Model-based semi-supervised clustering via finite-mixtures using proportional odds model for ordinal data

Ying Cui

School of Mathematics and Statistics,
Victoria University of Wellington, New Zealand

2024 Australasian Applied Statistics Conference
September, 2024

Outline

Introduction

Ordinal variable

* A type of categorical variable with fixed set of categories.
* has an ordered scale of categories (i.e. Likert scale responses to a survey question).

Three common used ordinal models:

* Proportional odds model (McCullagh, 1980).
* Ordered stereotype model (Anderson, 1984).
* Adjacent-categories logit model (Simon, 1974).

Model-based clustering

* An approach describes the clustering process via statistical densities.
* A method based on finite-mixture densities.

Introduction

Semi-supervised clustering for ordinal data:

* Unsupervised clustering method sometimes can not resulted in consistency between labeled and unlabeled data.
* Semi-supervised clustering can incorporate the information of known knowledge of labeled data to cluster the unlabeled data.
* Majority of semi-supervised clustering for analyzing the ordinal data is not appropriate (treating as continuous or nominal without considering the order).
* There is no likelihood-based semi-supervised clustering approach proposed for ordinal data.

Proportional odds model

* Consider an $n \times p$ data matrix, with entry y_{ij}.
* Each entry has fixed q response categories.
* Let the probabilities for the response categories for y_{ij} be $\theta_{ij1}, \theta_{ij2}, \dots, \theta_{ijq}$ such that $\sum_{k=1}^{q} \theta_{ijk} = 1, \ \forall i, j$.

$$\theta_{ijk} = \begin{cases} \dfrac{\exp(\mu_k - \alpha_i - \beta_j - \gamma_{rj})}{1 + \exp(\mu_k - \alpha_i - \beta_j - \gamma_{rj})} & k = 1 \\ \dfrac{\exp(\mu_k - \alpha_i - \beta_j - \gamma_{rj})}{1 + \exp(\mu_k - \alpha_i - \beta_j - \gamma_{rj})} - \dfrac{\exp(\mu_{k-1} - \alpha_i - \beta_j - \gamma_{rj})}{1 + \exp(\mu_{k-1} - \alpha_i - \beta_j - \gamma_{rj})} & 1 < k < q \\ 1 - \sum_{k=1}^{q-1} \theta_{ijk} & k = q. \end{cases}$$

Proportional odds model

Or we can express it using logistic form of the linear predictors:

$$\text{logit}\left[P(Y_{ij} \leq k)\right] = \begin{cases} \mu_k - \alpha_i - \beta_j - \gamma_{rj} & 1 \leq k < q \\ +\infty & k = q. \end{cases}$$

Proportional odds model with clustering

Proportional odds model with clustering

* Assume the rows with unlabeled cluster memberships come from finite mixture with R components.
* The previous logistic form of the linear predictors becomes:

$$\text{logit}\left[P(Y_{ij} \le k)\right] = \begin{cases} \mu_k - \alpha_r - \beta_j - \gamma_{rj} & 1 \le k < q \\ +\infty & k = q, \end{cases}$$

* The constraints are:
    * $\mu_1 < \mu_2 < \cdots < \mu_q = +\infty$.
    * $\sum_{r=1}^{R} \alpha_r = \sum_{j=1}^{p} \beta_j = 0$.
    * $\{\gamma_{ij}\} : \sum_{j=1}^{p} \gamma_{ij} = 0 \ \forall i$ and $\sum_{i=1}^{n} \gamma_{ij} = 0 \ \forall j$.

Data likelihood

$$
\begin{aligned}
& L\left[\boldsymbol{\Omega}, \boldsymbol{\pi} \mid \boldsymbol{Y}\right] \\
= \quad & \left(\prod_{i=1}^{n_\ell} \prod_{r=1}^{R} \prod_{j=1}^{p} \prod_{k=1}^{q} \theta_{r_i jk}^{I(y_{ij}=k) I(r_i=r)}\right) \left(\prod_{i=n_\ell+1}^{n_\ell+n_u} \sum_{r=1}^{R} \pi_r \prod_{j=1}^{p} \prod_{k=1}^{q} \theta_{rjk}^{I(y_{ij}=k)}\right).
\end{aligned}
$$

where:

* n_ℓ and n_u represent the number of cases with labeled and unlabeled cases respectively.
* $I\left(y_{ij}=k\right)$ is an indicator variable that is 1 if y_{ij} is in category k, and 0 otherwise;
* $I\left(r_i=r\right)$ is an indicator variable that is 1 if row i with known cluster membership r_i belongs to row cluster r, and 0 otherwise.
* θ_{rjk} is the probability of each entry y_{ij} has response in category k at row cluster r and column j.

