Clustering Multivariate Binary Outcomes with Restricted Latent Class Models: A Bayesian Approach

Zhenke Wu
Assistant Professor of Biostatistics
Schools of Public Health,
University of Michigan, Ann Arbor

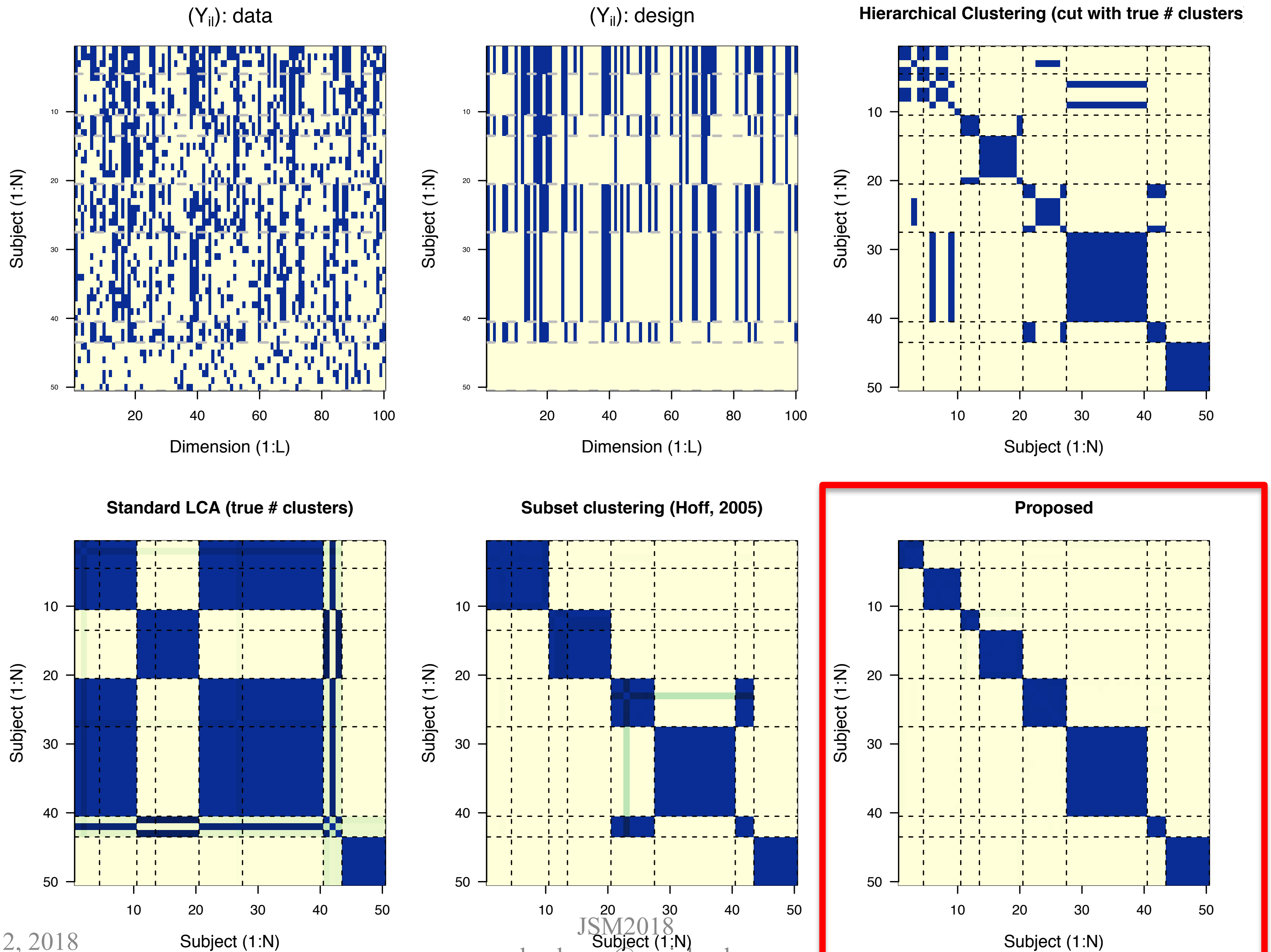
Joint Statistical Meetings 2018
Vancouver
August 2, 2018

(zhenkewu@umich.edu)

