k \in \{Z_i, Z_j\}$; At the end of intermediate Gibbs scan, we obtain $\boldsymbol{Z}^{\mathsf{launch}}$. In this step, one assigns a subject k in $S \setminus \{i, j\}$ to either the cluster of i or j with probability

$$\begin{aligned}
&\mathbb{P}(Z_k = z \mid \boldsymbol{Z}_{-k}, \mathbf{Y}, \text{ other parameters }) \\
&= \frac{(|C_z| + \gamma) g(C_z \cup \{k\}) / g(C_z)}{(|C_{Z_i}| + \gamma) g(C_{Z_i} \cup \{k\}) / g(C_{Z_i}) + (|C_{Z_j}| + \gamma) g(C_{Z_j} \cup \{k\}) / g(C_{Z_j})}, z \in \{Z_i, Z_j\},
\end{aligned} \tag{S7}$$

1c) Perform a *final* Gibbs scan restricted to observations $S \setminus \{i, j\}$ using (S7) and obtain updated clusters as the proposal states to be used in a Metroplis-Hasting step which we denote by $\boldsymbol{Z}^{\mathsf{cand}}$. We compute the proposal densities $q(\boldsymbol{Z}^{\mathsf{cand}} \mid \boldsymbol{Z})$

and $q(\boldsymbol{Z} \mid \boldsymbol{Z}^{\mathsf{cand}})$; For the non-trivial cases, the proposal densities depend on the random launch state $\boldsymbol{Z}^{\mathsf{launch}}$ and are products of Gibbs update densities in (S7).

1d) Accept or reject the proposed clustering $\boldsymbol{Z}^{\mathsf{cand}}$ with acceptance probability computed from prior ratio (based on two sets of clusters induced by $\boldsymbol{Z}^{\mathsf{cand}}$ vs $\boldsymbol{Z}^{\mathsf{launch}}$), likelihood ratio (given clusters $\boldsymbol{Z}^{\mathsf{cand}}$ vs $\boldsymbol{Z}^{\mathsf{launch}}$ and other population parameters), ratio of proposal densities (from 1c). See Jain and Neal (2004) for the general recipe of computing the acceptance probability.

1e) Perform one complete Gibbs scan (17) of $\boldsymbol{Z}$ for *all* individuals to refine the current state of cluster indicators.

The above is referred to as $(5, 1, 1)$ split-merge update where 5 intermediate Gibbs scans are used to reach launch states $\boldsymbol{Z}^{\mathsf{launch}}$, one Metroplis-Hasting step to accept or reject a candidate clustering $\boldsymbol{Z}^{\mathsf{cand}}$, and one final complete Gibbs scan for all observations to refine the newly obtained cluster (Jain and Neal, 2004).

3. Update individual machine usage profiles $H = \{\eta_{im}\}$. Because subjects within a cluster share latent states $\boldsymbol{\eta}_i = \boldsymbol{\eta}_j^*$, $i \in \{i : Z_i = j\}$ for cluster $j = 1, \ldots, T$, we sample from

$$[\boldsymbol{\eta}_j^* \mid \mathsf{others}] \propto \prod_{m=1}^{M} \{p_m\}^{\eta_{jm}^*} \{1 - p_m\}^{1-\eta_{jm}^*} \cdot \prod_{\ell:\xi_{j\ell}=0} \psi_\ell^{n_{j\ell 1}} (1 - \psi_\ell)^{n_{j\ell 0}} \cdot \prod_{\ell:\xi_{j\ell}=1} \theta_\ell^{n_{j\ell 1}} (1 - \theta_\ell)^{n_{j\ell 0}},$$

where $\xi_{j\ell} = \Gamma_{\boldsymbol{\eta}_j^*,\ell}$ indicates the active or inactive status at dimension ℓ in cluster C_j, $\boldsymbol{p} = \{p_m\}$ are within-cluster prevalence of M latent states and $n_{j\ell 1} = \sum_{i:Z_i=j} Y_{i\ell}$ and $n_{j\ell 0} = \sum_{i:Z_i=j}(1 - Y_{i\ell})$. Because $\boldsymbol{\eta}_j^* \in \{0, 1\}^M$, it is important to move around in this space fast. We currently use multinomial sampling in simplex Δ^{2^M-1}, which can be improved by either Hamming ball sampler or parallel tempering.

We remark on "partner latent states" that motivate merging a subset of rows in $Q^{(b)}$. Let $H^{(b)} = \{\eta_{im}^{(b)}\}$ be an N by M binary matrix that collects latent states for all subjects at iteration t. Let $M_{\mathsf{eff}}^{(b)} = \sum_{m=1}^{M} \mathbb{I}\{\mathbf{1}_N^\top H_{\star m}^{(b)} \neq 0\}$ be the number of nonzero columns in H at t-th MCMC iteration. The identifiability conditions apply only to the first $M_{\mathsf{eff}}^{(b)}$ rows of Q. Condition (C1) and (C3) hold at each iteration regardless of the value of $M_{\mathsf{eff}}^{(b)}$ because $Q \in \mathcal{Q}$ truncated to first $M_{\mathsf{eff}}^{(b)}$ rows remains in $\mathcal{Q}$. At

each iteration, conditions (C1) and (C3) also hold if we collapse two identical columns (m, m') of $H^{(b)}$ to combine two partner machines that are present or absent together among subjects ($\eta_{im}^{(b)} = \eta_{im'}^{(b)}$, $i = 1, \ldots, N$); We set $H_{\star m'}^{(b)} = \mathbf{0}_N$ and the other row $Q_{m\ell}^{(b)} = \max\{Q_{m\ell}^{(b)}, Q_{m'\ell}^{(b)}\}$, $\ell = 1, \ldots, L$. It is easy to verify that this scheme preserves conditions (C1) and (C3) and readily generalizes to cases where more than two columns of $H^{(b)}$ are identical. In the population, the diversity assumption $\mathcal{A} = \{0, 1\}^M$ does not hold if two latent states always positive together. When external knowledge is available for two "partner" states with separate known rows in Q, it can be readily integrated into posterior sampling.

4. Sample false positive rates from

$$[\psi_\ell \mid \mathsf{others}] \sim \mathsf{Beta}\left(\sum_i (1-\xi_{i\ell})Y_{i\ell} + a_\psi, \sum_i (1-\xi_{i\ell})(1-Y_{i\ell}) + b_\psi\right) \mathbb{I}\{(0, \theta_\ell)\}, \ell = 1, \ldots, L.$$

Sample true positive rates from

$$[\theta_\ell \mid \mathsf{others}] \sim \mathsf{Beta}\left(\sum_i \xi_{i\ell} Y_{i\ell} + a_\theta, \sum_i \xi_{i\ell}(1-Y_{i\ell}) + b_\theta\right) \mathbb{I}\{(\psi_\ell, 1)\}, \ell = 1, \ldots, L.$$

We also implemented in "`rewind`" specified upper bounds for $\{\psi_\ell\}$ and lower bounds for $\{\theta_\ell\}$ when needed.