Expectation Maximization Algorithm

* The EM algorithm is mostly applicable in calculating maximum likelihood estimates through providing an iterative procedure on incomplete data problems (McLachlan & Krishnan, 2015).
* **E-step:** is responsible for updating the latent variable z_{ir}, which is the posterior probability of cluster membership, to estimate missing cluster membership.
* **M-step:** updates the maximum likelihood estimates for parameters μ_k,α_r,β_j,γ_{rj}, and π_r using the estimates z_{ir} obtained from the E-step.

A new cycle starts when the parameters from the M-step are used in the E-step. This process repeats until estimates have converged.

Simulation study

Data set structure

* Fixed $p = 5$ columns and $q = 3$ ordinal response categories. Three possible choices of rows $n = (300, 1000, 3000)$ and rows are equally distributed among the $R = 3$ clustering groups.

* The true values of model's parameters are:

  - $\{\alpha_1, \alpha_2, \alpha_3\} = \{-2, 0, 2\}$;
  - $\{\beta_1, \beta_2, \beta_3, \beta_4, \beta_5\} = \{-2, -1.5, 0.3, 1.0, 2.2\}$;
  - $\{\mu_1, \mu_2\} = \{-0.693, 1.307\}$.

Scenarios

* fixed the percentage of cluster memberships that are known, denoted as $m\% = 10\%$.

* varied by the distribution of memberships within that labeled portion, denoted as $\{g_r\}$.

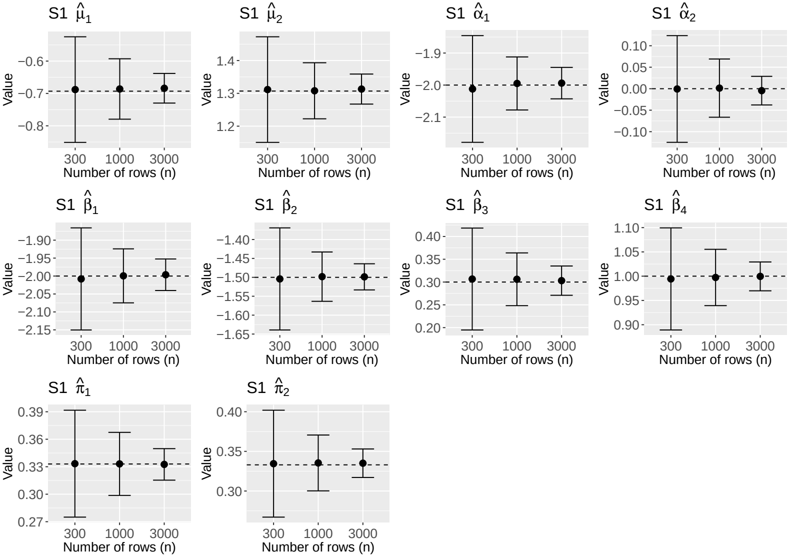
For each combination of scenario and n, we simulated 100 replicate datasets.

Simulation study: scenarios $1 \sim 3$

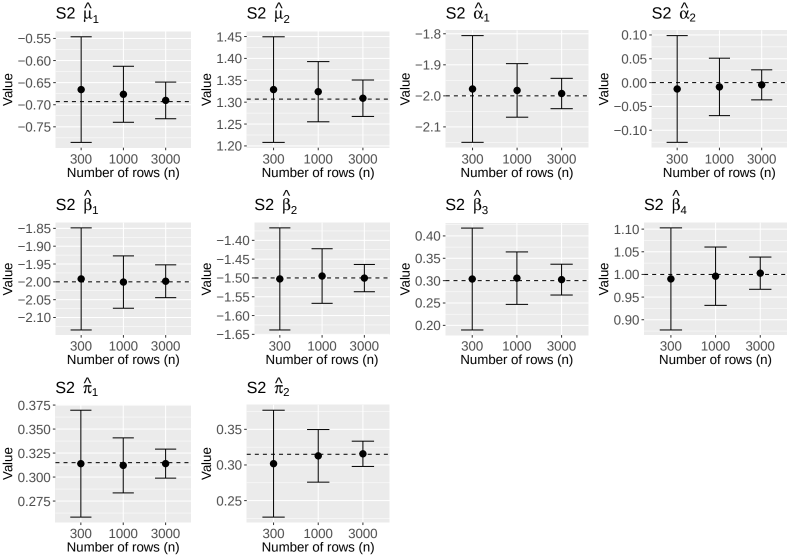
Table 1: Scenarios where all rows are equally distributed among the $R = 3$ clusters for semi-supervised row clustering approach.