zhenkewu.com

R Package: rewind
https://github.com/zhenkewu/rewind

Motivating Example



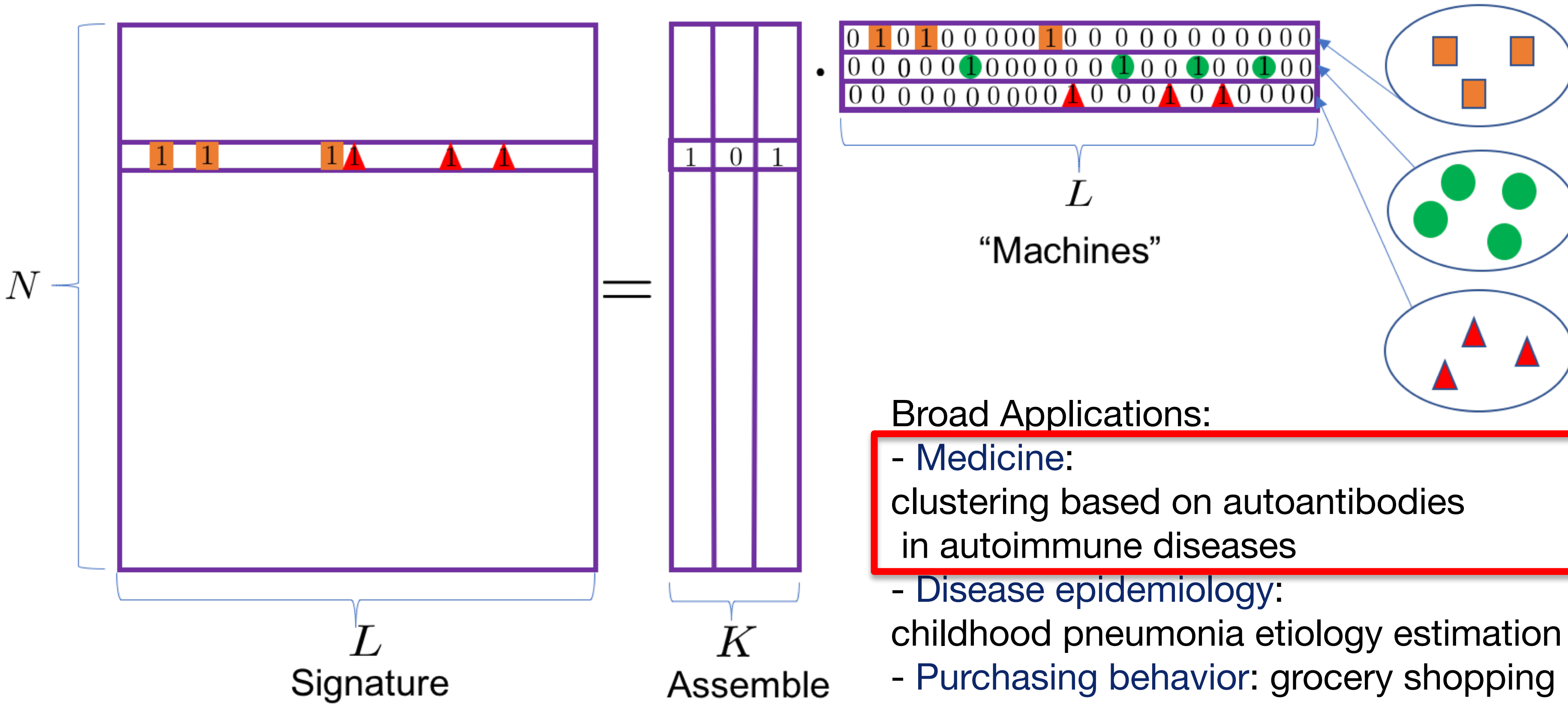
Take-away

Accurate clustering of multivariate binary data that

1) automatically selects feature subsets and
2) works well for **unbalanced** cluster sizes

We achieve this goal via boolean matrix decomposition, or more generally, restricted latent class models

Boolean Matrix Decomposition (noise-free version) (a special case of restricted latent class models)



Broad Applications:

- Medicine:
clustering based on autoantibodies in autoimmune diseases
- Disease epidemiology:
childhood pneumonia etiology estimation
- Purchasing behavior: grocery shopping
- Computer Science: text mining
- Educational assessment:
cognitive classifications
- Mobile health:
latent constructs, e.g., engagement with interventions, vulnerability and receptivity

Statistical Formulation

Model Setup: Quick Overview

- *Data*: $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{iL})^T \in \{0,1\}^L, i = 1, \ldots, N$
- *Latent state vector*: $\boldsymbol{\eta}_i \in A \subset \{0,1\}^M$
- *Latent dimension: M*
- *Latent class: K distinct patterns of* $\boldsymbol{\eta}_i$
- *The number of clusters, K, unknown (no greater than* 2^M*)*
- *Q-matrix (M by L; binary): Q*

Model Setup: Quick Overview

1) Given a latent state dimension M, specify likelihood $[\boldsymbol{Y}_i \mid \boldsymbol{\eta}_i, \boldsymbol{\Lambda}]$ via restricted latent class models (RLCM) ; with conditional independence

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

For example, for dimension *l:*

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

-Needs just one required state in ({$m : Q_{ml} = 1$}) for a positive ideal response $\Gamma_{il} = 1$.

- referred to as partially latent class model in epidemiology (Wu *et al*., 2016); Deterministic In and Noise Or gate (DINO) in psychology (Junker and Sijtsma, 2001); non-negative matrix factorization if rows of Q are orthogonal (Lee and Seung, 1999)

Model Setup: Quick Overview

In two steps,

1) Given a latent state dimension M, first specify the likelihood $[\boldsymbol{Y}_i \mid \boldsymbol{\eta}_i, \boldsymbol{\Lambda}]$ via restricted latent class models (RLCM) ; with conditional independence

2) A prior distribution $[\boldsymbol{\eta}_i, i = 1, \ldots, N]$ obtained from a clustering mechanism with unknown # of clusters *K* (represented by cluster assignment indicators $\{Z_i, i = 1, \ldots, N\}$);

We use mixture of finite mixtures (Miller and Harrison, 2017 JASA)

Challenges: Boolean Matrix Decomposition (an example of restricted latent class models)

- C1. High-dimensional discrete space
  Sparse priors that encourage:
  1. small # of latent state dimensions
  2. small # of distinct latent state patterns

- C2. Unknown number of latent state dimensions
  Infinite dimension model (based on semi-ordered formulation of Indian Buffet Process); Identifiability issue

- C3. Unknown number of clusters (i.e., # latent classes)
  Mixture of finite mixture model

- T1: Identifiability of model parameters based on likelihood only
  Open and frontier problem; exciting progress at Michigan