5. Update hyperparameter α. Suppose the hyperprior for α is $p(\alpha)$. Then by the marginal distribution of H^* from finite-M IBP (Ghahramani and Griffiths, 2006), we reparametrize in terms of $\beta = \frac{\alpha}{\alpha+1} \in (0, 1)$ and obtain

$$[\beta \mid H^*] \propto p(\beta) \cdot \left(\frac{\beta}{1-\beta}\right)^M \prod_{m=1}^M \frac{\Gamma(s_m + \beta/\{M(1-\beta)\})}{\Gamma(T + 1 + \beta/\{M(1-\beta)\}))},$$

which can be sampled from a dense grid over $(0, 1)$ and $s_m = \sum_{j=1}^T \eta_{jm}^*$ is the number of clusters that m-th latent state is positive. We use Beta distribution $\beta \sim \mathsf{Beta}(a_\beta, b_\beta)$ where $a_\beta = b_\beta = 1$ in our simulations and data analyses.

6. Update prevalence parameters $\boldsymbol{p} = \{p_1, \ldots, p_m\}$ from

$$[\boldsymbol{p} \mid \mathsf{others}] \propto \prod_{m=1}^{M} (p_m)^{n^*_{m1}} (1-p_m)^{n^*_{m0}} \mathsf{Beta}(p_m; \alpha/M, 1), \tag{S8}$$

which we sample independently $p_m \sim \mathsf{Beta}(n^*_{m1} + \alpha/M, n^*_{m0} + 1)$, $m = 1, \ldots, M$.

7. Update machine matrix Q via constrained Gibbs sampler. Update to $Q^{(b)}_{m\ell}$, $\ell = 1, 2, \ldots, L$, $m = 1, 2, \ldots, M$ under two mutually exclusive scenarios:

   1a) Keep $Q^{(t-1)}_{m\ell}$ if one of the three criteria holds: 1) $Q^{(t-1)}_{\star\ell} = \boldsymbol{e}_m$, 2) $1^\top_L Q^{(t-1)}_{m,\star} = 3$ and $Q_{m\ell} = 1$ or 3) $Q^{(t-1)}_{m\ell} = 0$, $Q^{(t-1)}_{\star\ell} = \boldsymbol{e}_m$ and there are only two $\boldsymbol{e}_m$ in the columns of Q.

   1b) Otherwise, flip $Q^{(t-1)}_{m\ell}$ to a different value z with probability $p(z \mid \mathsf{others})/(1-p(z \mid \mathsf{others}))$, where $p(z \mid \mathsf{others})$ is the full conditional distribution

$$\begin{aligned} pr(Q_{m\ell} = z \mid \mathsf{others}) &\propto \prod_{i=1}^{N} pr\left(Y_{i\ell} \mid \{\boldsymbol{\eta}_i\}, Q^{(b)}_{\mathsf{new}}, Q_{m\ell} = z, Q^{(t-1)}_{\mathsf{old}}, \theta_\ell, \psi_\ell\right) \\ &= \prod_{i:\xi_{i\ell}=1} \theta_\ell^{n'_{1\ell 1}} (1-\theta_\ell)^{n'_{1\ell 0}} \cdot \prod_{i:\xi_{i\ell}=0} \psi_\ell^{n'_{0\ell 1}} (1-\psi_\ell)^{n'_{0\ell 0}}, z = 0, 1, \end{aligned}$$

   where $n'_{1\ell 1} = \sum_{i=1}^N \xi_{i\ell} Y_{i\ell}$, $n'_{1\ell 0} = \sum_{i=1}^N \xi_{i\ell}(1-Y_{i\ell})$, $n'_{0\ell 1} = \sum_{i=1}^N (1-\xi_{i\ell}) Y_{i\ell}$, $n'_{0\ell 0} = \sum_{i=1}^N (1-\xi_{i\ell})(1-Y_{i\ell})$, and $Q^{(b)}_{\mathsf{new}}$ and $Q^{(t-1)}_{\mathsf{old}}$ represent entries of Q that have and have not been updated, respectively.

   2) Permute the rows of $Q^{(b)}$ by natural ordering of binary codes $\{Q_{m\star}, m = 1, \ldots, M\}$ represented in binary system. We order the rows of $Q^{(b)}$ by decreasing order of M-dimensional vector $Q^{(b)}\boldsymbol{v}$ where $\boldsymbol{v} = (2^{L-1}, 2^{L-2}, \ldots, 1)^\top$. We only do so after all the MCMC iterations.

   Condition (C1) guarantees that once $Q^\top$ is written in left-ordered form (Ghahramani and Griffiths, 2006), the bottom row of Q corresponds to a row with a positive *ideal* response at the smallest dimension $\ell_{(1)} = \arg\min_\ell \{Q_{m\ell} = 1, \forall m, \ell\}$, which if shared by more than one row, then the row having a postive ideal response at the second lowest dimension $\ell_{(2)} = \arg\min_{\ell:\ell>\ell_{(1)}} \{Q_{m\ell} = 1, \forall m, \ell\}$ is placed at the bottom row; this scheme of ordering the rows of Q will always

succeed according to (C1).

Finally, suppose at iteration s, the MCMC algorithm produces latent states unused by any observation: $\mathcal{M}^{\mathsf{non},(\mathsf{b})} = \{m' : \sum_i \eta_{im'}^{(b)} = 0\}$. We reset to zeros the subset of rows of Q corresponding to the unused latent states at an iteration. Given the sampled $\boldsymbol{\eta}_i^{(b)}$, the corresponding set of rows $Q_{\mathcal{M}^{\mathsf{non}}}^{(b)} = \{Q_{m\star}^{(b)}, m \in \mathcal{M}^{\mathsf{non},(\mathsf{b})}\}$ does not enter likelihood. We re-initiate $Q_{\mathcal{M}^{\mathsf{non}}}^{(b)}$ which upon sequential Gibbs scans create new machines that may enter and improve the likelihood at the next iteration. In our experiments, resetting $Q_{m\star}^{(b)}$ side-steps the difficulty of splitting a sampled machine that is populated with too many ones. Resetting is also practically easier to implement compared to a fine-tuned split-merge algorithm applied to the rows of Q in tandem with simulated annealing which are designed for a more complex time series segmentation tasks (e.g., Fox et al., 2014).

Convergence checks. In simulations and data analysis, we ran three MCMC chains each with a burn-in period of $10,000$ iterations followed by $10,000$ iterations stored for posterior inference. We look for potential non-convergence in terms of Gelman-Rubin statistic (Brooks and Gelman, 1998) that compares between-chain and within-chain variances for each model parameter where a large difference ($R_c > 1.1$) indicates non-convergence; We also used Geweke's diagnostic (Geweke and Zhou, 1996) that compare the observed mean for each unknown variable using the first 10% and the last 50% of the stored samples where a large Z-score indicates non-convergence ($|Z| > 2$). In our simulations and data analyses, we observed fast convergence (many satisfied convergence criteria within $2,000$ iterations) that led to well recovered clusters and Q matrices (results not shown here).

