Likelihood based Clustering via Finite Mixtures

Using adjacent-categories logit model for ordinal data

Lingyu Li*, A/Prof Ivy Liu, A/Prof Richard Arnold

Victoria University of Wellington, New Zealand

lilingyunz@hotmail.com

TE WHARE WĀNANGA O TE ŪPOKO O TE IKA A MĀUI
VICTORIA
UNIVERSITY OF WELLINGTON

Introduction

- Consider a questionnaire response, rows as the observations, columns as the questions.
- Data is formed into a $n \times m$ matrix with

$$Y_{ij} = k, \text{if individual } i \text{ answered k on question } j; \qquad k = 1, 2, \dots, q$$

- Response is all ordinal which has the same number of categories q.
- The suggested model adjacent-categories logit model is for ordinal response variables.
- Row clustering assumes rows are from R number of clusters; column clustering assumes columns are from C number of clusters.
- The goal is to cluster rows into different clusters if it is row clustering; to cluster columns into different clusters for column clustering; to cluster rows and columns simultaneously for bi-clustering.
- Finite mixtures are a successful way to do clustering analysis.
- Need to estimate the parameters for the model via EM algorithm [2].

Oridinal Data

- In statistics, a variable consists of an ordinal scale is called an ordinal variable [1].
- Examples of ordinal variables:
  - Family spending on food: high, medium, low
  - Degree: high school, college, undergraduate, master, PhD
  - How often do people do exercise: never, rarely, occasionally, often

Adjacent-categories logit models

- In this model, the probability that Y_{ij} takes category k is characterized by the following log odds:

$$\log\left(\frac{P[Y_{ij} = k|\boldsymbol{x}_{ij}]}{P[Y_{ij} = k-1|\boldsymbol{x}_{ij}]}\right) = \mu_k + \boldsymbol{\delta}^T \boldsymbol{x}_{ij},$$
$$i = 1, \dots, n, \quad j = 1, \dots, m, \quad k = 2, \dots, q,$$

The vector $\boldsymbol{x}_{ij}$ is a set of predictor variables which can be categorical or continuous. However, the vector of parameters $\boldsymbol{\delta}$ represents the effects of $\boldsymbol{x}$ on the log odds of the response variable for **the category k relative to the category $k-1$ instead of the baseline category**. We also restrict $\mu_1 = 0$ to be sure of identifiability.

Column Clustering

- Columns are assumed a priori to come from any of $c = 1, \dots, C$ column groups with probabilities $\kappa_1, \dots, \kappa_C$.
- That is, we assume that the columns come from a finite mixture with C components where both C and the column-cluster proportions κ_c are unknown.
- Note also that $C < m$ and $\sum_{c=1}^{C} \kappa_c = 1$, and $\kappa_c \geq 0$.
- Let $P[Y_{ij} = k|j \in c] = \theta_{ick}$, which means the probability that observation $Y_{ij} = k$ given that column j belongs to column-cluster c.
- **The adjacent-categories logit model with column clustering has the form**:

$$\log\left(\frac{P[Y_{ij} = k|j \in c]}{P[Y_{ij} = k-1|j \in c]}\right) = \mu_k + \beta_c,$$
$$i = 1, \dots, n, \quad c = 1, \dots, C, \quad k = 2, \dots, q,$$

where μ_k is the kth intercept, β_c is the cth column-cluster effect.

- Through some mathematical induction, we have:

$$\theta_{ick} = P[Y_{ij} = k|j \in c] = \frac{\exp\left[\mu_k^* + (k-1)\beta_c\right]}{\sum_{\ell=1}^{q} \exp\left[\mu_\ell^* + (l-1)\beta_c\right]}$$
$$i = 1, \dots, n, \quad c = 1, \dots, C, \quad k = 1, \dots, q,$$

where $\beta_1 = 0, \mu_1 = 0$, and

$$\mu_k^* = \sum_{h=2}^{k} \mu_h = \mu_2 + \mu_3 + \dots + \mu_k.$$

- Assuming independence among the columns and, conditional on the columns, independence over the rows, the likelihood with column-clustering becomes:

$$L(\boldsymbol{\Omega}|\mathbf{Y}) = \prod_{j=1}^{m} [\sum_{c=1}^{C} \kappa_c \prod_{i=1}^{n} \prod_{k=1}^{q} (\theta_{ick})^{I(y_{ij}=k)}]$$