C: computational
T: theoretical

Comparison of variants of latent class analysis of multivariate binary data

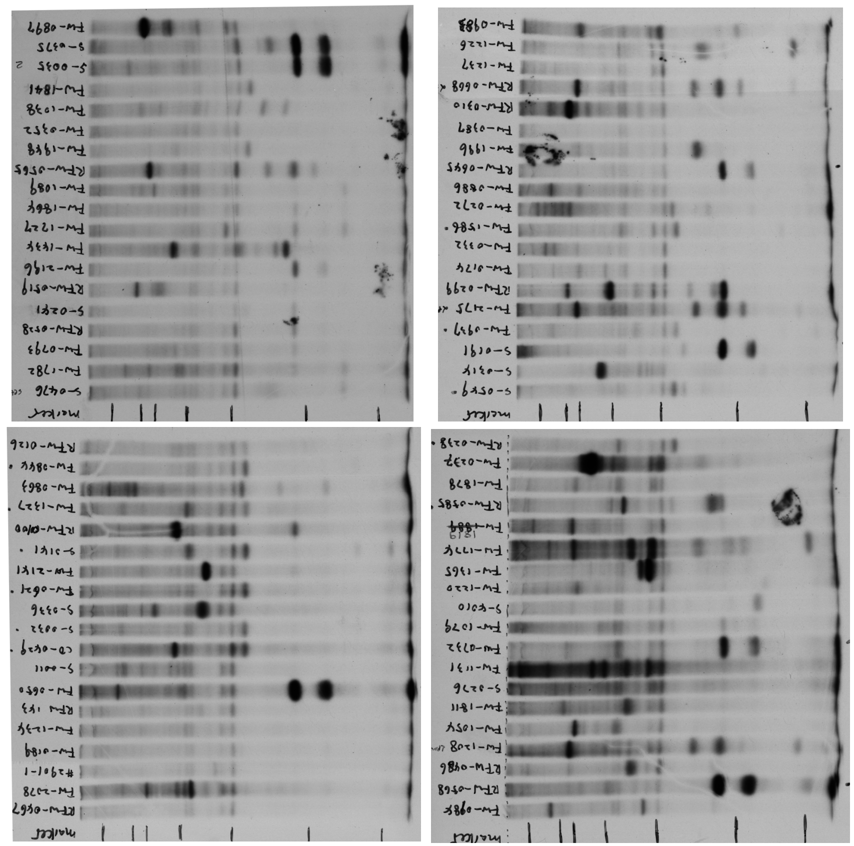
Model Specification			Methods (examples): Restricted LCM: Bayesian	Methods (examples): Restricted LCM: non-Bayesian	Methods (examples): Classical LCM	Methods (examples): Nested Partially LCM †
latent state variables ($\boldsymbol{\eta}_i \in \mathcal{A} \subset \{0,1\}^M$; **#latent classes:** $\widetilde{K} = \mid \mathcal{A} \mid$)	$\widetilde{K}$ known	$\mathcal{A}$ known	$\mathcal{A} = \{0,1\}^M$: Chen et al. (2017)	$\mathcal{A} = \{0,1\}^M$: Xu (2017); $\mathbf{0}_M \in \mathcal{A} \neq \{0,1\}^M$: Leighton et al. (2004), Gu and Xu (2018)	Green (1951)*, Anderson (1954)*, Lazarsfeld and Henry (1968)*, Goodman (1974)*, Erosheva et al. (2007)†,‡, Bhattacharya and Dunson (2012)†,‡	$\mathbf{0}_M \in \mathcal{A}$ and partially observed some of $\{i : \boldsymbol{\eta}_i = \mathbf{0}_M\}$: Wu et al. (2017b)
		$\mathcal{A}$ unknown	-	Miettinen et al. (2008)#	-	-
	$\widetilde{K}$ unknown	$\mathcal{A}$ unknown (M known or 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 (2017), Chen et al. (2015)	-	-
measurement process ($[\mathbf{Y}_i \mid \boldsymbol{\eta}_i, \Gamma, \boldsymbol{\Lambda}]$)	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. *: early applications. #: non-probabilistic.
✓: applies to all in the column (except for other rows in the same row block)

Table 1: Comparison of variants of latent class analysis of multivariate binary data.

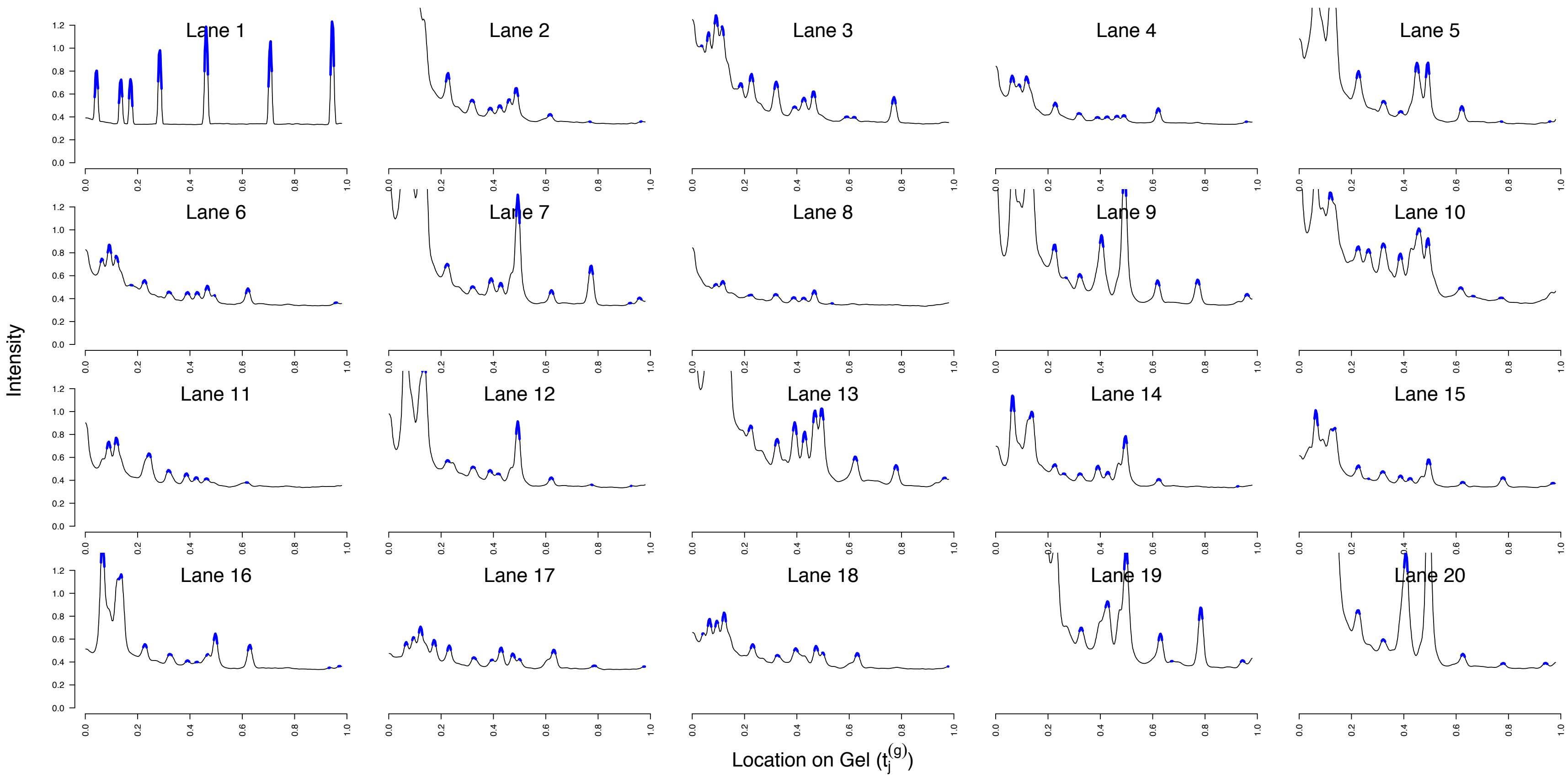
Data

- 76 autoantibody patterns from patients with rheumatic disease & cancer
- all were negative for autoantibodies against prominent defined specificities



Can an algorithm be developed to identify common autoantibody signatures? And estimate clusters among patients?