C.2 Algorithm under $M = \infty$

This section presents the algorithm without the need to pre-specify the exact or an upper bound of the number of factors M. The algorithm adapts the slice sampler for infinite factor model (Teh et al., 2007) which performs adaptive truncation of the infinite model to finite dimensions and avoids approximation of the Indian Buffet Process (IBP) prior for H^*. The algorithm builds on the *semi-ordered* representation of the IBP, where the probabilities of

active states are *non-ordered* and the probabilities of inactive states truncated to a random number M^0 are *ordered*. We use this algorithm to infer the number of active states.

0. Initialize the number of active states M^+, the random truncation level for inactive states $M^0 = 0$. Initialize Q with an appropriate $M^* = M^+ + M^0$ by L binary matrix; Initialize the IBP hyperparameter α; Initialize p of length M^* to be the vector of the probabilities for each state being used (if the initial $M^0 = 0$ as recommended, then p needs not be ordered). Initiate H^* as $(T_{\max} + 3)$ by $M_{\max}$ matrix with all zeros, where $T_{\max}$ and $M_{\max}$ are the guessed maximum number of clusters and truncated number of states the algorithm will visit across iterations. Neither $T_{\max}$ nor $M_{\max}$ is introduced to approximate any probabilistic distribution: one can increase both numbers as appropriate at the expense of extra memory.

   <u>Repeat steps 1 to 10 below for iterations $b = 1, \ldots, B$:</u>

1. Gibbs update cluster indicators $\boldsymbol{Z} = \{Z_i, i = 1, \ldots, N\}$ and the cluster-specific sizes $|\mathcal{C}_j|, j = 1, \ldots, t$, where t is the number of unique values in $\boldsymbol{Z}$

2. For Iteration 1, update H^* elementwise for $t \cdot M^*$ elements corresponding to the currently non-empty clusters and the current truncation level M^* for the number of factors; Otherwise, update H^* by the full conditional distribution given other parameters including the slice variable s:

$$pr(\eta^*_{jm} = z \mid \text{others}) \propto \frac{p_m}{p^+_{\min}} \times$$
$$\prod_{m=1}^{M} \{p_m\}^{\eta^*_{jm}} \{1 - p_m\}^{1-\eta^*_{jm}} \cdot \prod_{\ell:\xi_{j\ell}=0} \psi_\ell^{n_{j\ell 1}} (1 - \psi_\ell)^{n_{j\ell 0}} \cdot \prod_{\ell:\xi_{c\ell}=1} \theta_\ell^{n_{j\ell 1}} (1 - \theta_\ell)^{n_{j\ell 0}},$$

   for $z = 0, 1$, $m = 1, \ldots, M^+$, where $p^+_{\min} = p^+_{\min}(H^*, \{p_m, m = 1, 2, \ldots, \}) = \min_{1 \le m \le M^+}\{p_m\}$ depends on η^*_{jm} and is the normalizing constant for the uniform distribution of the slice variable: $pr(s \mid H^*, \{p_m, m = 1, 2, \ldots, \}) = \mathbf{1}_{\{0 \le s \le p^+_{\min}\}}/p^+_{\min}$. For example, given s one must set to zero any column $m \in \{1, \ldots, M^+\}$ in H^*, $\{\eta^*_{jm}, j = 1, \ldots, t\}$ whenever $p_m < s$.

3. Update Q matrix (M^* by L) as in Step 6 in Section C.1;

4. Update the number of active factors (M^+) by finding the number of columns in H^* with non-zero column sums.

5. Update unordered $\{p_m, m = 1, \ldots, M^+\}$ by $p_m \sim \mathsf{Beta}(\sum_{j=1}^t \eta_{jm}^*, 1 + t - \sum_{j=1}^t \eta_{jm}^*)$, $m = 1, \ldots, M^+$;

6. Update slice variable $s \sim \mathsf{Uniform}(0, \min_m p_m)$;

7. Starting from $m = 1$, sample

$$p_{(m)}^0 \mid p_{(m-1)}^0 \sim \exp\{\alpha \sum_{j=1}^t (1 - p_{(m)}^0)^j\}(p_{(m)}^0)^\alpha (1 - p_{(m)}^0)^N \cdot \mathbf{1}_{\{0 \le p_{(m)}^0 \le p_{(m-1)}^0\}},$$

   until $p_{(M^0+1)}^0 < s$, where $p_{(0)}^0 = 1$. Use adaptive rejection sampling (ARS, Gilks and Wild, 1992) to sample from this distribution iteratively for $m = 1, \ldots, M^0$, where $M^0 > 0$ only when $p_{(1)}^0 > s$;

8. If $M^0 > 0$, update p by concatenating the old p and p^0; update $M^* = M^+ + M^0$;

9. Pad H^* with M^0 columns of zeros to its right; Subset the rows of Q to those M^+ factors and pad it with M^0 extra rows sampled from an appropriate initialization sampler;

10. Update other parameters $\boldsymbol{\theta}$, $\boldsymbol{\psi}$, α as in Section C.1.

D Likelihood-based identifiability conditions given $\widetilde{K}$, M and Γ (or Q)

Given Γ, Gu and Xu (2018) established that the separability of Γ is sufficient and necessary for identifying $\boldsymbol{\pi}_{\widetilde{K}}$ under two-parameter models for *known* conditional response probabilities Λ; If Γ is inseparable, $\boldsymbol{\pi}_{\widetilde{K}}$ is identified up to equivalent classes defined by identical rows in Γ (in this paper, we transposed Γ used in Gu and Xu (2018)). When Λ is unknown, Gu and Xu (2018) established sufficient conditions for $\boldsymbol{\pi}_{\widetilde{K}}$-partial identifiability (strictly identify Λ but identify $\boldsymbol{\pi}_{\widetilde{K}}$ up to equivalent classes defined by identical rows in Γ). For Q-restricted two-parameter models, if $\mathcal{A}$ is saturated and Γ is separable, then these conditions become

minimal, i.e. sufficient and necessary conditions: 1) ≥ 3 items per latent state and 2) $Q = [I_M, Q_1^\top]$ where Q_1 has distinct columns.

For multi-parameter models, separability of Γ is sufficient for identifying $\boldsymbol{\pi}_{\widetilde{K}}$ given known Λ. $\boldsymbol{\pi}_{\widetilde{K}}$ will be strictly identifiable given two technical conditions (Gu and Xu, 2018, C3 and C4) - Condition (C3) implies separability of Γ which could be true for Q-RLCM induced Γ with unsaturated $\mathcal{A}$ and without single-attribute items in Q. They also established "generic identifiability" results for Λ and $\boldsymbol{\pi}_{\widetilde{K}}$ when Γ is inseparable: as long as one can flip entries to satisfy two technical conditions. The notion of "generic identifiability" is introduced, because the identifiability results for multi-parameter models hold except on a Lebesgue measure-zero set where the models are reduced to two-parameter models. For the special cases of Q-restricted model (saturated), the two technical conditions do not require the Q-matrix to contain an identity submatrix and provides a flexible new condition for generic identifiability under various Q-matrix structures; the results are generically identifiable up to label swapping among those latent classes that have the same row vectors in the Γ-matrix.

