Supplementary Materials to "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.

Summary

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 A1 contains remarks, Section A2 details the posterior algorithms for pre-specified M (Section A2.1) and infinite M (Section A2.3), respectively. Section A3 presents posterior summaries that we use in simulation and data analysis. Section A4 illustrates through simulations the benefit of removing irrelevant features. Section A5 contains additional data analysis results. Finally, Section A6 collects a table for variants of LCMs as well as figures for model results on the data analysis in Main Paper.

A1 Remarks

A1.1 Other Examples in Psychology and Epidemiology that Require Scientifically-Structured Clustering

We refer to the motivating example of estimating subsets of autoimmune disease patients in Main Paper as “Example 1”. Example 2 is related 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).

A1.2 General Technical Formulation of RLCMs: Imposing Restrictions for Scientifically-Structured Classes

RLCMs impose equality among a subset of response probabilities across classes. The specification of the RLCM likelihood involves two aspects, restriction of the order of between-class response probabilities and the parameterizations.

Order restriction by design matrix. The response probabilities in RLCMs must satisfy certain order constraints specified by a binary design matrix $\Gamma = \{\Gamma_{\boldsymbol{\alpha},\ell}\} \in \{0,1\}^{\widetilde{K}\times L}$ with latent classes and measurements in the rows and columns, respectively. Let $\mathcal{A}_\ell = \{\boldsymbol{\alpha} \in \mathcal{A} : \Gamma_{\boldsymbol{\alpha},\ell} = 1\}$ denote the latent classes with the highest response probability for dimension ℓ according to Γ (Gu and Xu, 2018). That is, if $\mathcal{A}_\ell \neq \emptyset$, we restrict the response probabilities at feature ℓ by

$$\max_{\boldsymbol{\alpha}\in\mathcal{A}_\ell} \lambda_{\boldsymbol{\alpha},\ell} = \min_{\boldsymbol{\alpha}\in\mathcal{A}_\ell} \lambda_{\boldsymbol{\alpha},\ell} > \lambda_{\boldsymbol{\alpha}',\ell}, \ell = 1,\ldots,L, \boldsymbol{\alpha}' \in \mathcal{A}_\ell^c, \tag{S1}$$

where $\mathcal{A}_\ell^c = \mathcal{A} - \mathcal{A}_\ell = \{\boldsymbol{\alpha} \in \mathcal{A} : \Gamma_{\boldsymbol{\alpha},\ell} = 0\}$. Further, there can exist a class $\boldsymbol{\alpha} \in \mathcal{A}$ that gives rise to an all-zero row $\Gamma_{\boldsymbol{\alpha}\star} = \mathbf{0}_{1\times L}$. Restrictions (S1) only specify the relative magnitudes but not the actual values of the response probabilities. Besides, subjects in classes $\mathcal{A}^c$ may have multiple levels of response probabilities. Supplementary Material A1.2.1 below presents an equivalent formulation based on parameters that represent distinct levels of response probabilities.

Q-based design matrix. In this paper, we focus on the special and useful case where design matrix Γ further depends on latent state vectors $\boldsymbol{\alpha}$ and a possibly unknown binary matrix Q of dimension M by L as follows:

$$\Gamma_{\boldsymbol{\alpha},\ell} = \Gamma(\boldsymbol{\alpha}, Q_{\star\ell}),\ \text{for all } \boldsymbol{\alpha} \in \mathcal{A}, \ell = 1,\ldots,L, \tag{S2}$$

where the scientific context motivates specific mathematical forms of $\Gamma(\cdot,\cdot)$ (e.g., $\Gamma_{\boldsymbol{\eta}_i,\ell} = \boldsymbol{\eta}_i^\top Q_{\star\ell}$ in Figure 1, Main Paper).

Finally, it remains to specify the class-specific measurement likelihood function (1) in Main Paper. Given Γ defined by (S2), we parameterize the response probabilities $\lambda_{i\ell} = \lambda_\ell(\boldsymbol{\eta}_i)$ by

$$\lambda_\ell(\boldsymbol{\eta}_i) = \lambda_\ell^R(\boldsymbol{\eta}_i; Q_{\star\ell}, \boldsymbol{\beta}_\ell) \in [0,1], \boldsymbol{\eta}_i \in \mathcal{A}, \ell = 1,\ldots,L, \tag{S3}$$

where we require that $\{\lambda_\ell^R(\boldsymbol{\alpha}) : \boldsymbol{\alpha} \in \mathcal{A}\}$ must satisfy the relative magnitude restriction (S1) across latent classes.

A1.2.1 Equivalent Formulation of RLCM

It is sometimes convenient to formulate RLCM using distinct levels of response probabilities. 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 $\boldsymbol{\eta}_i = \boldsymbol{\alpha} \notin \mathcal{A}_\ell$, 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 (S3). For models with $K_\ell^- = 1$, $v_i = 1$; Otherwise, $\nu(\cdot,\cdot)$ depends on $\mathcal{A}$ (the set of possible patterns 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 (S5) in Supplementary Materials A1.3; The traditional LCM results by setting $Q = 1_{M\times L}$ and under $K^+ + K^- = \widetilde{K}$ for each ℓ).

We therefore have an equivalent formulation for the response probability parameters λ_ℓ^R in (S3):

$$\lambda_\ell^R(\boldsymbol{\eta}_i; Q_{\star\ell}, \boldsymbol{\beta}_\ell) = \{\theta_\ell\}^{\Gamma_{\boldsymbol{\eta}_i,\ell}} \cdot \left\{\boldsymbol{\psi}_{\ell,v(\boldsymbol{\eta}_i,\ell)}\right\}^{1-\Gamma_{\boldsymbol{\eta}_i,\ell}} \in [0,1], \tag{S4}$$

where $\boldsymbol{\beta}_\ell = \{\boldsymbol{\theta} = \{\theta_\ell\}, \Psi = \{\boldsymbol{\psi}_\ell\}\}$ with constraints $\theta_\ell > \psi_{\ell,v}, \forall v = 1, \ldots, K_\ell^-$.

A1.3 Other Examples of RLCM in the Literature

Two-parameter examples of RLCM. Model (2) in Main Paper is a two-parameter RLCM because $K_\ell^+ + K_\ell^- = 2$. 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 error-free 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).