Raw Intensity Scan Data
(20 lanes on a single gel)



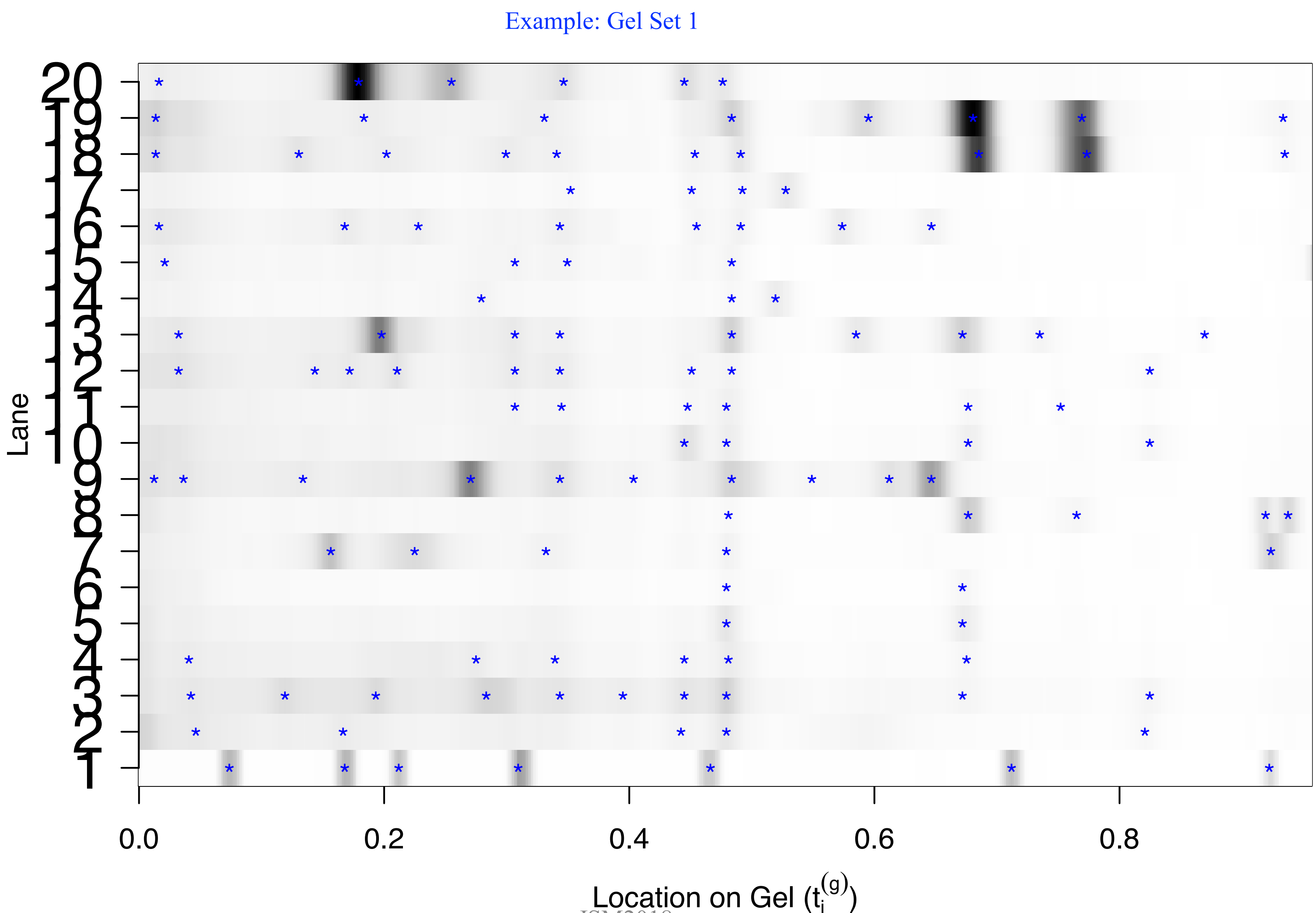
Scientific Questions

- How many clusters? What are the clusters?
  [the clustering problem]

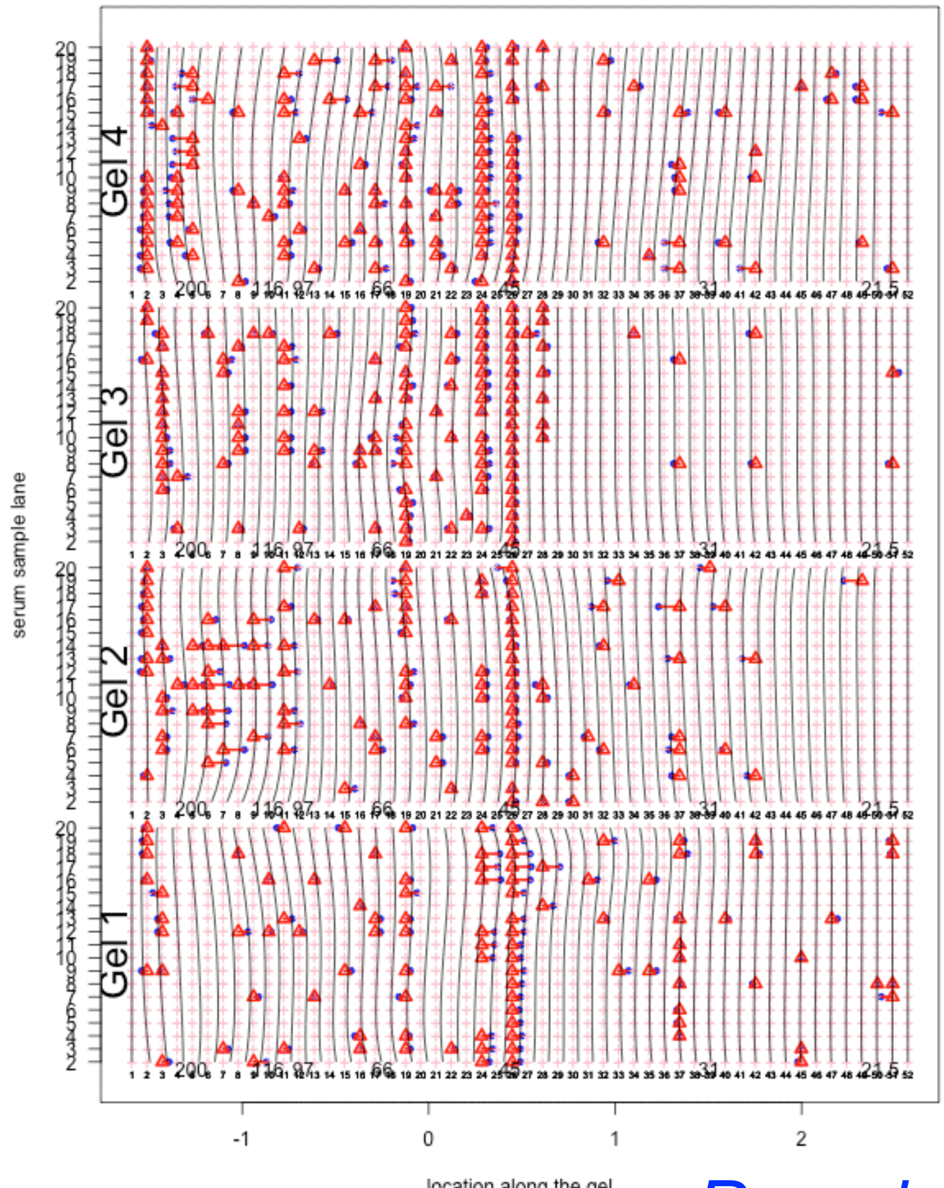
- How many machines are there and what are the component auto-antigens?
  [estimation of latent state dimensions]