Scenario1 m%= 10%	**Scenario2** m%= 10%	**Scenario3** m%= 10%
π_1=0.333	π_1=0.315	π_1=0.260
π_2=0.333	π_2=0.315	π_2=0.370
π_3=0.334	π_3=0.370	π_3=0.370
g_1=0.333	g_1=0.500	g_1=1.000
g_2=0.333	g_2=0.500	
g_3=0.334		

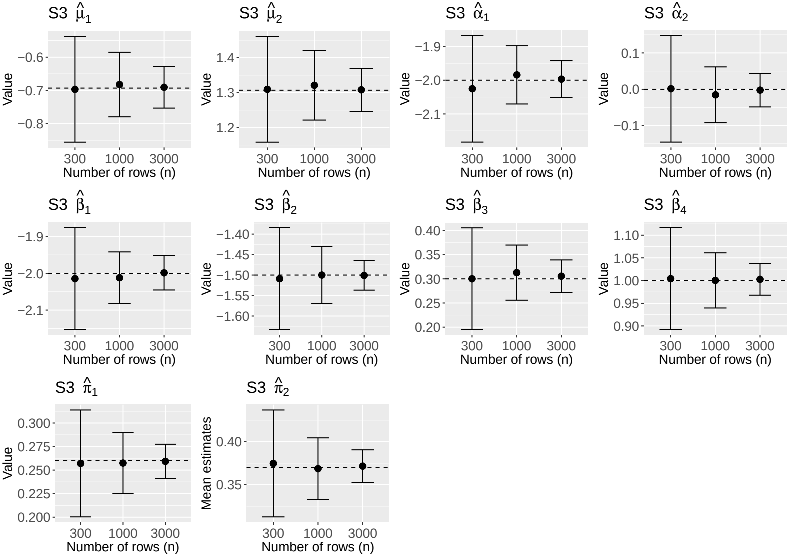
Parameter estimates error bars: Scenario 1



Parameter estimates error bars: Scenario 2



Parameter estimates error bars: Scenario 3



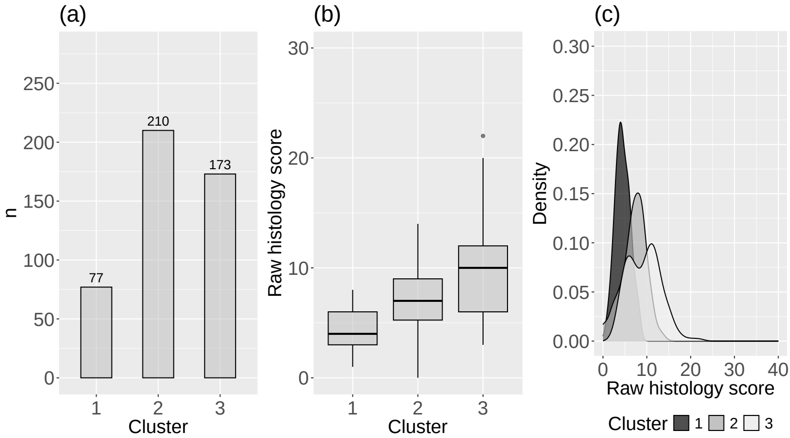
Case study: Salmon fish from Cawthron

* New Zealand's largest independent science organization, the Cawthron Institute in the aquaculture sector.
* Cawthron Institute runs many different trials and collects data from salmon in commercial farms in New Zealand.
* Cawthron collected a variety of health markers, such as blood, growth performance, feeding condition, nutrient composition, and histology of individual tissues for fish.
* Some of the markers are gathered in a destructive manner which makes the corresponding markers expensive to collect. Thus, the Cawthron Institute would like to know which other non-destructive markers can be used as proxies for fish health.

Case study: labeled clusters generation

- The initial known cluster memberships are generated from previous existing **unsupervised** model-based row clustering approach using the proportional odds model (Matechou et al., 2016).
- The data has 460 fish and 9 destructively-collected histology measurement variables, each ordinal response has 4 categories which represent the level of abnormality.
- AIC and BIC choose the Model with $R = 3$ row clusters (linear predictor: $\mu_k - \alpha_r - \beta_j - \gamma_{rj}$).