Multi-parameter examples of RLCM. Two-parameter models assume that "$\Gamma_{\boldsymbol{\alpha},\ell} = \Gamma_{\boldsymbol{\alpha}',\ell} = 0$ implies identical response probabilities $\lambda_{\boldsymbol{\alpha},\ell} = \lambda_{\boldsymbol{\alpha}',\ell} = \psi_\ell$", regardless of the distinct patterns $\boldsymbol{\alpha} \neq \boldsymbol{\alpha}'$. In practice, deviation from such assumptions occurs if $\boldsymbol{\alpha}$ has more nonzero elements than $\boldsymbol{\alpha}'$ and requires additional levels of response probability in the class with latent states $\boldsymbol{\alpha}$, 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: $\lambda_{i\ell} = \lambda^R_\ell(\boldsymbol{\eta}_i; \boldsymbol{\beta}_\ell, Q_{\star\ell}) = \operatorname{expit}\left\{\boldsymbol{\beta}_\ell^\top \boldsymbol{h}(\boldsymbol{\eta}_i, Q_{\star\ell})\right\} =$

$$\operatorname{expit}\left\{\beta_{\ell 0} + \sum_{m=1}^{M} \beta_{\ell m}(Q_{m\ell}\eta_{im}) + \sum_{m<m'} \beta_{\ell m m'}(Q_{m\ell}\eta_{im})(Q_{m'\ell}\eta_{im'}) + \ldots + \beta_{\ell 12\ldots M}\prod_{m}(Q_{m\ell}\eta_{im})\right\} \tag{S5}$$

that includes higher order interactions among latent states required by an item (Henson et al., 2009); Here $\operatorname{expit}(x) = \frac{\exp(x)}{1+\exp(x)}$. 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 $Q_{m\ell} = 1$.

A1.4 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 Bernoulli products" with each latent class (defined by $\boldsymbol{\eta}_i$) having *relevant* features at possibly overlapping subsets of features $\mathcal{S}_{\boldsymbol{\alpha}} = \{\ell : \Gamma_{\boldsymbol{\alpha},\ell} = 1\}$, $\boldsymbol{\alpha} \in \mathcal{A}$ (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 $\lambda^*_{i\ell} = \mathbb{P}(Y^*_{i\ell} = 1 \mid -) = 1 - \mathbb{P}(Y_{i\ell} = 1 \mid -) = 1 - \lambda_{i\ell} = \left\{\psi^*_{\ell,v}\right\}^{1-\Gamma^*_{i\ell}} (\theta^*_\ell)^{\Gamma^*_{i\ell}}$.

A1.5 Identifiability Considerations: Posterior Algorithm Design

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 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 few clusters hence a small number of parameters via mixture of finite mixture models (Miller and Harrison, 2017). We now discuss identifiability results based on likelihood function.

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 in data analysis 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": $\pi_{\boldsymbol{\alpha}} > 0, \forall \boldsymbol{\alpha} \in \mathcal{A} = \{0,1\}^M$), 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 Supplementary Material A1.3); 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) $(\Lambda_{\boldsymbol{\alpha},\ell}, \ell > 2M)^\top \neq (\Lambda_{\boldsymbol{\alpha}',\ell}, \ell > 2M)^\top$ for any $\boldsymbol{\alpha} \neq \boldsymbol{\alpha}'$ and $\boldsymbol{\alpha} \succeq \boldsymbol{\alpha}'$,

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.

Xu and Shang (2018) studied identifiability issues for 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) $(\Lambda_{\boldsymbol{\alpha},\ell}, \ell > 2M)^\top \neq (\Lambda_{\boldsymbol{\alpha}',\ell}, \ell > 2M)^\top$ for any $\boldsymbol{\alpha} \neq \boldsymbol{\alpha}'$ and $\boldsymbol{\alpha} \succeq \boldsymbol{\alpha}'$,

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) below 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 A1.3) 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 A2).

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

$$\mathcal{Q} = \{Q \in \{0,1\}^{M\times L} : Q = P_1 Q^{\dagger} P_2,\ Q^{\dagger} = [I_M; I_M; \widetilde{Q}],\ \widetilde{Q}\mathbf{1}_{L-2M} \succeq \mathbf{1}_{L-2M}\}, \tag{S6}$$

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.

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 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 A2.1). 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).

Finally, 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) 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).

A1.5.1 Additional 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.

A1.6 On the prior distribution on H^*

By Beta-Bernoulli conjugacy, we integrate the joint distribution in (4)-(5) in Main Paper $[H^* \mid \boldsymbol{p}][\boldsymbol{p} \mid \alpha_1, \alpha_2]$ over $\boldsymbol{p}$ to obtain the marginal prior:

$$pr(H^* \mid \alpha_1, \alpha_2) = \prod_{m=1}^{M} \frac{(\alpha_1\alpha_2/M)\Gamma(s_m + \alpha_1\alpha_2/M)\Gamma(T - s_m + \alpha_2)}{\Gamma(T + \alpha_2 + \alpha_1/M)}, \tag{S7}$$

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.

A1.7 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 A2.1), 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 our experience in simulation studies. 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 (S7) in Main Paper to infinity and obtain infinite-column prior for H^*; This construction defines the infinite Indian Buffet process (Ghahramani and Griffiths, 2006). Supplementary Material A2.3 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).

A1.8 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 (S11) below in Supplementary Material A2. 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 only 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). This is our strategy in Section 2.3 in Main Paper to specify priors for clusters with an unknown number of classes and unkonwn latent state space $\mathcal{A}$.

A1.9 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 A2 below, we will first 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

$$p(\widetilde{\mathcal{C}} \mid \alpha_1, \gamma) = \sum_{\mathcal{C}:\mathcal{C}\preceq\widetilde{\mathcal{C}}} p(\widetilde{\mathcal{C}} \mid \mathcal{C}, \alpha) \cdot p(\mathcal{C} \mid \gamma) \tag{S8}$$

$$= \sum_{\mathcal{C}:\mathcal{C}\preceq\widetilde{\mathcal{C}}} \binom{2^M}{\widetilde{T}} (\widetilde{T})! \left\{ \int p(H^* \mid \mathcal{S}, \boldsymbol{p}) p(\boldsymbol{p} \mid \alpha_1) \mathrm{d}\boldsymbol{p} \right\} \cdot p(\mathcal{S} \mid \gamma) \cdot T!, \tag{S9}$$

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.