- What makes the clusters different in terms of presence or absence of machines?
  [interpretability of the clusters]

Preprocessing Step I-a: Automated Peak Detection



Align the peaks (Wu et al., 2017)



R package: spotgear

Posterior Computation

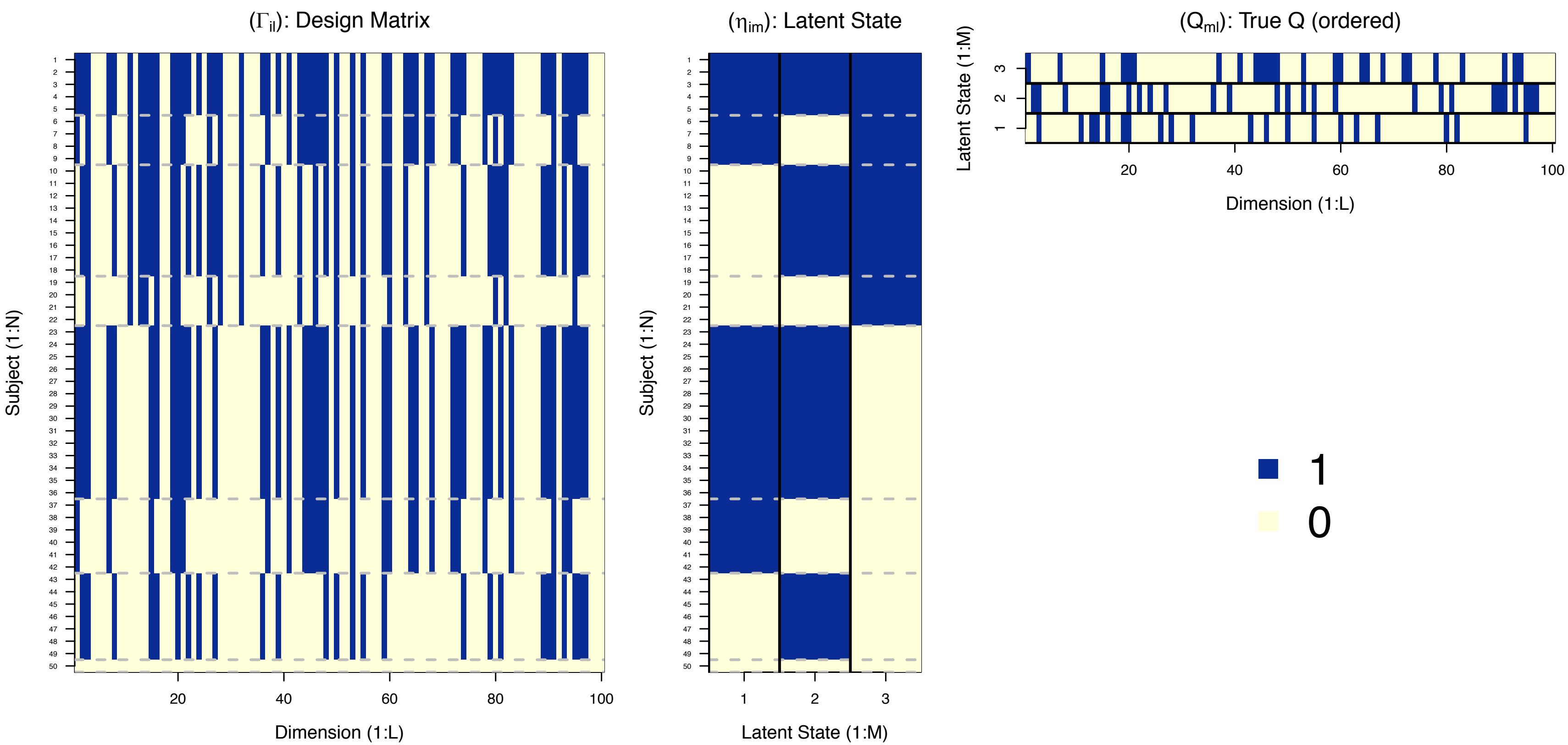
Posterior Computation

- Designed and implemented MCMC algorithms that deal with
  - a) unknown number of clusters (mixture of finite mixture models; split-merge), and
  - b) unknown number of machines (slice sampler for infinite Indian Buffet Process). Also works for pre-specified number of machines.