Estimation by using EM algorithm

We define the unknown column group memberships through the following indicator latent variables:

$$X_{jc} = I(j \in c) = \begin{cases} 1 \text{ if } j \in c \\ 0 \text{ if } j \notin c \end{cases} \quad j = 1, \dots, m \quad c = 1, \dots, C$$

where $j \in c$ indicates that column j is in column group c. It follows that:

$$\sum_{c=1}^{C} X_{jc} = 1, \; j = 1, \dots, m,$$

Given a value for the number of the mixture components C, the EM algorithm proceeds as follows:

E step:

Update $\widehat{\boldsymbol{x}}$. Given $\boldsymbol{Y}$ and values for $\kappa_c, \mu_k, \alpha_r$, estimate $E[X_{jc}|\{y_{ij}\}, \boldsymbol{\Omega}] = x_{jc}$ as:

$$\widehat{x}_{jc}^{(t)} = \frac{\widehat{\kappa}_c^{(t-1)} \prod_{i=1}^{n} \prod_{k=1}^{q} (\widehat{\theta}_{ick}^{(t-1)})^{I(y_{ij}=k)}}{\sum_{g=1}^{C} [\widehat{\kappa}_g^{(t-1)} \prod_{i=1}^{n} \prod_{k=1}^{q} (\widehat{\theta}_{ick}^{(t-1)})^{I(y_{ij}=k)}]} \tag{1}$$

M step:

The M-step has two parts:

(1) Update the column cluster propotions using:

$$\widehat{\kappa}_c^{(t)} = \frac{1}{m} \sum_{j=1}^{m} E[X_{jc}|\{y_{ij}\}, \boldsymbol{\Omega}^{(t-1)}] = \frac{1}{m} \sum_{j=1}^{m} \widehat{x}_{jc}^{(t)}.$$

(2) Numerically maximize the complete data log-likelihood:

$$Q^{(t)} = \sum_{j=1}^{m} \sum_{c=1}^{C} \widehat{x}_{jc}^{(t)} \log(\widehat{\kappa}_c^{(t-1)}) + \sum_{i=1}^{n} \sum_{j=1}^{m} \sum_{k=1}^{q} \sum_{c=1}^{C} \widehat{x}_{jc}^{(t)} I(y_{ij} = k) \log(\theta_{ick}).$$

given $\widehat{x}_{jc}$ from the E-step. We maximize $Q^{(t)}$ to obtain new values for the parameters μ_k, β_c.

A new cycle starts from using the parameters getting from the M-step in the E-step. This process repeats until estimates have converged. There is a risk of convergence to local maxima due to multimodality on the likelihood surface, and thus it is important to use several initial values to start the EM algorithm.

Row Clustering

- Row clustering is very similar to column clustering since they are both one-way clustering.
- Setting R as the number of row clusters in our dataset. Each cluster with proportion $\pi_1, \pi_2, \dots, \pi_R$. We assume the rows come from a finite mixture with R components where both R and π_r are all unknown. Note that $R < n$ and $\sum_{r=1}^{R} = 1$.
- Let $P[Y_{ij} = k|i \in r] = \theta_{rjk}$,