E Additional simulated example: removing irrelevant features reduces the noise and improves cluster estimation

When Q is unknown, the proposed method for scientifically structured clustering includes an additional step for sampling Q. A zero column in Q, say column ℓ, indicates irrelevance of ℓ-th dimension because all positive observations at that dimension will be false positives. By estimating which columns are zeros, our algorithm removes irrelevant features when clustering observations.

Clustering multivariate binary data on a subset of features reduces the impact of noise introduced by less important features and therefore can be superior to all-feature clustering methods such as the standard latent class analysis. For example, in model (2) with $Q = I_{L\times L}$, irrelevant features $\mathcal{L}^c = \{\ell : \Gamma_{\star\ell} = \mathbf{0}\}$ ideally would not enter likelihood ratio calculations when assigning observations to clusters. Indeed, let $R_{kk'}(\boldsymbol{Y} = \boldsymbol{y})$ be the log relative probabilities of assigning an observation $\boldsymbol{Y}_i$ to cluster k ($\mathcal{C}_{-i}^{(k)}$) versus k' ($\mathcal{C}_{-i}^{(k')}$) given

other parameters and clustering $\mathcal{C}_{-i}$ can be Taylor approximated by

$$R_{kk'}(\boldsymbol{Y}_i) \approx \log \frac{|\mathcal{C}_{-i}^{(k)}| + \gamma}{|\mathcal{C}_{-i}^{(k')}| + \gamma} + \sum_{\ell=1}^{L} p_\ell \log \left(\frac{\widehat{\theta}_{(k)\ell}}{\widehat{\theta}_{(k')\ell}} \right)^{Y_{i\ell}} \left(\frac{1 - \widehat{\theta}_{(k)\ell}}{1 - \widehat{\theta}_{(k')\ell}} \right)^{1-Y_{i\ell}}, \tag{S9}$$

where the terms corresponding to irrelevant features become negligible if $\widehat{\theta}_{(k)\ell} \approx \psi_\ell$. The response probabilities at irrelevant dimensions ($\{\psi_\ell : \ell \in \mathcal{L}^c\}$) are nevertheless estimated with error and contribute to noise in assigning each observation to an existing cluster. $R_{kk'}(\boldsymbol{Y}) > 0, = 0, < 0$ indicate assignment of observation $\boldsymbol{Y}$ to cluster k more, equally and less likely than to cluster k', respectively. Consider a triple of observations $(\boldsymbol{Y}_1, \boldsymbol{Y}_2, \boldsymbol{Y}_3)$ where the first (cluster k') and the rest (cluster k) belong to two distinct clusters, respectively. The probability of clustering $\boldsymbol{Y}_1$ and $\boldsymbol{Y}_2$ into their respective true clusters is $p_{12} = (1 - \mathsf{expit}\{R_{kk'}(\boldsymbol{y}_1)\})\mathsf{expit}\{R_{kk'}(\boldsymbol{y}_2)\}$; the probability of assigning $\boldsymbol{Y}_2$ and $\boldsymbol{Y}_3$ into the same true cluster is $p_{23} = \mathsf{expit}\{R_{kk'}(\boldsymbol{y}_2)\}\mathsf{expit}\{R_{kk'}(\boldsymbol{y}_3)\}$. Here we have used lower case $\boldsymbol{y}_i$ to represent the sub-vector of $\boldsymbol{Y}_i$ that entered the calculation in (S9).

We simulated $L_1 = 5$ relevant dimensions and $L_2 = 40$ irrelevant dimensions $\mathcal{L}^c = \{6, \ldots, 45\}$. To mimic the noisy estimates of the response probabilities in cluster k and k', we simulated $\widehat{\theta}_{(k)\ell} = (\log \boldsymbol{r}, \log \boldsymbol{\epsilon})$ and $\widehat{\theta}_{(k')\ell} = (\log \boldsymbol{r}', \log \boldsymbol{\epsilon}')$ where $r_{\ell 1}, \ldots, r_{\ell, L_1} \overset{d}{\sim} \mathsf{Beta}(0.1N_k, 0.9N_k)$, $r'_{\ell 1}, \ldots, r'_{\ell, L_1} \overset{d}{\sim} \mathsf{Beta}(0.9N_{k'}, 0.1N_{k'})$ and $\epsilon_{\ell 1}, \ldots, r_{\ell, L_2} \overset{\mathsf{iid}}{\sim} \mathsf{Beta}(0.1N_k, 0.9N_k)$ and $\epsilon'_{\ell 1}, \ldots, \epsilon'_{\ell, L_2} \overset{\mathsf{iid}}{\sim} \mathsf{Beta}(0.1N_{k'}, 0.9N_{k'})$. We set $N_k = N_{k'} = 20$. Given $\{\widehat{\theta}_{(k)\ell}\}$ and $\{\widehat{\theta}_{(k')\ell}\}$, we draw observations from two classes that have response probability profiles ($\boldsymbol{Y}_2$ and $\boldsymbol{Y}_3$ from $\{\theta_{(k)\ell}, \ell = 1, \ldots, L\} = (\underbrace{0.9, \ldots, 0.9}_{L_1}, \underbrace{0.1, \ldots, 0.1}_{L_2})$ and $\boldsymbol{Y}_1$ from $\{\theta_{(k')\ell}, \ell = 1, \ldots, L\} = (\underbrace{0.1, \ldots, 0.1}_{L_1}, \underbrace{0.1, \ldots, 0.1}_{L_2})$).

Based on $R = 100$ replications, Figure S2 shows $R = 100$ values of p_{12} (left) and $R = 100$ values of p_{23} (right) computed by setting $\{\boldsymbol{y}_i, i = 1, 2, 3\}$ to be the irrelevant, all and relevant features in the data vector $\{\boldsymbol{Y}_i, i = 1, 2, 3\}$, respectively.

By selecting relevant features, the model improves our ability to separate observations from distinct clusters and group observations that belong to the same cluster. On the left panel, the all-feature p_{12} values are pulled towards zero (towards left) that favors assigning $\boldsymbol{Y}_2$ to cluster k and $\boldsymbol{Y}_1$ to cluster k'. On the right panel, the all-feature p_{23} values are pulled towards one (towards right) that favors clustering $\boldsymbol{Y}_2$ and $\boldsymbol{Y}_3$ together in the true cluster (k).

In practice, the relevant features are of course to be inferred from data, by their observed

marginal independence from the rest of the measured features. The improvements of clustering using subset clustering with *inferred* subsets can be seen from in Figure 2 in Main Paper by the superior clustering performance in (f) under feature selection compared to (e) obtained without selecting features.