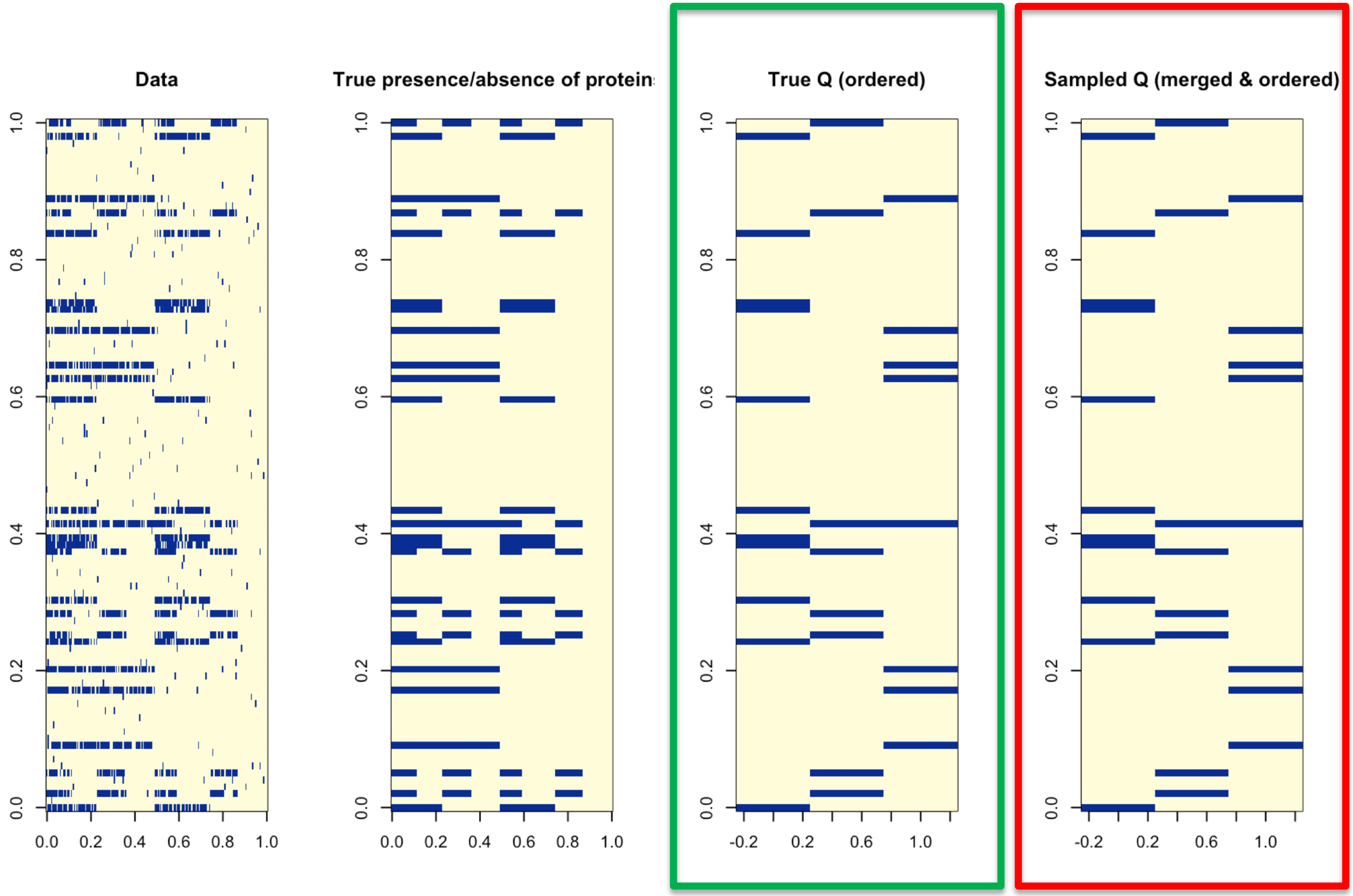
R package: rewind

Simulation

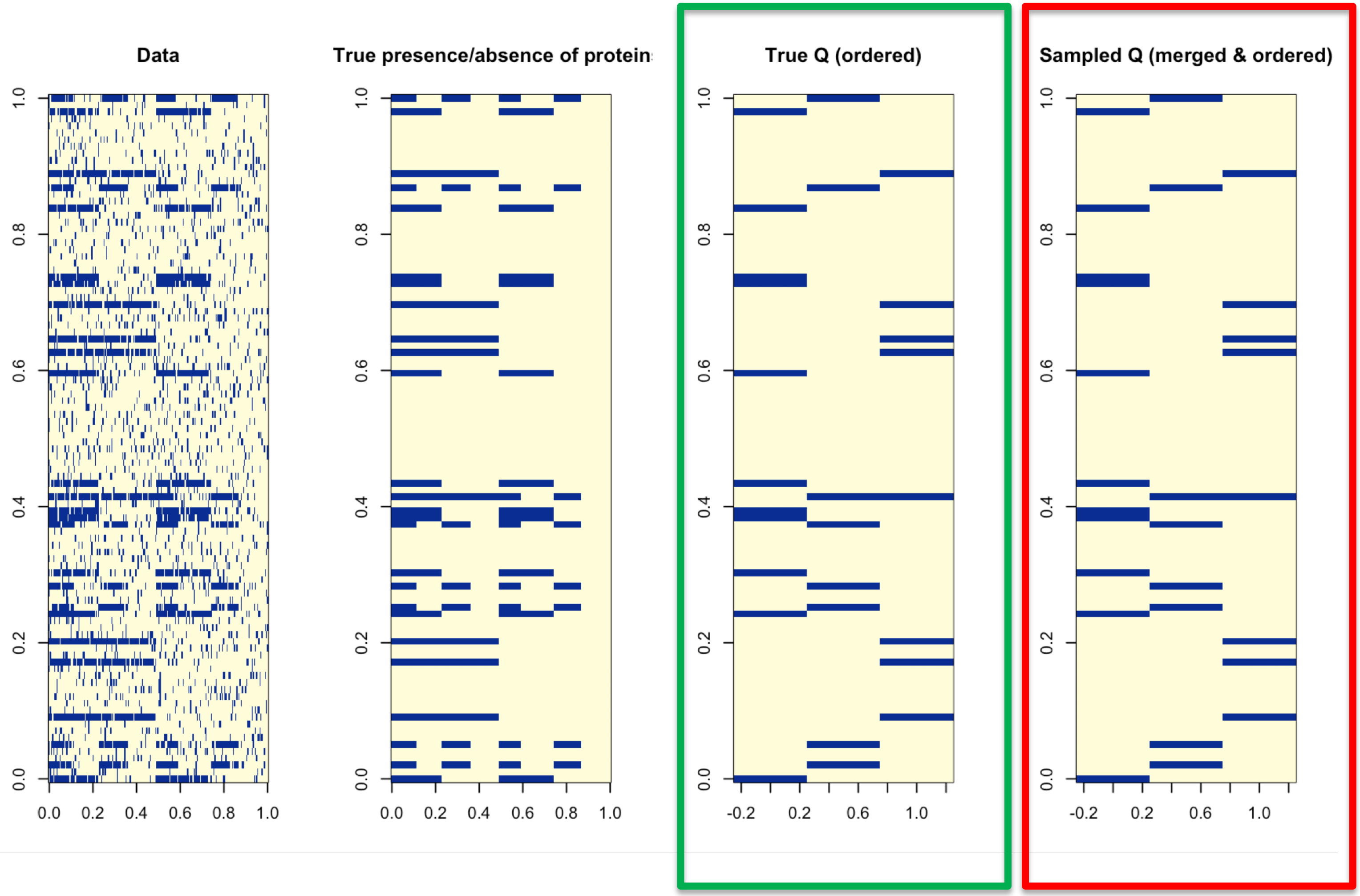
Simulation Setup