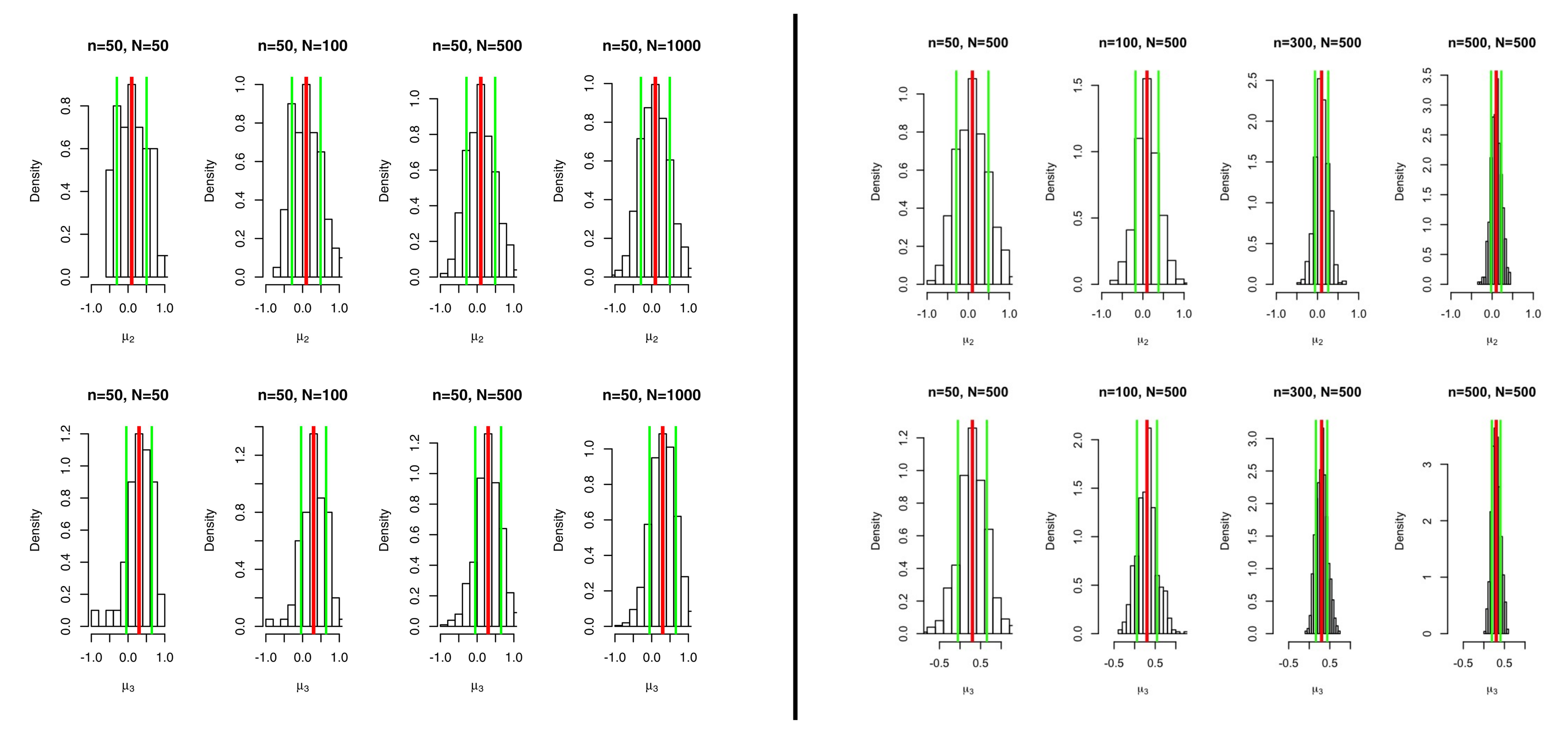
$$\log\left(\frac{P[Y_{ij} = k|i \in r]}{P[Y_{ij} = k-1|i \in r]}\right) = \mu_k + \alpha_r,$$
$$i = 1, \dots, n, \quad j = 1, \dots m, \quad r = 1, \dots, R, \quad k = 2, \dots, q,$$

Simulation

- A simplest adjacent-categories logit model has the form as follows:

$$\log\left(\frac{P[Y_i = k]}{P[Y_i = k-1]}\right) = \mu_k, \qquad k = 2, \dots, q$$

- Simulation results when the true parameter value $\mu_2 = 0.1, \mu_3 = 0.3$. The number of response in each datasets is n, while the number of simulation datasets (replicates) is N



Future Work

- Row clustering, column clustering and bi-clustering using adjacent-categories logit model via a finite mixture model.
- Use simulation study and heat maps to evaluate our proposed model on row/column clustering and biclustering. Apply model selection methods such as AIC and BIC.
- Evaluate and compare finite mixture clustering models and logistic regression models through an application in Linguistics.
- Using randomised quantile residuals to construct a goodness-of-fit test for fuzzy clustering: Use $\widehat{X}_{jc}$ as the weight, then calculate the weighted randomised quantile residual:

$$E_j = \sum_{c=1}^{C} \widehat{X}_{jc} \epsilon_{jc}$$

- Apply LASSO [4] on clustering and compare it with fuzzy clustering via finite mixtures. By solving the quasi-likelihood equations such as GEE [3] subject to

$$\sum_{j<h}^{m} \omega_{jh} |\beta_j - \beta_h| \leq s \qquad \text{and} \qquad \sum_{j=1}^{m} \beta_j = 0$$

where ω_{jh} is the weight, β_j is the column effect of the jth column. If we have very similar values of $\widehat{\beta}_j$, we can merge them and cluster the corresponding columns into the same clusters.

References

[1] Agresti, A. (2010). *Analysis of ordinal categorical data*, volume 656. John Wiley & Sons.

[2] Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society. Series B (methodological)*, pages 1–38.

[3] Hardin, J. W. and Hilbe, J. M. (2002). *Generalized estimating equations*. Chapman and Hall/CRC.

[4] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288.

Acknowledgements

This poster presence for this conference is supported by a Marsden grant from the Royal Society of New Zealand.