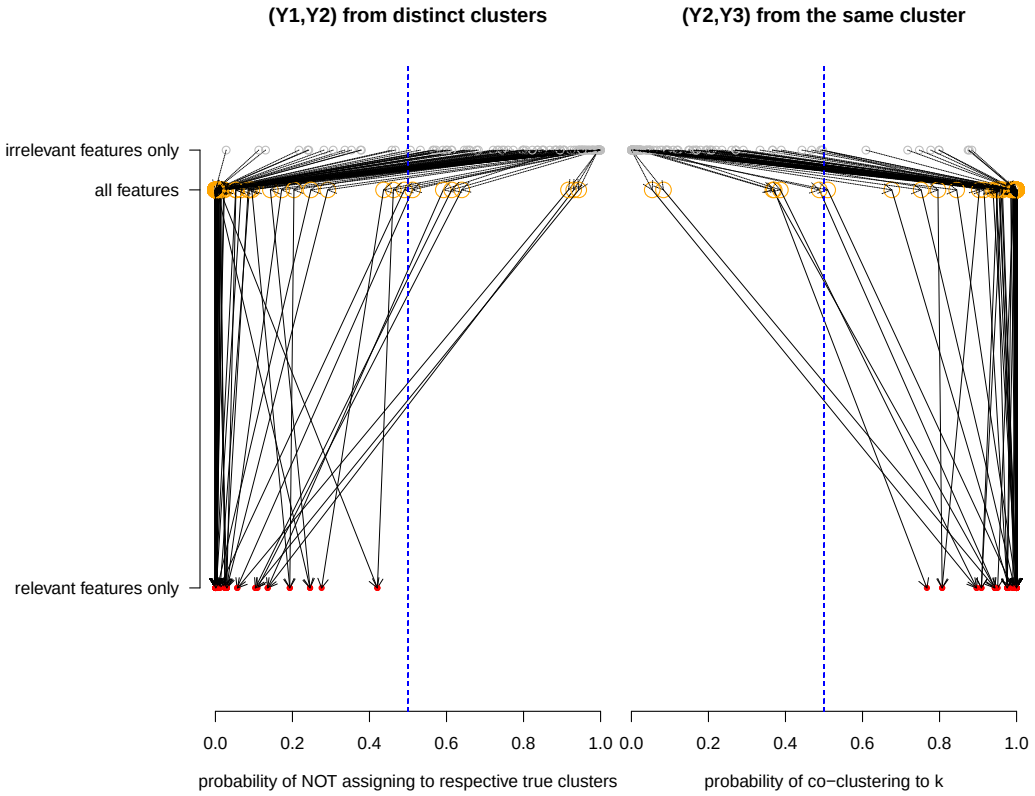

Figure S2: Removing irrelevant features improves estimation of clusters. *Left*) 100 random pairs of observations drawn from distinct clusters; the probability of them not being clustered correctly is lowered (pulled towards zero) once the irrelevant features are removed. *Right*) 100 random pairs of observations drawn from the same cluster; the probability of co-clustering to the correct cluster is increased towards one once the irrelevant features are removed.

E.1 Additional Figures and Tables

Model Specification			Methods (examples): Restricted LCM, Bayesian	Restricted LCM, non-Bayesian	LCM, Classical	LCM, Nested Partially†
latent state variables ($\boldsymbol{\eta}_i \in \mathcal{A} \subset \{0,1\}^M$; **#latent classes:** $\widetilde{K} = \mid \mathcal{A} \mid$)	$\widetilde{K}$ known	$\mathcal{A}$ pre-specified	$\mathcal{A} = \{0,1\}^M$: Chen et al. (2017)	$\mathcal{A} = \{0,1\}^M$: Xu (2017); $\mathbf{0}_M \in \mathcal{A} \subsetneq \{0,1\}^M$: Leighton et al. (2004), Gu and Xu (2018)	Lazarsfeld (1950)⋆, Anderson (1954)⋆, Goodman (1974)⋆ Erosheva et al. (2007)†,‡, Garrett and Zeger (2000)†	$\mathbf{0}_M \in \mathcal{A}$ and partially observed some of $\{i : \boldsymbol{\eta}_i = \mathbf{0}_M\}$: Wu et al. (2017b)
		$\mathcal{A}$ unknown	(proposed)	Miettinen et al. (2008)# ($Q = I_{L\times L}$)	-	-
	$\widetilde{K}$ unknown	$\mathcal{A}$ unknown	(proposed)	-	Dunson and Xing (2009)†	Hoff (2005)
design matrix ($\Gamma = (\Gamma_{\boldsymbol{\eta},\ell}) \in \{0,1\}^{\widetilde{K}\times L}$)	Q-matrix ($\Gamma = \Gamma(\boldsymbol{\eta}, Q)$)	known	(proposed)	Xu (2017)	✓: $Q = \mathbf{1}_{M\times L}$	Wu et al. (2017b); Hoff (2005): $Q = I_{L\times L}$
		unknown	(proposed), Chen et al. (2017), Rukat et al. (2017)	Xu and Shang (2018), Chen et al. (2015)	-	-
measurement process ($[\boldsymbol{Y}_i \mid \boldsymbol{\eta}_i, \Gamma, \boldsymbol{\Lambda}]$)	local indep. given $\boldsymbol{\eta}_i$	yes	(proposed)	✓	✓	Wu et al. (2016)
		no	-	-	Pepe and Janes (2006), Albert et al. (2001)	Wu et al. (2017b)
	(K_ℓ^+, K_ℓ^-)	$(=1, =1)$	(proposed), Chen et al. (2017), Rukat et al. (2017), Wu et al. (2016)	Junker and Sijtsma (2001), Templin and Henson (2006)	-	Wu et al. (2016)
		$(\geq 1, =1)$	-	-	-	Hoff (2005)
		$(=1, \geq 1)$	(proposed)	De La Torre (2011), Henson et al. (2009)	-	-
		$(\geq 1, \geq 1)$	-	-	-	Wu et al. (2017b)
		$(\geq 1, =0)$	-	-	✓	-

†: Bayesian approach. ‡: has equivalent LCM formulation. #: non-probabilistic ⋆: early applications.
✓: applies to all in the column (except for other rows in the same row block)

Table S2: Comparison of variants of latent class analysis of multivariate binary data.