Remark S1. *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. $\widetilde{\boldsymbol{\alpha}}_k, k = 1, \ldots, \widetilde{K}$ 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 (4)-(5) in Main Paper 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.*

A1.10 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 (S6). 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. 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 $\zeta = 0$.

We specify the priors for response probabilities $\Lambda = \{\lambda_{i\ell}\}$ in (S4) to satisfy the monotonic constraints in (S1) 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 $\boldsymbol{\alpha}$ with $\Gamma_{\boldsymbol{\alpha},\ell} = 0$ defined in (S2) and the truncation of θ_ℓ follows from the definition of RLCM (S1). 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 (S5)), 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. 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 A2).

A1.11 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

$$\begin{aligned}
&\left\{\prod_{i=1}^{N}\prod_{\ell=1}^{L}\left[\Gamma_{\boldsymbol{\eta}_i,\ell}\theta_\ell^{Y_{i\ell}}(1-\theta_\ell)^{1-Y_{i\ell}} + (1-\Gamma_{\boldsymbol{\eta}_i,\ell})\psi_{\ell,v_i}^{Y_{i\ell}}(1-\psi_{\ell,v_i})^{1-Y_{i\ell}}\right]\right\} \\
&\times \prod_{\ell=1}^{L}\left[\mathsf{TruncatedBeta}(\theta_\ell; a_\theta, b_\theta, (\max_{1\le v\le K_\ell^-}\psi_{\ell v}, 1))\prod_{v}\mathsf{Beta}(\psi_{\ell v}; a_\psi, b_\psi)\mathbf{1}\{\boldsymbol{\psi}_\ell \in \Delta\}\right]\cdot \\
&\times f(\alpha_1)\cdot \mathsf{IBP}_M(H^*; \alpha_1, K)\cdot \mathbb{P}(\mathcal{C}; \gamma, p_K(\cdot)), \qquad \text{(S10)}
\end{aligned}$$

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.

A2 Details of Posterior Algorithm

A2.1 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 (1-3) 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.

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 cluster estimates) 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 $[\mathbf{Y} \mid \boldsymbol{Z}, \boldsymbol{\theta}, \boldsymbol{\psi}, Q, \boldsymbol{p}]$. 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:

$$\mathbb{P}(Z_i = j \mid -) \propto \begin{cases} (|C| + \gamma) \cdot \frac{g(C \cup \{i\})}{g(C)}, & \textsf{if } C \in \mathcal{C}_{-i}, j = 1, \ldots, |\mathcal{C}_{-i}|, \textsf{ or} \\ \gamma \frac{V_N(t+1)}{V_N(t)} \cdot g(C), & \textsf{if } C = \{i\}, j = |\mathcal{C}_{-i}| + 1, \end{cases} \tag{S11}$$

where $g(C) = g(C; \boldsymbol{\theta}, \boldsymbol{\psi}, Q, \boldsymbol{p}) = \prod_{\ell=1}^{L} pr(\{Y_{i\ell} : i \in C\} \mid \boldsymbol{\theta}, \boldsymbol{\psi}, Q, \boldsymbol{p})$ is the marginal likelihood for data in cluster C (see (S14) in Supplementary Material A2.2 for an illustration using DINO model). If adding subject i to any existing $C \in \mathcal{C}_{-i}$ 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 $\{i\}$.

Because the Gibbs update (S11) 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 (S11) restricted to observations in the same clusters as i or j. That is, use (S11) 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

$$\begin{aligned}
&\mathbb{P}(Z_k = z \mid \boldsymbol{Z}_{-k}, \mathbf{Y}, \text{ other parameters }) \\
&= \frac{(|C_z| + \gamma) g(C_z \cup \{k\}) / g(C_z)}{(|C_{Z_i}| + \gamma) g(C_{Z_i} \cup \{k\}) / g(C_{Z_i}) + (|C_{Z_j}| + \gamma) g(C_{Z_j} \cup \{k\}) / g(C_{Z_j})}, z \in \{Z_i, Z_j\},
\end{aligned} \tag{S12}$$

1c) Perform a *final* Gibbs scan restricted to observations $S \setminus \{i, j\}$ using (S12) 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 (S12).

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 (S11) 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 $H^{(b)} = \{\eta^{(b)}_{im}\}$ be an N by M binary matrix that collects latent states for all subjects at iteration t. Let $M^{(b)}_{\mathsf{eff}} = \sum_{m=1}^{M} \mathbb{I}\{\mathbf{1}_N^\top H^{(b)}_{\star m} \neq 0\}$ be the number of nonzero columns in H at t-th MCMC iteration. The identifiability conditions apply only to the first $M^{(b)}_{\mathsf{eff}}$ rows of Q. Condition (C1) and (C3) hold at each iteration regardless of the value of $M^{(b)}_{\mathsf{eff}}$ because $Q \in \mathcal{Q}$ truncated to first $M^{(b)}_{\mathsf{eff}}$ 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 ($\eta^{(b)}_{im} = \eta^{(b)}_{im'}$, $i = 1, \ldots, N$); We set $H^{(b)}_{\star m'} = \mathbf{0}_N$ and the other row $Q^{(b)}_{m\ell} = \max\{Q^{(b)}_{m\ell}, Q^{(b)}_{m'\ell}\}$, $\ell = 1, \ldots, L$. 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

$$[\boldsymbol{p} \mid \mathsf{others}] \propto \prod_{m=1}^{M} (p_m)^{n^*_{m1}} (1 - p_m)^{n^*_{m0}} \mathsf{Beta}(p_m; \alpha/M, 1), \tag{S13}$$

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) $1_L^\top Q^{(t-1)}_{m,\star} = 3$ 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).

