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

Zhenke Wu | zhenkewu.com | Department of Biostatistics and Michigan Institute for Data Science

Introduction

- **Y:** $N \times L$ binary data matrix of N observations with L dimensions or "features".
- Multivariate binary data frequently arise as noisy measurements of presence or absence of a list of unobservable or latent binary variables $\boldsymbol{\eta}$ called "***states***".
- Seek to estimate clusters of observations from **Y**
- [**Want to incorporate *Scientific Structures!***] It is hypothesized that there are underlying subgroups defined by distinct patterns of $\boldsymbol{\eta}$ vectors that take values on a relatively small number of elements in $\{0, 1\}^M$, $M \leq L$.
- That is, members of the same cluster (class) share values of a relatively smaller number of states.

Traditional Latent Class Models (LCM)

- latent state vectors $\{\boldsymbol{\eta}_i\}$ take value from $\mathcal{A}$, where $K = |\mathcal{A}|$ is the number of distinct patterns $\mathcal{A}=\{\boldsymbol{\eta}_k^*, k = 1, \ldots, K\}$.
- Latent classes have distinct $\boldsymbol{\eta}$ patterns.
- Given $\boldsymbol{\eta}_i$, response probability of feature ℓ for subject i is $P(Y_{i\ell} = 1 | \Lambda) = \lambda_{i\ell}$, $\ell = 1,\ldots,L$, where $\Lambda=\{\lambda_{i\ell}\}$ is a $N \times L$ matrix.
- $\lambda_{i\ell} = \lambda_\ell(\boldsymbol{\eta}_i)$ where $\lambda_\ell : \mathcal{A} \to [0, 1]$ computes the response probability of feature ℓ given $\boldsymbol{\eta}_i$. $\lambda_{i\ell}$ shifts between K distinct values, or *between-class differential errors*.
- *Conditional independence* assumption: the measurements from distinct dimensions are independent of one another given the latent class and response probabilities in that class, i.e. $Y_{i\ell} \perp Y_{i\ell'} \mid \boldsymbol{\eta}_i, \lambda_\ell(\cdot)$.

Simulation II: Superior Clustering Performance

Figure 3. The proposed Bayesian RLCM most accurately recovers the true clusters compared to hierarchical clustering (HC), traditional Bayesian latent class analysis (LCA) and subset clustering (Hoff, 2005). Under each true parameter setting, the adjusted Rand Index (aRI) is computed for the estimated clusters obtained by each method under $R = 60$ replications.

- adjusted Rand index (aRI): assess the agreement between two clusterings (e.g., estimated versus truth)
- $-1 \leq \text{aRI} \leq 1$ and corrects for chance agreement. It equals one for identical clusterings and is on average zero for two random partitions; larger values indicate better agreements between the two clusterings.
- Simulations cover distinct levels of measurement errors, feature dimensions, sparsity levels of Q, sample sizes and population fractions of latent state patterns.
- Overall winner: ***Bayesian RLCM***; Accurate clustering through feature selection and scientific structures.

Application to Gel Electrophoresis and Autorad Data

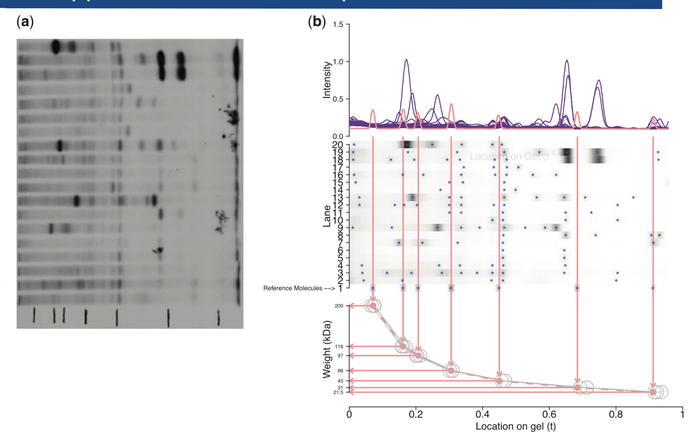

Figure 4. Gel electrophoresis autoradiography (GEA) data for 20 samples on one gel. (**a**) Raw GEA image. (**b**) *Top* : Radioactive intensities for all the samples; *Middle* : Heatmap of the radioactive intensities for all the samples. The asterisks (*) denote the detected peaks. Seven vertical lines indicate the locations of the seven reference molecules observed on lane 1. *Bottom* : Actual molecular weights (Y-axis) as read from the location along the gel (X-axis). Four location- to-weight curves are shown here, each corresponding to reference lane 1s in the four gels analyzed here (the dashed curve "- - -" is for the gel shown in the middle). Note the reference molecule misalignment shown by the scattered "◯".

Take-away Messages

- Proposed a method for estimating scientifically-structured clusters (SSC) from multivariate binary data.
- Most useful for **large** dimension L (needs feature selection) with **unknown** number of clusters (needs posterior distribution over # of clusters).
- Advantages of SSC
  - If the underlying clusters differ from one another only at subsets of features, SSC can more accurately estimate these clusters than standard clustering methods such as latent class analysis and hierarchical clustering.
  - SSC also produces more interpretable clusters (Fig. 5).