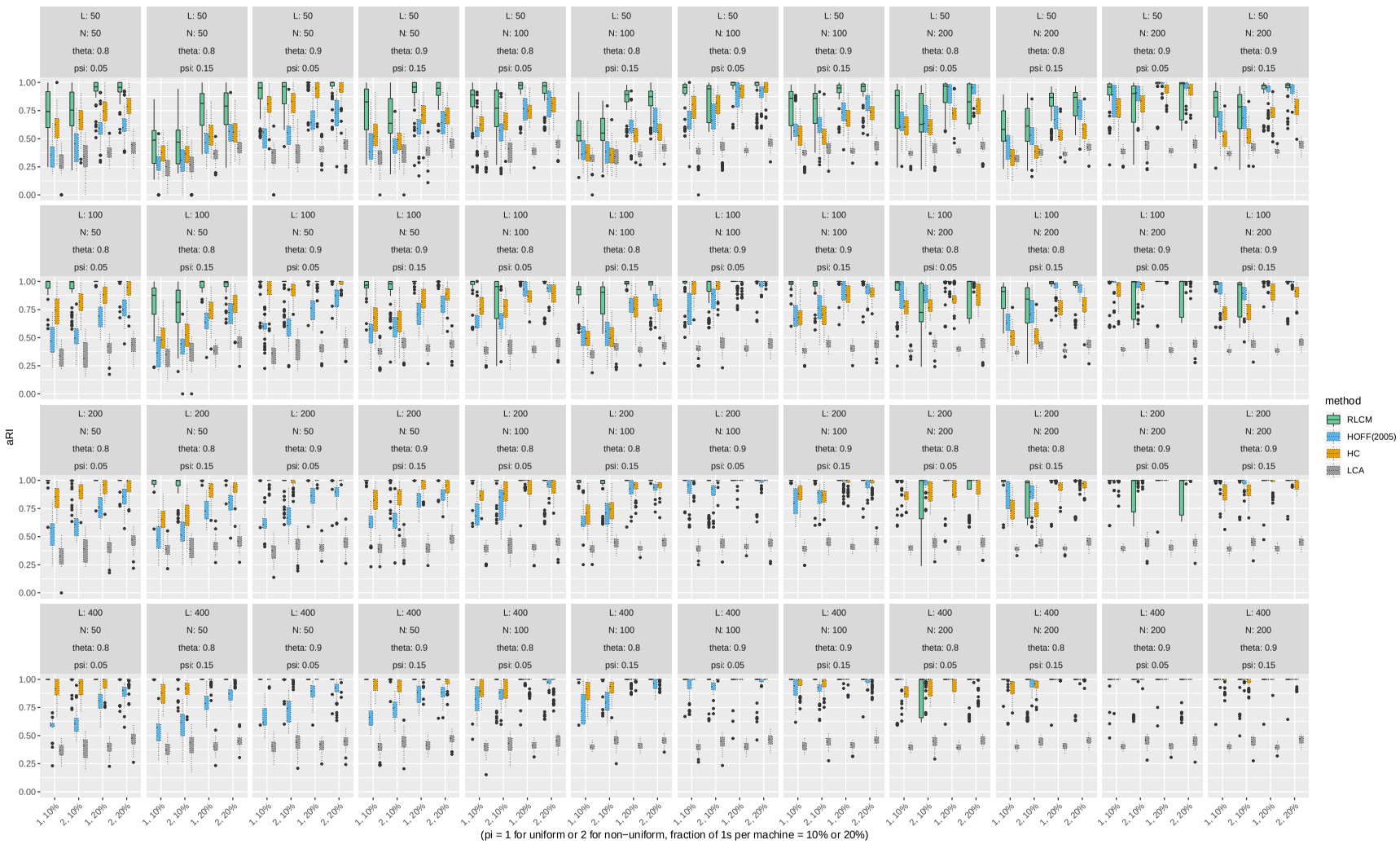

Figure S3: Based on $R = 60$ replications for each parameter setting, Bayesian RLCM (boxplots with solid lines) most accurately recovers the true clusters compared to subset clustering (Hoff, 2005) hierarchical clustering (HC) and traditional Bayesian latent class analysis (LCA) (from the left to the right in each group of four boxplots). This figure expands Figure 2 in Main Paper over more parameter settings.

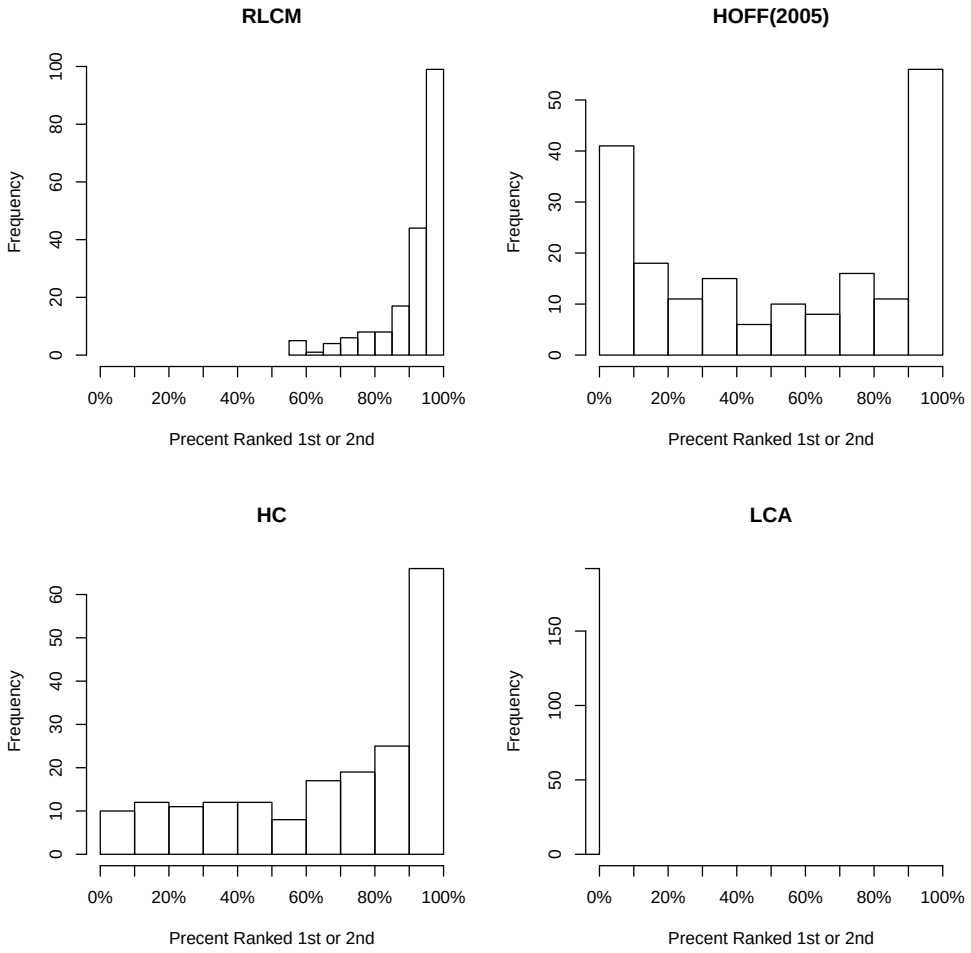

Figure S4: For each of four clustering methods (Bayesian RLCM, Hoff (2005), HC, Bayesian LCA), the percent being ranked the first or the second in terms of the mean aRI averaged across $R = 60$ replications (Section 4.2.1 in Main Paper). Each histogram is produced for the $1,920$ combinations of parameters investigated in Section 4.1 in Main Paper.