A2.2 Marginal Likelihood $g(C)$

To illustrate the calculation of marginal likelihood $g(C)$ central to the posterior sampling of clusters in (S11), we focus on two-parameter DINO model; see Remark S3 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

$$pr\left(\{\boldsymbol{Y}_i, i \in C_j\} \mid \boldsymbol{\eta}_j^*, \Theta, \Psi, Q\right) = \prod_{\ell:\xi_{j\ell}=0} \psi_\ell^{n_{j\ell 1}} (1-\psi_\ell)^{n_{j\ell 0}} \cdot \prod_{\ell:\xi_{j\ell}=1} \theta_\ell^{n_{j\ell 1}} (1-\theta_\ell)^{n_{j\ell 0}}, \qquad \text{(S14)}$$

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 (S14):

$$g(C) = \sum_{\boldsymbol{\alpha}\in\{0,1\}^M} pr\left(\{\boldsymbol{Y}_i, i \in C_j\} \mid \boldsymbol{\alpha}, \Theta, \Psi, Q\right) \mathbb{P}(\boldsymbol{\eta}_j^* = \boldsymbol{\alpha} \mid \boldsymbol{p}), \qquad \text{(S15)}$$

where $\mathbb{P}(\boldsymbol{\eta}_j^* = \boldsymbol{\alpha} \mid \boldsymbol{p}) = \prod_{m=1}^{M} p_m^{\eta_m}(1-p_m)^{1-\eta_m}$. 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 S2. *Computational considerations. One of the computational costs results from the summation under a large M in (S15), or "add" operation over $\boldsymbol{\alpha} \in \{0,1\}^M$. 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 S3. *To generalize (S14) from two-parameter models to general restricted LCMs, simply replace the first product with $\prod_{\ell:\Gamma_{\boldsymbol{\eta}_j^*,\ell}=0}\left(\psi_{\ell,v(\boldsymbol{\eta}_i^*,\ell)}\right)^{n_{j\ell 1}}\left(1-\psi_{\ell,v(\boldsymbol{\eta}_i^*,\ell)}\right)^{n_{j\ell 0}}$.*

A2.3 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.

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

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:

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

   for $z = 0, 1, m = 1, \ldots, M^+$, where $p^+_{\text{min}} = p^+_{\text{min}}(H^*, \{p_m, m = 1, 2, \ldots, \}) = \min_{1 \le m \le M^+}\{p_m\}$ depends on η^*_{jm} and is the normalizing constant for the uniform distribution of the slice variable: $pr(s \mid H^*, \{p_m, m = 1, 2, \ldots, \}) = \mathbf{1}_{\{0 \le s \le p^+_{\text{min}}\}}/p^+_{\text{min}}$. 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 A2.1;

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

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

until $p^0_{(M^0+1)} < s$, 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 $p^0_{(1)} > s$;

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 A2.1.

A3 Posterior summaries of co-clustering and latent states

Here we focus on posterior co-clustering probabilities $\pi_{ii'} = \mathbb{P}(Z_i = Z_{i'} \mid \mathbf{Y})$, for subjects i, $i' = 1, \ldots, N$. We estimate $\pi_{ii'}$ by the empirical frequencies $\widehat{\pi}_{ii'}$ of subjects i and i' being clustered together across MCMC iterations. For point estimation, we compute the 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^* that minimizes the loss: $b^* = \arg\min_b \|Q^{(b)\top}Q^{(b)} - \frac{1}{B}\sum_{b'=1}^{B} Q^{(b')\top}Q^{(b')}\|_F$ where $\|\cdot\|_F = \sqrt{\sum a_{ij}^2}$ computes matrix Frobenius norm. $Q^\top Q$ is a L by L matrix invariant to relabeling of latent states and represents with its (ℓ, ℓ')-th element the number of positive states required by feature ℓ only when $\ell = \ell'$ or by both of the feature pair (ℓ, ℓ') when $\ell \neq \ell'$. Turning to the inference of $H = \{\boldsymbol{\eta}_i\}$, we generate more posterior samples and reduce Monte Carlo errors of approximating $[H \mid Q = Q^{(b*)}, \mathbf{Y}]$ by refitting a model with $Q = Q^{(b*)}$.

Remark S4. *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* $\boldsymbol{\alpha}$, $\mathbb{P}(\boldsymbol{\eta}_i = \boldsymbol{\alpha} \mid \mathbf{Y})$, *as estimated by the relative frequency of the event* $\boldsymbol{\eta}_i = \boldsymbol{\alpha}$ *across MCMC iterations:* $\frac{1}{B}\sum_{b=1}^{B} \mathbf{1}\{\boldsymbol{\eta}_i^{(b)} = \boldsymbol{\alpha}\}, \forall \boldsymbol{\alpha} \in \mathcal{A}$ *where b indexes the stored MCMC samples obtained in Supplementary Material A2.1. 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 A2.1. The number of scientific clusters* $\widetilde{K}$ *can also be summarized by its empirical frequencies based on posterior samples.*

A4 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 (1-3) in Main Paper with $Q = I_{L\times L}$ in (2), 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 $R_{kk'}(\boldsymbol{Y}_i)$ 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

$$R_{kk'}(\boldsymbol{Y}_i) \approx \log \frac{|\mathcal{C}_{-i}^{(k)}| + \gamma}{|\mathcal{C}_{-i}^{(k')}| + \gamma} + \sum_{\ell=1}^{L} p_\ell \log \left(\frac{\widehat{\theta}_{(k)\ell}}{\widehat{\theta}_{(k')\ell}}\right)^{Y_{i\ell}} \left(\frac{1-\widehat{\theta}_{(k)\ell}}{1-\widehat{\theta}_{(k')\ell}}\right)^{1-Y_{i\ell}}, \tag{S16}$$

where $\widehat{\theta}_{(k)\ell}$ is an estimated true positive rate in cluster k at feature ℓ and 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 interfere with 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 (S16).

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 S1 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.