Incorporate Scientific Structured Classes ("restricted")

- Restricted LCMs (RLCM) assume equality of a subset of response probabilities across classes.
- The set of RLCM parameters comprises a *Lebesgue measure zero set* in the parameter space of the LCM.
- The restrictions are specified by introducing a binary design matrix $\Gamma = \{\Gamma_{\eta,\ell}\} \in \{0,1\}^{K \times L}$ with latent classes and dimensions in the rows and columns, respectively.
- Let $\mathcal{A}_\ell = \{\boldsymbol{\eta} \in \mathcal{A} : \Gamma_{\eta,\ell} = 1\}$ where $\mathcal{A}_\ell$ collects latent classes with the highest response probability for dimension ℓ. If $\mathcal{A}_\ell \neq \emptyset$, restrict the response probabilities at feature ℓ by $\max_{\boldsymbol{\eta} \in \mathcal{A}_\ell} \lambda_{\eta,\ell} = \min_{\boldsymbol{\eta} \in \mathcal{A}_\ell} \lambda_{\eta,\ell} > \lambda_{\eta',\ell}$, $\ell = 1,\ldots,L$, $\boldsymbol{\eta}' \notin \mathcal{A}_\ell$
- Further, there can exist a class $\boldsymbol{\eta} \in \mathcal{A}$ that gives rise to all-zero ideal responses $\Gamma_{\eta\star} = \mathbf{0}_{1\times L}$.

A Noise-Free Example ("Boolean Matrix Factorization")

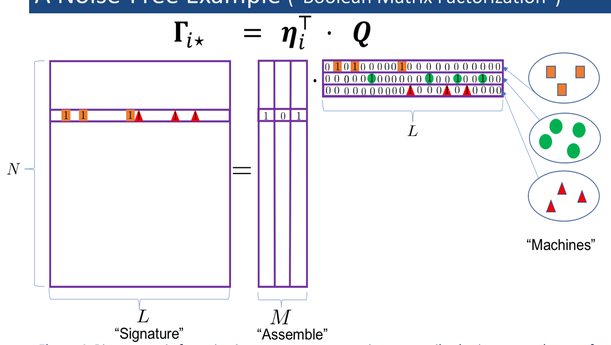

Figure 1. Binary matrix factorization generates composite autoantibody signatures that are further subject to misclassification errors. The signature $\boldsymbol{\Gamma}_{i\star} = \boldsymbol{\eta}_i^T \boldsymbol{Q}$ assembles three *orthogonal* machines with 3, 4 and 3 landmark proteins, respectively. The highlighted individual, if without error, will mount immune responses against antigens in Machines 1 and 3.

Simulation I: SSC and Feature Selection

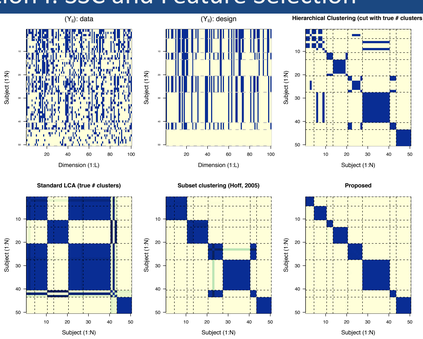

Figure 2. In the 100-dimension multivariate binary data example, the eight classes differ with respect to subsets of measured features. Bayesian restricted latent class analysis accounts for measurement errors, selects the relevant subsets and filters the subsets by a low-dimensional model (Fig.1) and therefore yields superior clustering results.

Open-Source Software (R Packages)

1) ***rewind***: **R**econstructing **E**tiology **w**ith B**in**ary **D**ecomposition
   - Fit Bayesian restricted latent class models based on mixture of finite mixture models (issues addressed: unknown # of clusters, discrete component parameters)
   - https://github.com/zhenkewu/rewind
2) ***spotgear:*** **S**ubset **P**rofiling and **O**rganizing **T**ools for **G**el **E**lectrophoresis **A**utoradiography in **R**
   - 2D Bayesian image dewarping (registration) method for preprocessing GEA data ***prior*** to statistical analysis
   - https://github.com/zhenkewu/spotgear

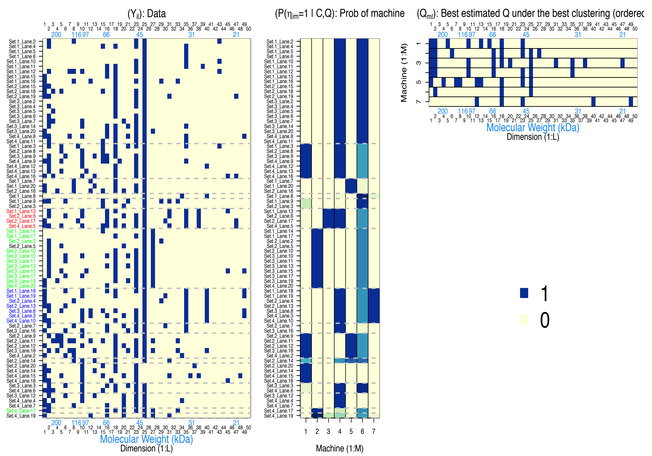

Figure 5. Model Results for GEA data. *Left*) Aligned data matrix for band presence or absence; row for 76 serum lanes, reordered into optimal estimated subgroups separated by gray horizontal lines "—–"; columns for L = 50 protein landmarks. A blue vertical line "|" indicates a band; *Middle*) lane-machine matrix for the probability of a lane (serum sample) having a particular machine. The blue cells correspond to high probability of having a machine in that column. Smaller probabilities are shown in lighter blue; *Right*) The estimated machine profiles. Here seven estimated machines are shown, each with component proteins shown by a blue bar "|".

References:

- **Wu Z** and Zeger SL (2018+). A Bayesian Approach to Restricted Latent Class Models for Scientifically-Structured Clustering of Multivariate Binary Outcomes. Working paper.
- **Wu Z** et al. (2017b). Estimating Autoantibody Signatures to Detect Autoimmune Disease Patient Subsets. Biostatistics. doi:10.1093/biostatistics/kxw037.