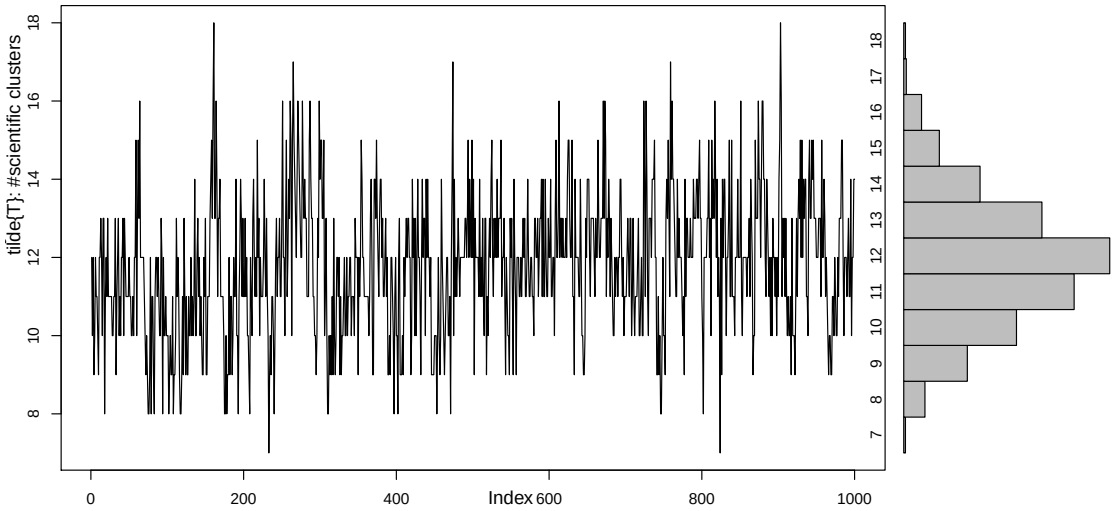

Figure S5: MCMC samples of the number of scientific clusters ($\widetilde{C}$) with its marginal posterior on the right margin.

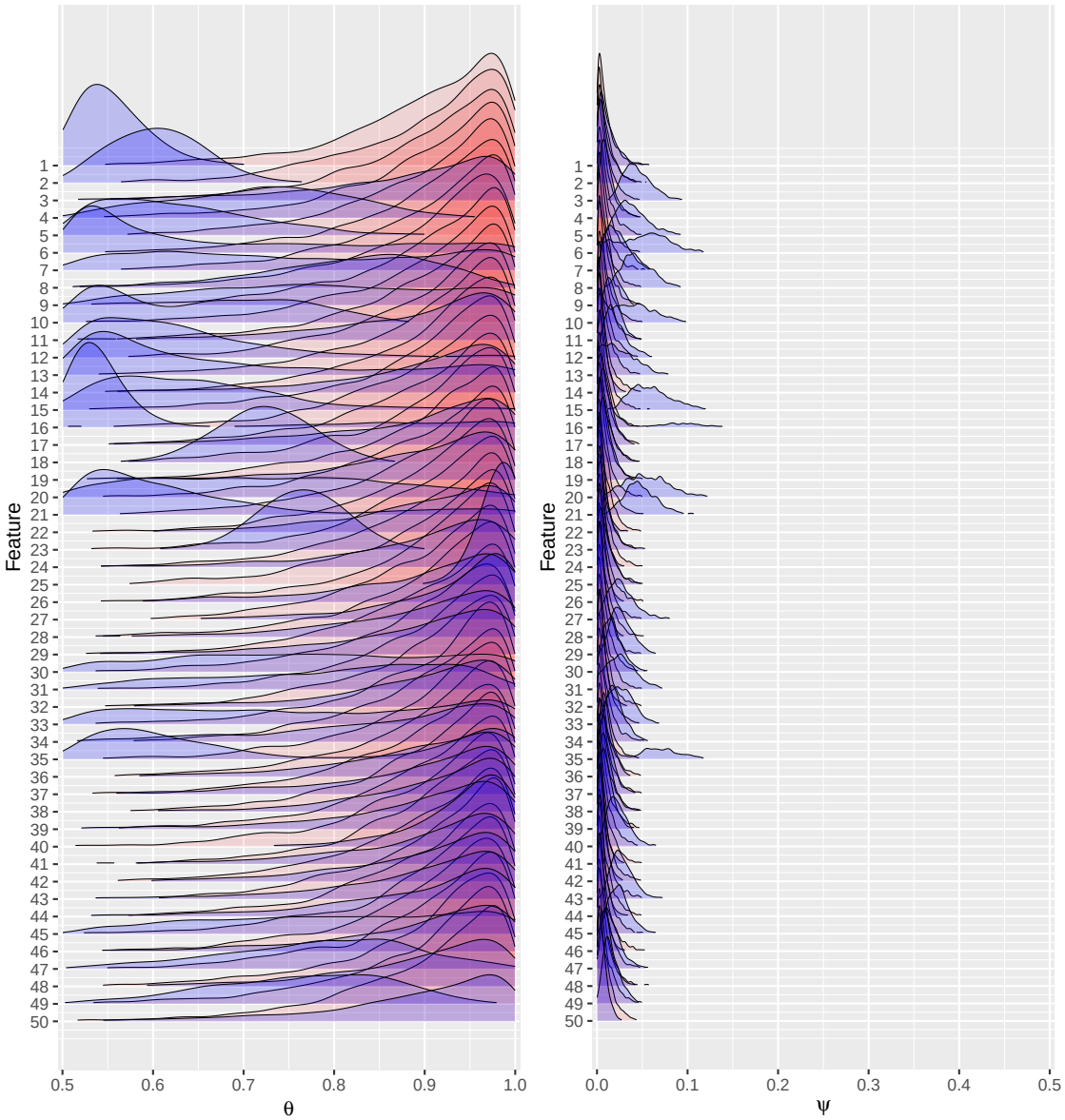

Figure S6: Prior vs posterior for all true positive rates $\{\theta_\ell\}$ (left) and false positive rates $\{\psi_\ell\}$ (right).