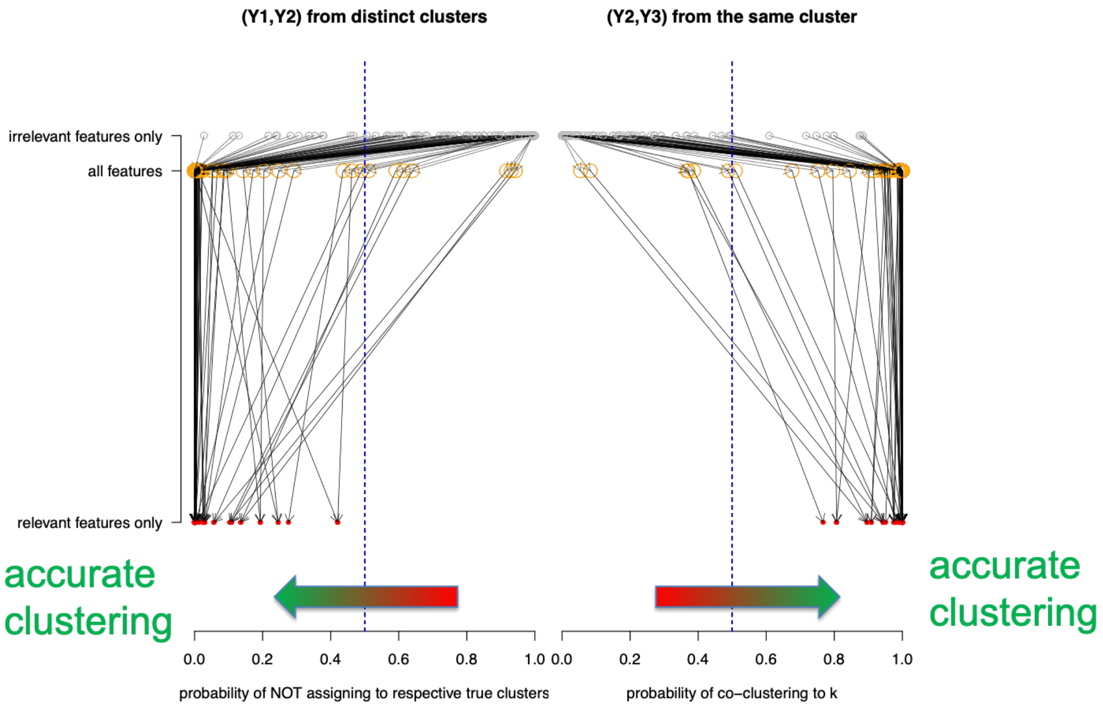

Figure S1: Removing irrelevant features improves estimation of clusters. *Left*) 100 random pairs of observations drawn from distinct clusters; the probability of them not being clustered correctly is lowered (pulled towards zero) once the irrelevant features are removed. *Right*) 100 random pairs of 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.

A5 Additional Analysis Results

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.

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).

A6 Additional Figures and Tables

There are three major aspects of a specification of RLCM: a) whether the latent states ($\boldsymbol{\eta}_i$) that define the clusters take values from a known or unknown subset $\mathcal{A} \subseteq \{0,1\}^L$; b) whether it is known, partially known, or unknown about how a binary design matrix Γ depends on $\{\boldsymbol{\eta}_i\}$ along with other parameters, where Γ specifies which and how measurements exhibit between-class differential response probabilities, and c) the parametric form of the conditional distribution of measurements given latent states and response probabilities (Λ): $[\boldsymbol{Y}_i \mid \boldsymbol{\eta}_i, \Lambda]$, where $\Lambda = \{\lambda_{i\ell}\}$ is parameterized and must satisfy restrictions imposed by the design matrix Γ. Table S1 in Supplementary Materials summarizes the above and some other variants of LCMs by specifications of the latent state space ($\mathcal{A}$), design matrix (Γ), and measurement likelihood.

Model Specification			Methods (examples)			
			Restricted LCM		LCM	
			Bayesian	non-Bayesian	Classical	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)	-	-
	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)
measurement process ($[\mathbf{Y}_i \mid \boldsymbol{\eta}_i, \boldsymbol{\Lambda}]$) **$\boldsymbol{\Lambda}$ must respect** Γ	(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 S1: Comparison of variants of latent class analysis of multivariate binary data.