Case study: Three levels of fish health status



Case study: Large data with Growth measurement features

* This large dataset has 3488 salmon fish (with 460 labeled fish) as the rows, values of condition factor (Froese, 2006) at 8 time stages as the columns.
*

$$\mathrm{CF} = \left(\frac{\mathrm{W}}{\mathrm{L}^3}\right) \times 100,000. \tag{4.1}$$

where:

- W represents the fish weight(g).
- L is the fish fork length(mm).

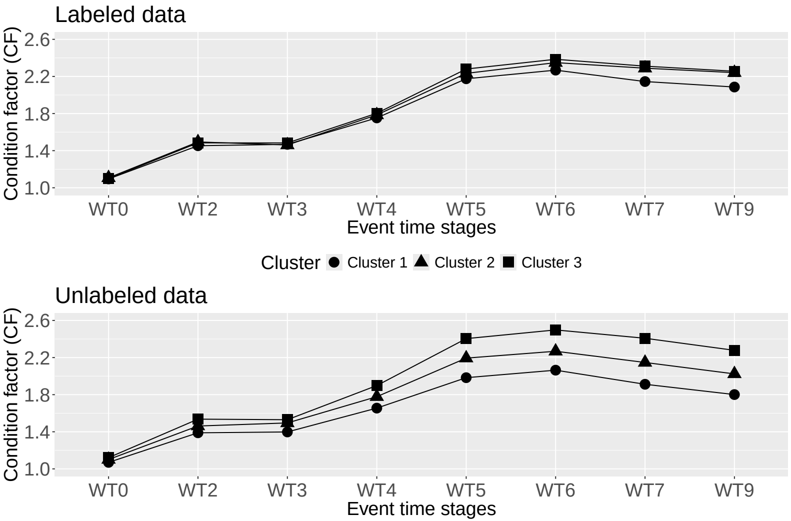
* For CF in each time stage, we code the value from quantile 0% to 25%, 25% to 50%, 50% to 75%, and 75% to 100% as ordinal response 1, 2, 3, and 4 correspondingly.

Case study: Model selection

Table 2: Suit of semi-supervised row clustering models with fitted $\widehat{R} = 3$ applied on the data with variable condition factor (CF) at 8 different time stages.

Information Criteria	$\operatorname{logit}\left[P\left(Y_{ij} \leq k\right)\right], \ 1 \leq k \leq q$		
	$\mu_k - \alpha_r$	$\mu_k - \alpha_r - \beta_j$	$\mu_k - \alpha_r - \beta_j - \gamma_{rj}$
AIC	48336.2	48257.9	**47792.9**
AICc	48336.2	48257.9	**47793.0**
AICu	48344.2	48272.9	**47822.0**
AIC3	48343.2	48271.9	**47820.9**
BIC	48393.8	48373.2	**48023.5**

Case study: Scatterplots of CF at 8 stages for three clusters



Discussion

* The semi-supervised model-based clustering approach takes into account the ordinal nature of the response data and incorporates information about existing clustering memberships to cluster data with unknown memberships.
* A simulation study was conducted and the results indicate the model parameter estimation perform well in defined scenarios.
* Clustering pattern detected for classifying the health status of fish Trial data collected from Cawthron. The unhealthy fish are likely to be fat and short when they grow up.

Further study

* Evaluate the performance of parameter estimation in other scenarios.
* Aim to develop another semi-supervised clustering strategy using the ordered stereotype model as the basic structure, and the corresponding R package will be built.
* Conduct the clustering analysis for fish farm data collected from Cawthron to classify the fish health.


thank you!

Anderson, J. A. (1984). Regression and ordered categorical variables. *Journal of the Royal Statistical Society. Series B, Methodological*, **46**(1), 1–30.

Froese, R. (2006). Cube law, condition factor and weight-length relationships: history, meta-analysis and recommendations. *Journal of applied ichthyology*, **22**(4), 241–253.

Matechou, E., Liu, I., Fernández, D., Farias, M., & Gjelsvik, B. (2016). Biclustering models for two-mode ordinal data. *Psychometrika*, **81**(3), 611–624.

McCullagh, P. (1980). Regression models for ordinal data. *Journal of the Royal Statistical Society. Series B, Methodological*, **42**(2), 109–142.

McLachlan, G. J. & Krishnan, T. (2015). *The EM Algorithm and Extensions*. John Wiley and Sons, Inc., 2nd edition.

Simon, G. (1974). Alternative analyses for the singly-ordered contingency table. *Journal of the American Statistical Association*, **69**(348), 971–976.