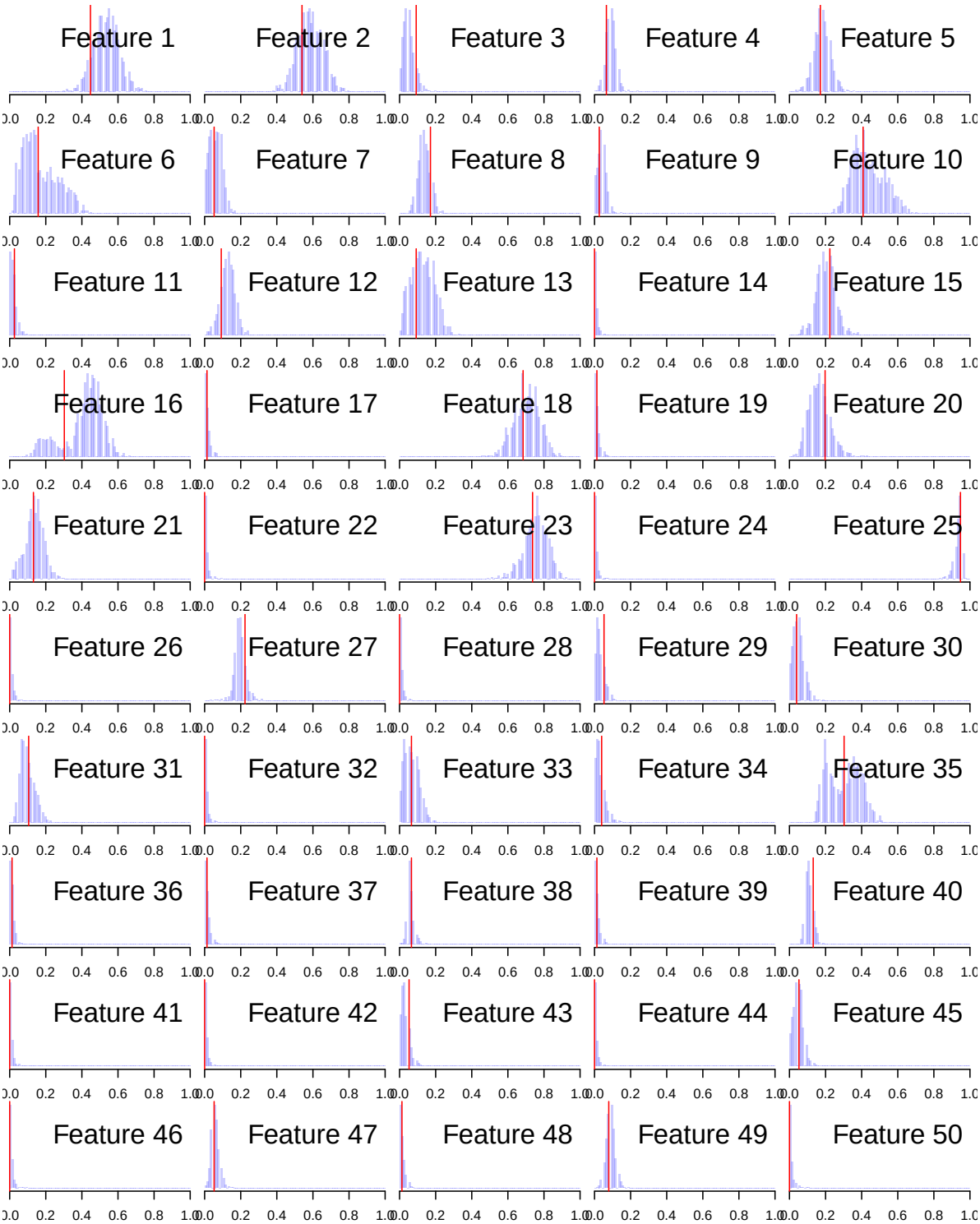

Figure S7: Observed marginal positive rate (solid vertical line) plotted against the posterior predictive distributions for $L = 50$ landmarks in Example 1.

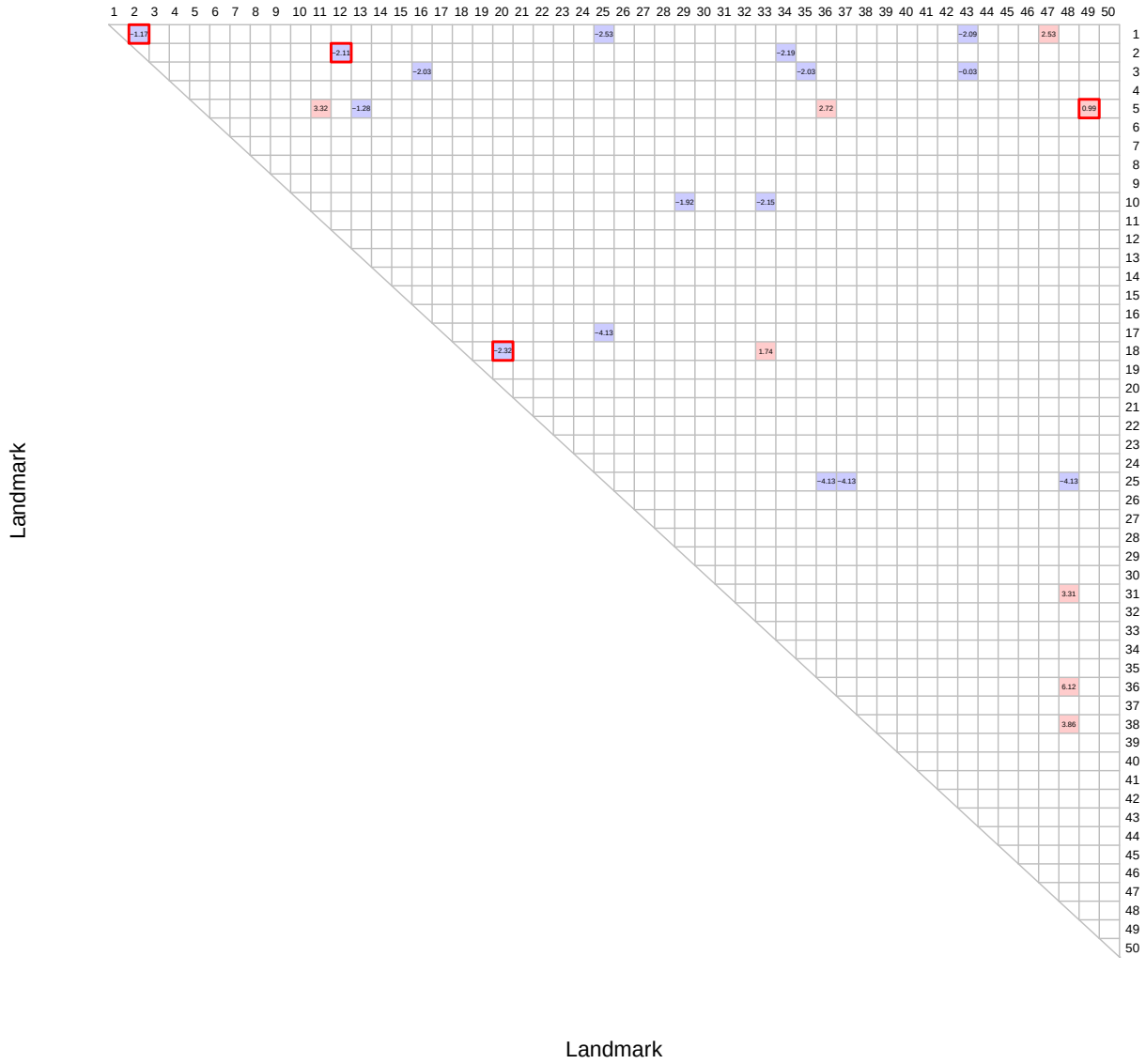

Figure S8: Significant deviations of model predicted log odds ratios (LOR) from the observed LOR. A blank cell indicates a good model prediction for the observed pairwise LOR ($|\mathsf{SLORD}| < 2$); A red (blue) cell indicates model under- (over-) fitting $\mathsf{SLORD} > 2 (< -2)$, where standardized LOR difference (SLORD) is defined as the observed LOR for a pair of landmarks minus the mean LOR for the predictive distribution value divided by the standard deviation of the LOR predictive distribution. A red box indicate that the pair of landmarks have cell counts in the 2 by 2 observed marginal table all greater than 5.