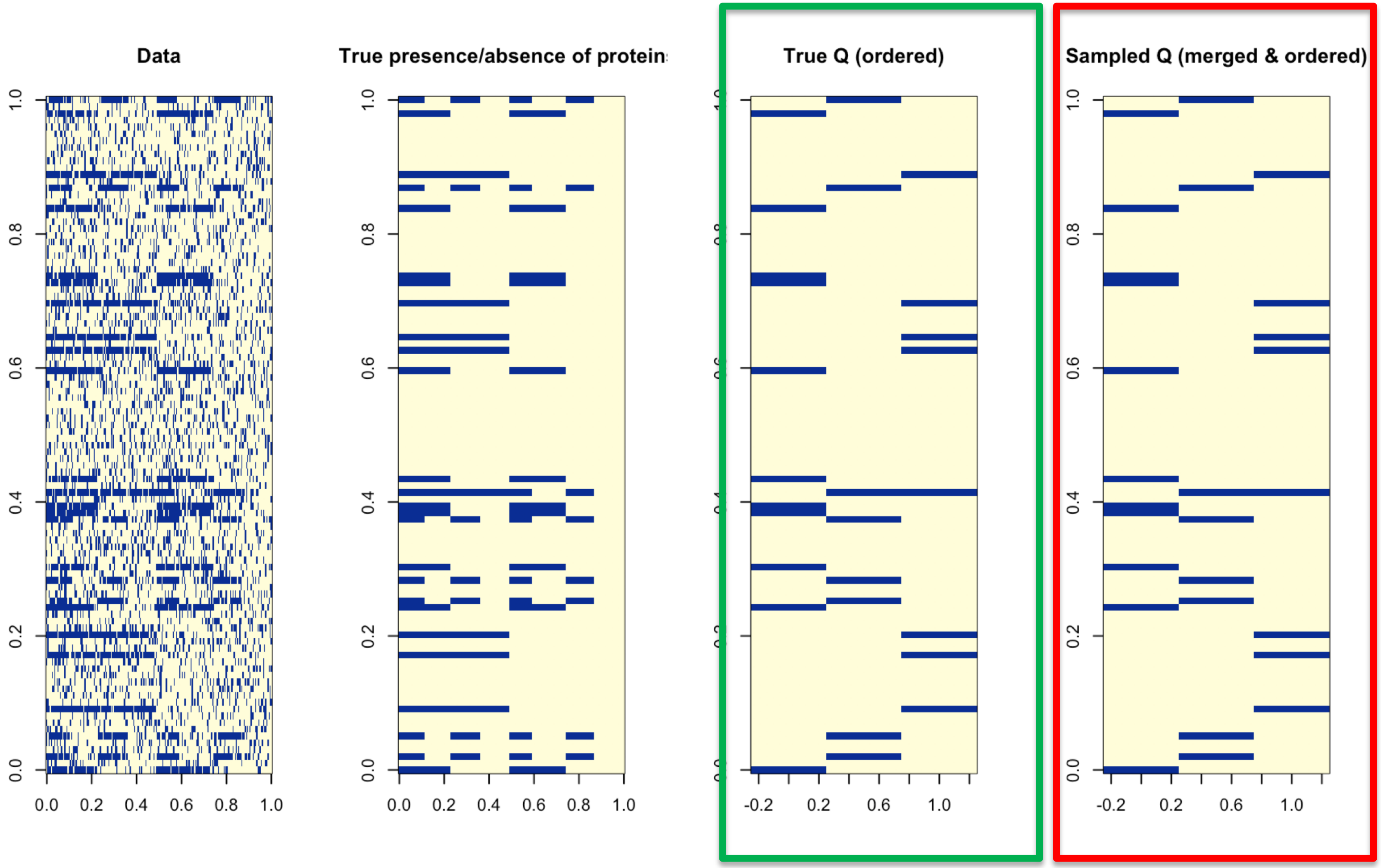
Recovery of the matrix Q (low noise)



Recovery of the matrix Q (intermediate noise)



Recovery of the matrix Q (high noise)



