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

Introduction

• Consider a questionnaire response, rows as the observations, columns as the questions.
• Data is formed into a n x m matrix with
  \[ Y_{ij} = k, \] if individual i answered k on question j;
  \[ k = 1, 2, \ldots, q \]
• 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].

Ordinal 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:
  \[
  \log \left( \frac{P(Y_{ij} = k|x_{ij})}{P(Y_{ij} = k - 1|x_{ij})} \right) = \mu_k + \beta_x x_{ij},
  \]
  \[ i = 1, \ldots, n, \quad j = 1, \ldots, m, \quad k = 2, \ldots, q. \]

The vector \( x_{ij} \) is a set of predictor variables which can be categorical or continuous. However, the vector of parameters \( \beta \) represents the effects of \( x \) 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 to come from any of c₁, c₂, ..., C column groups with probabilities \( \pi_{c_1}, \pi_{c_2}, ..., \pi_c \).
• That is, we assume the columns come from a finite mixture with C components where both C and the column-cluster proportions \( \pi_c \) are unknown.
• Note also that \( C < m \) and \( \sum_{c=1}^{C} \pi_c = 1 \), and \( \pi_c \geq 0 \).
• Let \( R(Y_{ij} = k|j \in c) = \theta_{ck} \), 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:
  \[
  \log \left( \frac{P(Y_{ij} = k|j \in c)}{P(Y_{ij} = k - 1|j \in c)} \right) = \mu_k + \beta_c, 
  \]
  \[ i = 1, \ldots, n, \quad c = 1, \ldots, C, \quad k = 2, \ldots, q, \]

where \( \mu_k \) is the \( k \)th intercept, \( \beta_c \) is the \( c \)th column-cluster effect.
• Through some mathematical induction, we have:
  \[
  \theta_{ck} = P(Y_{ij} = k|j \in c) = \frac{\exp \left[ \mu_k + (k - 1)\beta_c \right]}{\sum_{i=1}^{m} \exp \left[ \mu_k + (i - 1)\beta_c \right]}, 
  \]
  \[ i = 1, \ldots, n, \quad c = 1, \ldots, C, \quad k = 1, \ldots, q, \]

where \( \beta_1 = 0, \mu_0 = 0, \) and \( \mu_k = \frac{1}{C} \sum_{c=1}^{C} \mu_c + \mu_1 + \ldots + \mu_k \).
• Assuming independence among the columns and, conditional on the columns, independence over the rows, the likelihood with column-clustering becomes:
  \[
  L(Y) = \prod_{j=1}^{n} \prod_{c=1}^{C} \prod_{k=1}^{q} \left( \pi_c \frac{\exp \left[ \mu_k + (k - 1)\beta_c \right]}{\sum_{i=1}^{m} \exp \left[ \mu_k + (i - 1)\beta_c \right]} \right).
  \]

Estimation by using EM algorithm

We define the unknown column group memberships through the following indicator latent variables:
\[
X_{c,j} = I(j \in c) - 1 \quad \text{if} \quad c \in \mathcal{C}_{j}, 
0 \quad \text{otherwise}
\]
where \( j \in \mathcal{C}_{j} \) indicates that column j is in column group \( c \). It follows that:
\[
\sum_{c=1}^{C} X_{c,j} = 1, \quad j = 1, \ldots, m.
\]

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

\textbf{E step:}
Update \( \mathbf{e} \). Given \( \mathbf{Y} \) and values for \( \pi_0, \mu_0, \beta_0 \), estimate \( \mathbb{E}[X_{c,j}|\mathbf{Y}] \). Recall \( \mathbf{e} = x_{jk} \) as:
\[
\mathbb{E}[X_{c,j}|\mathbf{Y}] = \frac{\gamma_c \exp \left( \sum_{k=1}^{q} (\mu_k + (k - 1)\beta_c) x_{jk} \right)}{\sum_{c=1}^{C} \gamma_c \exp \left( \sum_{k=1}^{q} (\mu_k + (k - 1)\beta_c) x_{jk} \right)}.
\]

\textbf{M step:}
(1) Update the column cluster proportions using:
\[
\gamma_c = \frac{1}{m} \sum_{j=1}^{m} \mathbb{E}[X_{c,j}|\mathbf{Y}], \quad \forall c = 1, \ldots, C
\]
(2) Numerically maximize the complete data log-likelihood:
\[
Q(\theta) = \sum_{j=1}^{m} \sum_{k=1}^{C} \mathbb{E}[\log(P(Y_{ij} = k|\theta))|\mathbf{Y}].
\]
given \( \theta \), from the E-step. We maximize \( Q(\theta) \) to obtain new values for the parameters \( \mu_k, \beta_c \).

A new cycle starts from using the parameters getting from the M-step in the E-step. This process repeats until estimates become stable. There is a risk of convergence to local maxima due to multidimensionality 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 \( \pi_1, \pi_2, \ldots, \pi_R \).
• We assume the rows come from a finite mixture with \( R \) components where both \( R \) and \( \pi_r \) are all unknown. Note that \( C < n \) and \( \sum_{r=1}^{R} \pi_r = 1 \).
• Let \( P(Y_{ij} = k|i \in r) = \theta_{rk} \).

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

Simulation

• A simplest adjacent-categories logit model has the form as follows:
  \[
  \log \left( \frac{P(Y_{ij} = k)}{P(Y_{ij} = k - 1)} \right) = \mu_k, \quad k = 2, \ldots, q.
  \]

• Simulation results when the true parameter value \( \mu_2 = 0, \mu_3 = 0.3 \). The number of response in each dataset 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 \( \hat{X}_X \) as the weight, then calculate the weighted randomised quantile residual:
  \[
  E_T = \sum_{i=1}^{n} \hat{X}_X Y_{ij}.
  \]

• Apply LASSO [4] on clustering and compare it with fuzzy clustering via finite mixtures. By solving the quasi-likelihood equations such as GEE [1] subject to
  \[
  \sum_{j=1}^{m} \omega_{jk} | \beta_j - \beta_j^0 | \leq s \quad \text{and} \quad \sum_{j=1}^{m} \beta_j = 0
  \]
where \( \omega_{jk} \) is the weight, \( \beta_j \) is the column effect of the \( j \)th column. If we have very similar values of \( \beta_j \), we can merge them and cluster the corresponding columns into the same clusters.

References


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