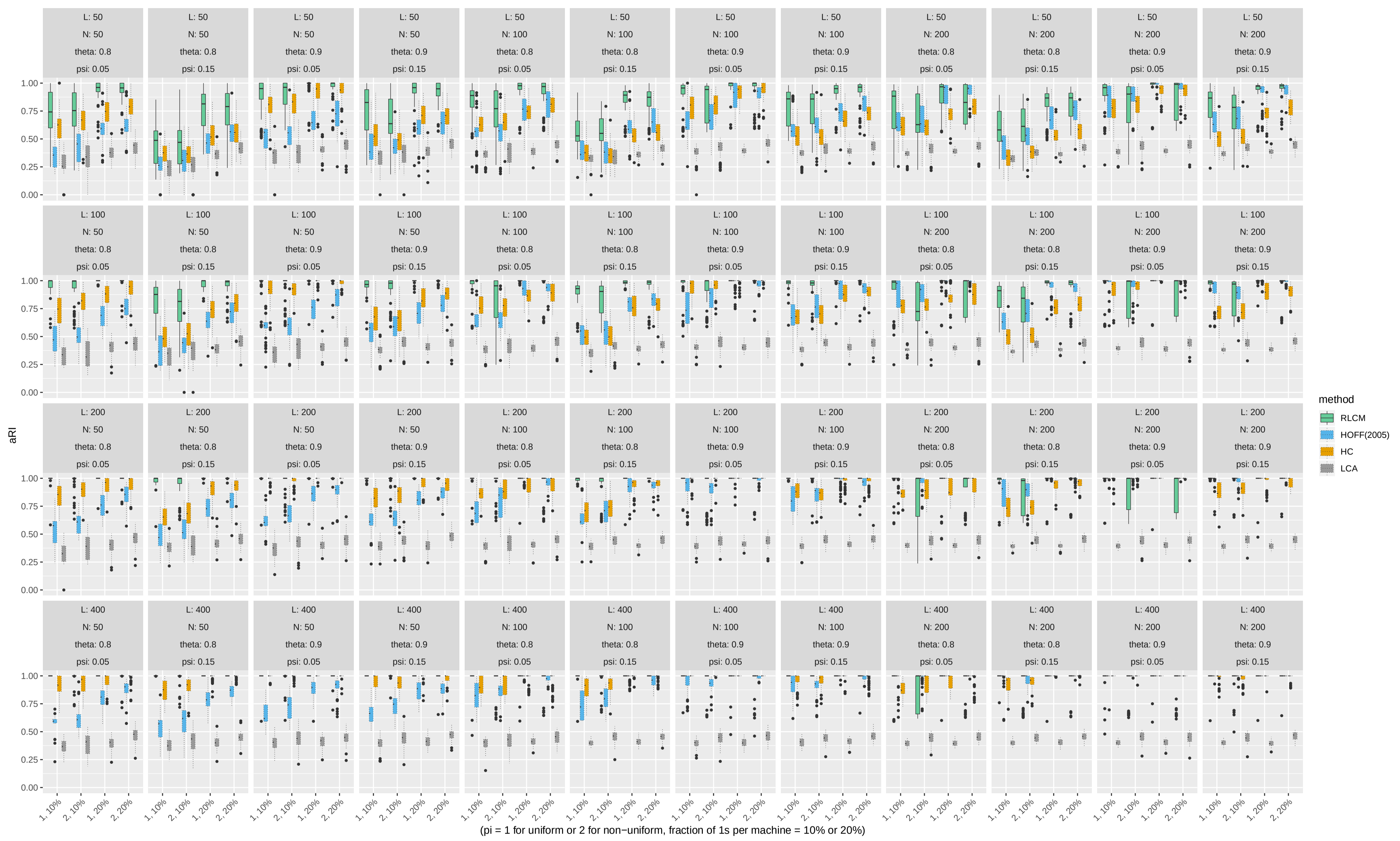

Figure S2: Based on $R = 60$ replications for each parameter setting, Bayesian RLCM (boxplots with solid lines) most accurately recovers the true clusters compared to subset clustering (Hoff, 2005) hierarchical clustering (HC) and traditional Bayesian latent class analysis (LCA) (from the left to the right in each group of four boxplots). This figure expands Figure 2 in Main Paper over more parameter settings.

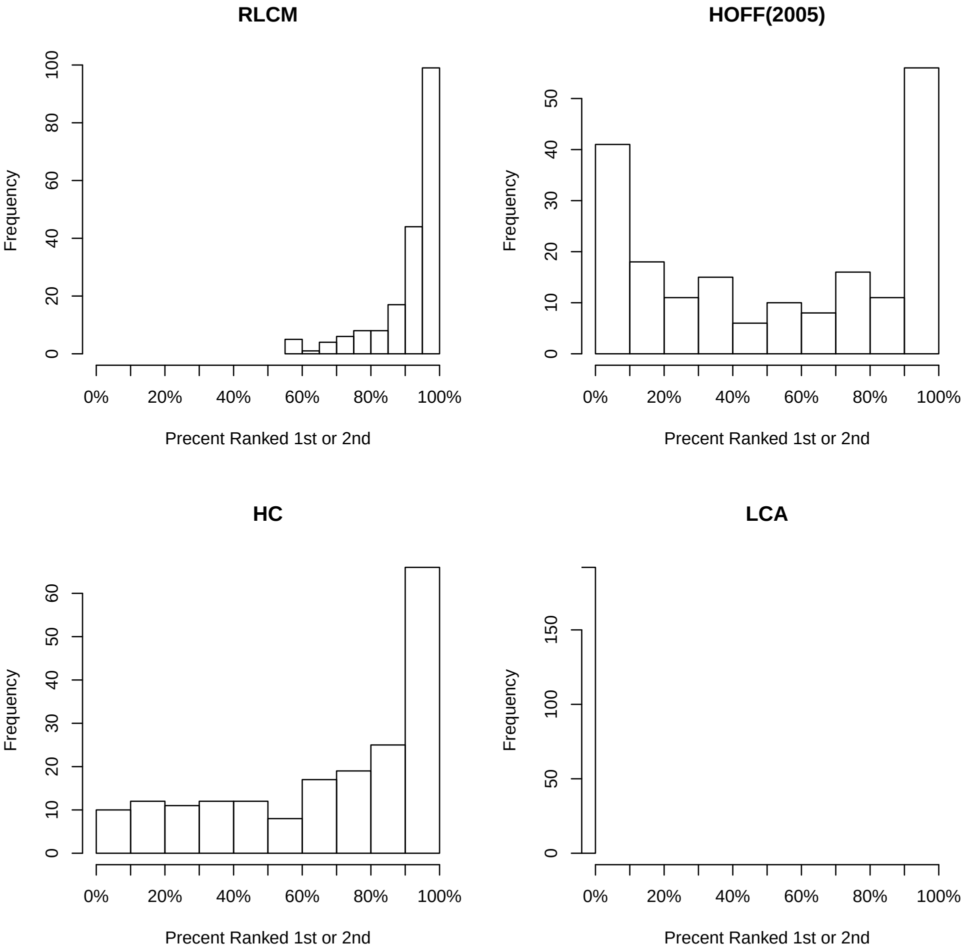

Figure S3: For each of four clustering methods (Bayesian RLCM, Hoff (2005), HC, Bayesian LCA), the percent being ranked the first or the second in terms of the mean aRI averaged across $R = 60$ replications (Section 4.1 in Main Paper). Each histogram is produced for the $1,920$ combinations of parameters investigated in Section 4.1 in Main Paper.

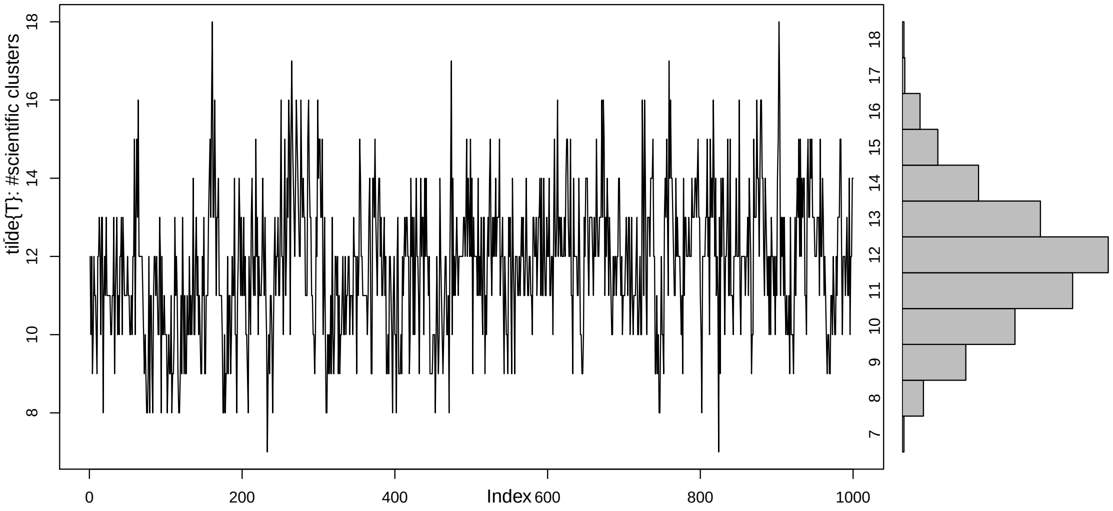

Figure S4: MCMC samples of the number of scientific clusters ($\widetilde{C}$) with its marginal posterior on the right margin.