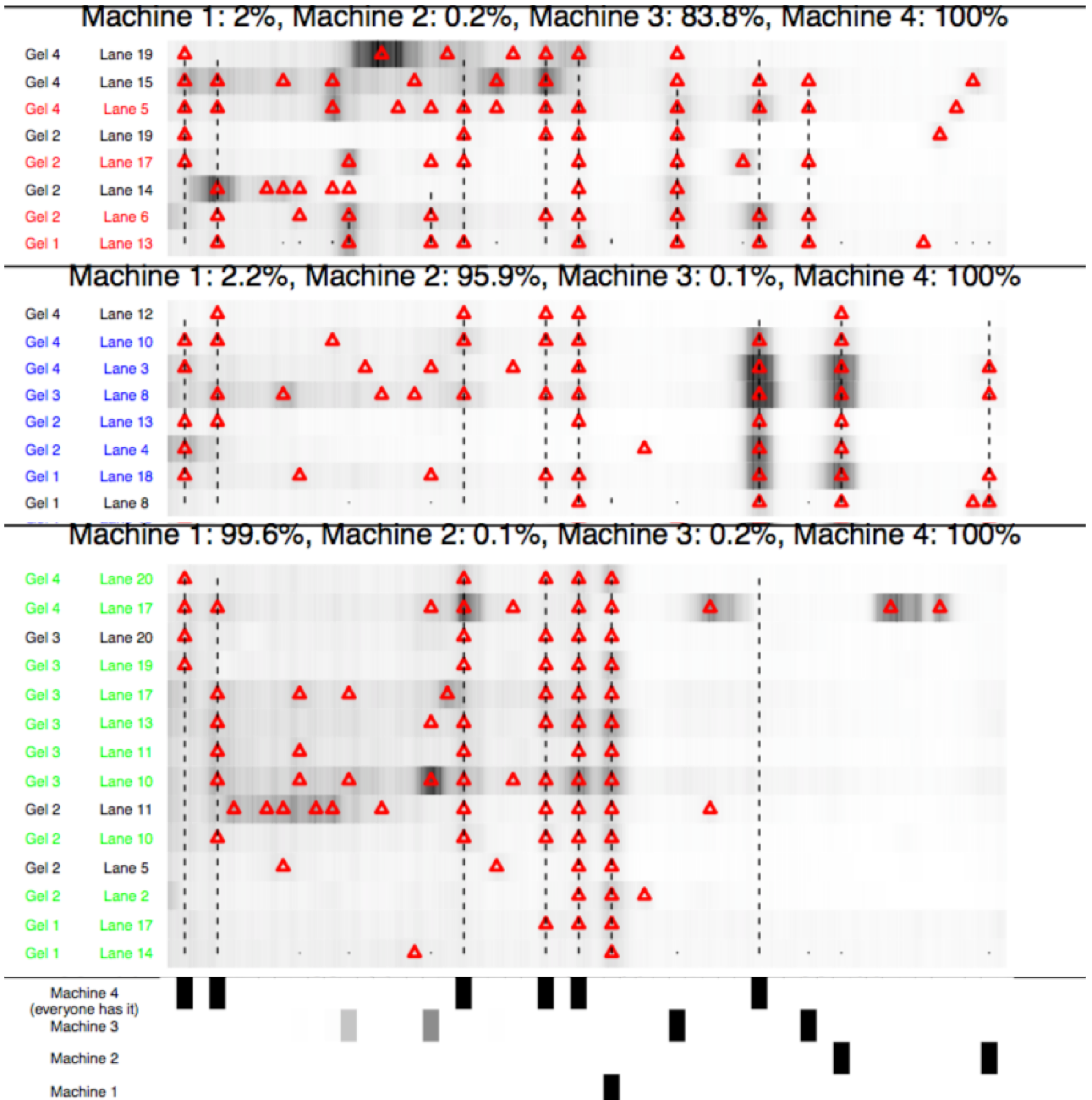
Preliminary clustering results based on machine models

Data: CTP negative sera

Method: Bayesian machine-based restricted latent class analysis

Figure: Three estimated clusters (top three panels) with distinct enrichment of three distinct estimated machines (bottom panel)

Colored labels: red, blue, green - for clusters obtained by standard method; this algorithm is agnostic to them.

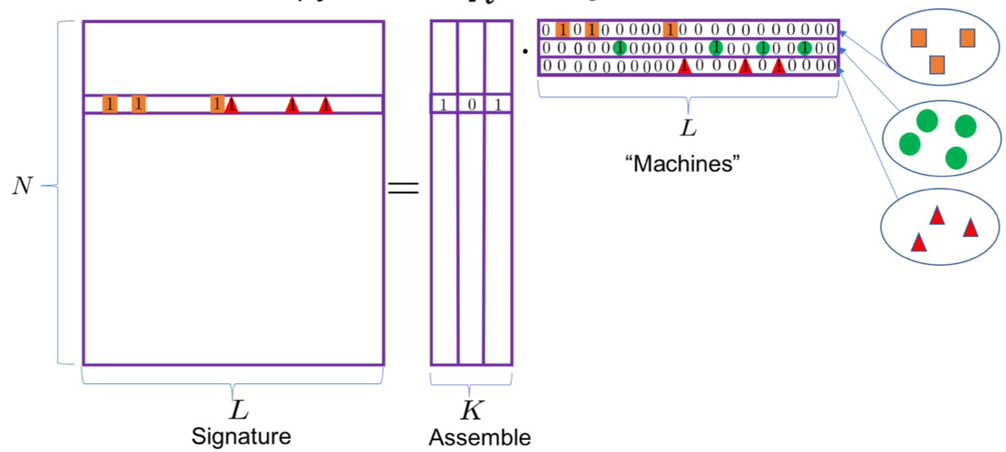


Main Points Once Again

- Goal: Based on multivariate binary data, find scientifically structured, interpretable clusters

- Proposed a framework for clustering using restricted latent class models

$$\xi_i = \eta_i^\top \cdot Q$$



- Designed and implemented MCMC algorithms that deal with unknown number of clusters and machines; Bayesian binary factorization algorithm

- Superior clustering performance compared to standard analyses; Improved estimations under **unbalanced** cluster sizes.

References

Wu Z, Casciola-Rosen L, Shah A, Rosen A, Zeger SL. Estimating autoantibody signatures to detect autoimmune disease patient subsets. Submitted for publication. *Biostatistics*. In Press. doi: 10.1093/biostatistics/kxx061

Wu Z, Zeger SL. Clustering Multivariate Binary Outcomes with Restricted Latent Class Models: A Bayesian Approach. Working paper.

Wu Z, Deloria-Knoll M, Hammitt LL, Zeger SL for the PERCH Core Team. Partially-latent class models (pLCM) for case-control studies of childhood pneumonia etiology. *Journal of the Royal Statistical Society, Series C*. 65: 97-114, 2016.

Wu Z, Deloria-Knoll M, Zeger SL. Nested, partially-latent class models for dependent binary data with application to estimating disease etiology. *Biostatistics* 18 (2), 200-213. 2016

Open Source Software

- ***spotgear:*** Subset Profiling and Organizing Tools for Gel Electrophoresis Autoradiography in R

- ***rewind:*** Reconstructing Etiology with Binary Decomposition

Available from *https://github.com/zhenkewu*

Thank you