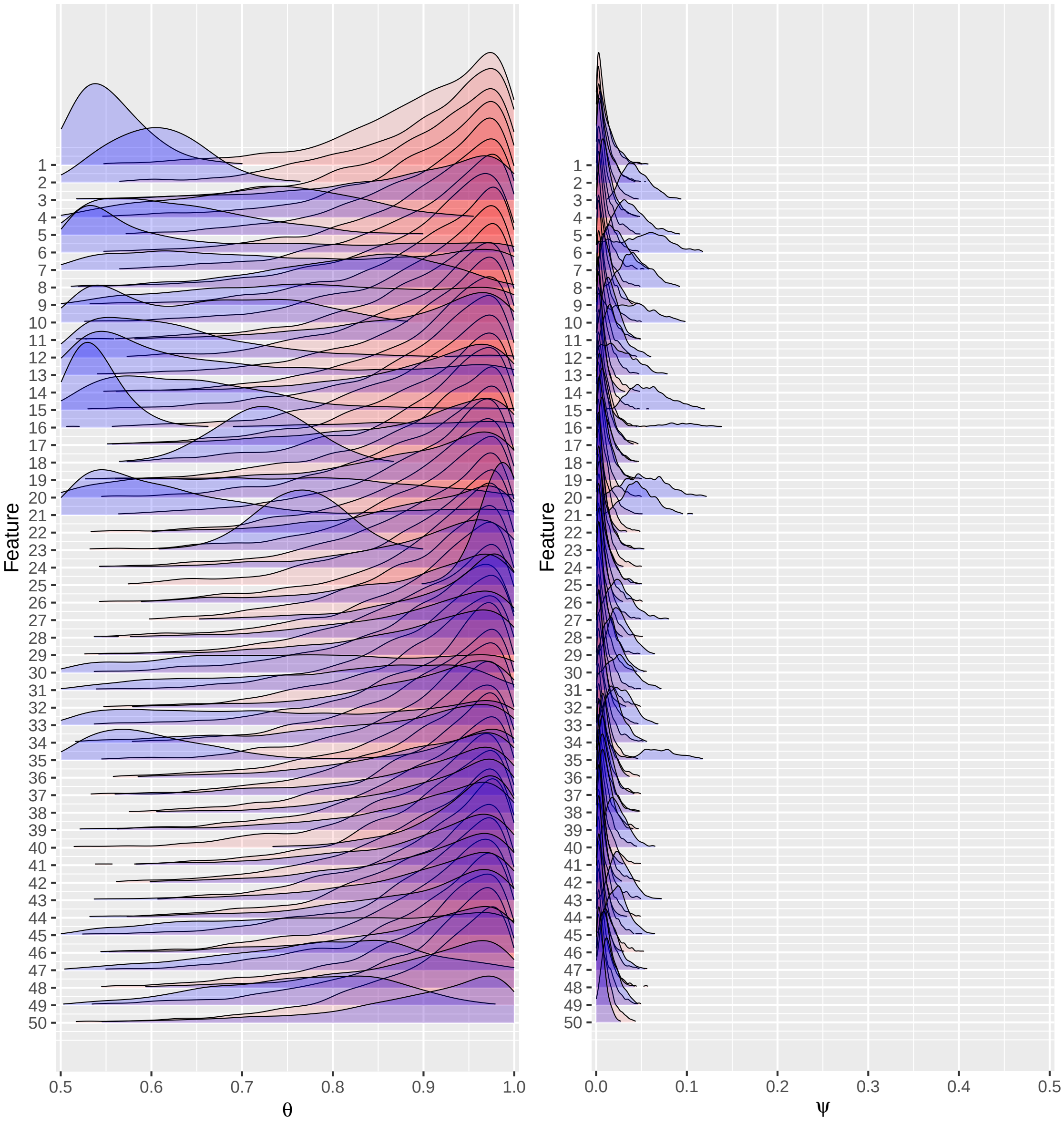

Figure S5: Prior vs posterior for all true positive rates $\{\theta_\ell\}$ (left) and false positive rates $\{\psi_\ell\}$ (right).

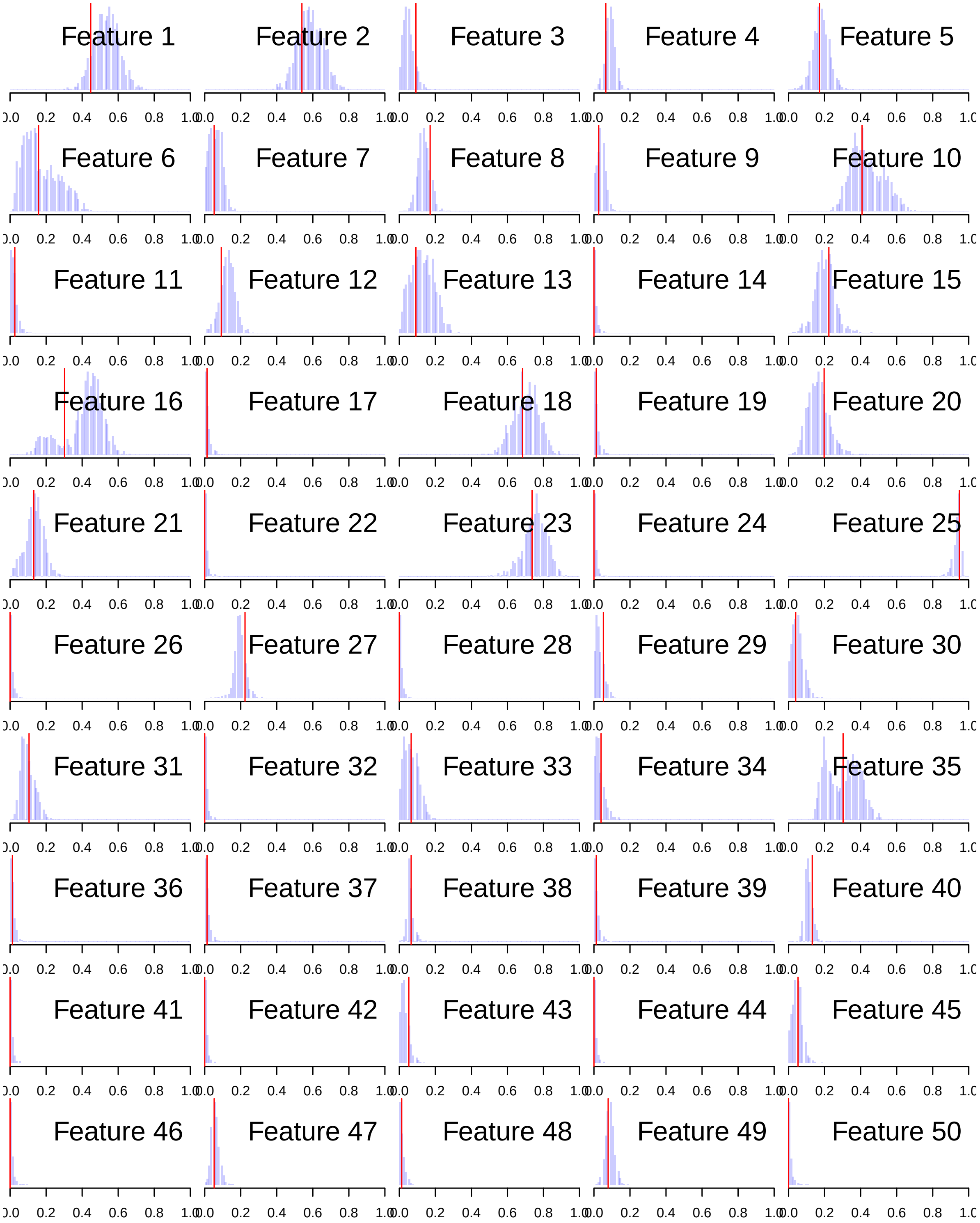

Figure S6: Observed marginal positive rate (solid vertical line) plotted against the posterior predictive distributions for $L = 50$ landmarks in Example 1.

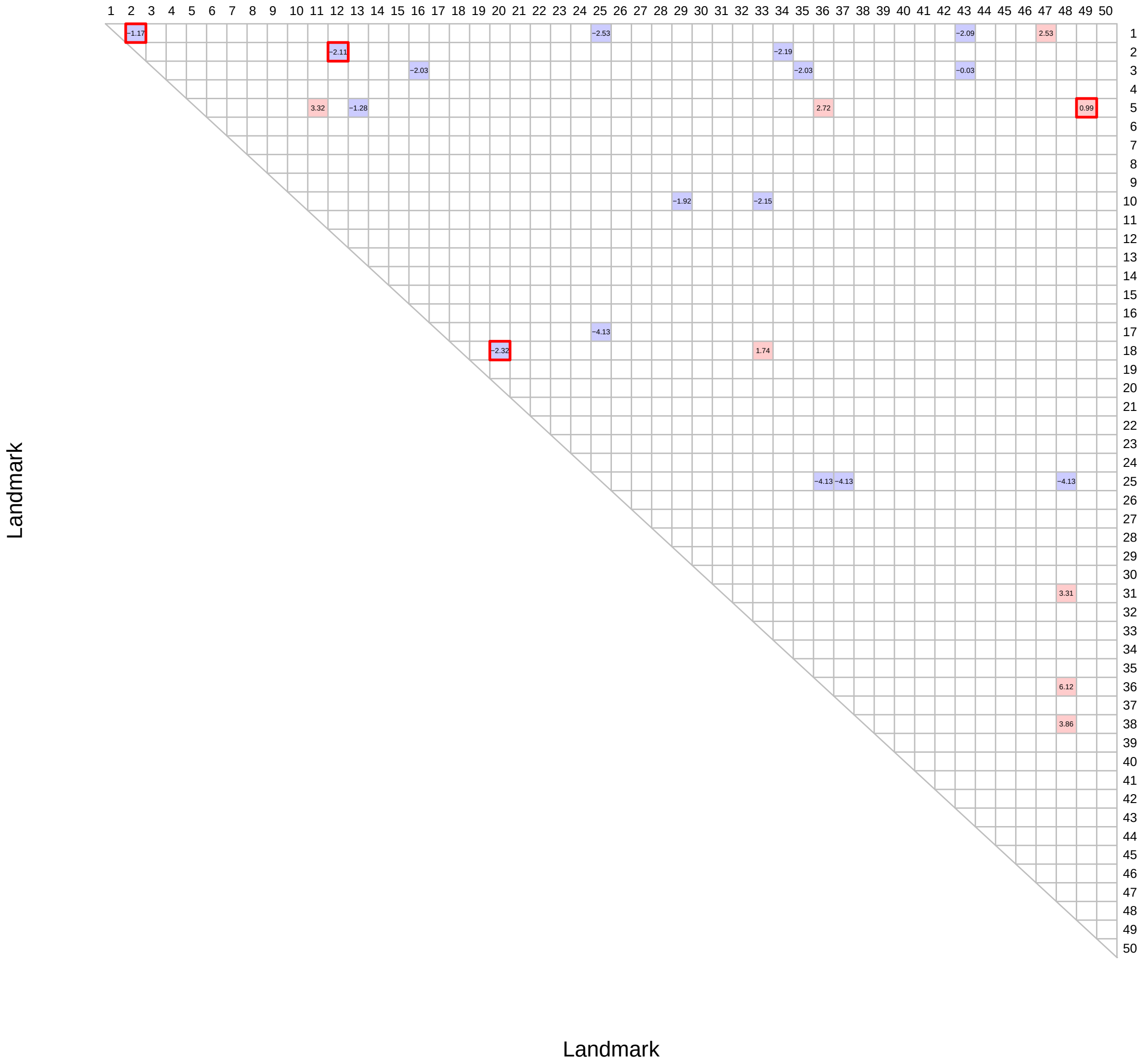

Figure S7: Significant deviations of model predicted log odds ratios (LOR) from the observed LOR. A blank cell indicates a good model prediction for the observed pairwise LOR ($|\mathsf{SLORD}| < 2$); A red (blue) cell indicates model under- (over-) fitting $\mathsf{SLORD} > 2(< -2)$, where standardized 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.

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.

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.

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.

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.

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.

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.

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.

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.

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.