Model-Based Clustering using Adjacent-Categories Logit Models via Finite Mixture Model

Lingyu Li* Ivy Liu * Richard Arnold*

Abstract

This paper presents cluster analysis of ordinal data utilising the natural order information of ordinal data. Three models commonly used in ordinal modelling are discussed: the proportional odds model, the adjacent-categories logit model and the ordered stereotype model.

In our research, the data take the form of a matrix where the rows are subjects, and the columns are a set of ordinal responses by those subjects to, say, the questions in a questionnaire. We implement model-based fuzzy clustering via a finite mixture model, in which the subjects (the rows of the matrix) and/or the questions (the columns of the matrix) are grouped into a finite number of clusters. We will explain how to use EM (Expectation-Maximisation) algorithm to estimate the model parameters. Specifically, we illustrate the details of using adjacent-categories logit model to perform row/column and bi-clustering. This clustering method differs from other typical clustering methods such as K-means or hierarchical clustering, because it is a likelihood-based model, and thus statistical inference is possible.

Key Words: clustering, finite mixture model, ordinal data, EM algorithm, categorical data analysis

1. Introduction

There are two basic types of data: numerical (quantitative) and categorical (qualitative) variables. Numerical variables are typically measurement (continuous) or counts (discrete) (Clark and Randal, 2011). For example a person's height, weight, IQ or blood pressure; or counts, such as the number of customers visiting a store in one day, how many dogs each household has, or how many computers in a lab. Categorical variables are the variables which have a measurement scale consisting of a set of categories (Agresti, 2013). Categorical variables can be separated into nominal and ordinal types. Nominal variables do not have a natural order, such as ethnicity (European, Chinese, Māori, etc.), gender (female, male and bisexual) and travel method (bus, car, walk, etc.). Unlike nominal data, ordinal data have a specific ordering. Examples of ordinal variables include health status (poor, reasonable, good and excellent), letter grades (A, B, C, D), socioeconomic status (high, middle, and low), satisfaction level to a service (very unsatisfied, somewhat unsatisfied, neutral, somewhat satisfied, very satisfied) and any other Likert scale.

Many researchers treat ordinal variables as continuous or nominal (Fernández et al., 2019). When treating them as continuous variables, by ignoring the categorical nature and assigning numerical scores to the ordered categories, methods such as ordinary least squares (OLS) can be used to fit the model. However, this can lead to unsatisfactory results which include: predicted values below the lowest category score or above the highest category score (Agresti, 2010); assuming equal numerical distance among different categories is not suitable for some data (for instance, the difference in the severity of a pain expressed in level 5 rather than 4 may be much more than the difference between level 2 and level 1). Another option is treating ordinal variables as nominal variables. This makes all the analysis methods that work on nominal variables can be applied to ordinal variables as well: Chi-square test (Franke et al., 2012, McHugh, 2013), multinomial logistic regression (Kwak and Clayton-Matthews, 2002, Chan, 2005), log-linear models (Christensen, 2006),

*Victoria University of Wellington, Kelburn Parade, Wellington 6140, New Zealand

etc. However, if the ordering is an important part of our research question, we will not find the insights we need. Losing the ordering information also makes statistical analysis and inference inappropriate.

Clustering for ordinal data faces a lot of challenges as well. Clustering or cluster analysis aims to group a data set into different clusters by searching and analysing the response patterns (Hand, 2007, Chapter 5). Traditional clustering methods such as hierarchical clustering (Murtagh, 1983, Murtagh and Contreras, 2012), centroid-based clustering, association analysis (Strehl et al., 1999) have well-developed and documented (Jain et al., 1999, Han et al., 2011, Aggarwal and Reddy, 2013, Witten et al., 2016). However, these traditional clustering methods are not likelihood-based, and statistical inference is not available.

In our research, the dataset can be organised into a matrix. Consider questionnaire data: rows stand for respondents and columns stand for questions. Fuzzy clustering via a finite mixture model allows to group the rows and/or columns of such data matrices, see the example, Pledger and Arnold (2014) who clustered matrices of binary and count data. Through likelihood-based fuzzy clustering, maximum likelihood estimation of parameters can be carried out, and likelihood ratio tests or information criteria such as AIC (Akaike, 1987), BIC(Schwarz et al., 1978) and DIC (Berg et al., 2004), etc. are available for model selection.

Our research applies fuzzy clustering via finite mixtures to the adjacent-categories logit model (Agresti, 1999) for ordinal data. This work is an extension of likelihood-based models in Pledger and Arnold (2014).

This paper is structured as follows. Section 2 reviews relevant literature related to our research. The topics include ordinal data, ordinal data modelling, finite mixtures and model-based clustering. Section 3 introduces the adjacent-categories logit model and how it applies to fuzzy clustering via a finite mixture. Section 4 describes how we can use a simulation study to evaluate our proposed model. Section 5 lists our next steps of the research.

2. Literature Review

In this chapter, we review the existing literature relevant to our research.

2.1 Ordinal modeling

Ordinal trend models are a class of regression models for ordinal data. They can be separated into three important types: logistic regression models using cumulative logits, logistic regression models without using cumulative logits, and other ordinal multinomial response models. Logistic regression models using cumulative logits use the cumulative probabilities of response categories. The most commonly used is the proportional odds version of the cumulative logit model, which is reviewed below. Logistic regression models without using cumulative logits include adjacent-categories logit models, continuation-ratio logit models and stereotype models. These models do not use the accumulative probabilities for the response categories. Instead, they use single response probabilities to specify the model structure. We give a review of the continuation-ratio logit models as an example of this type of model. Other ordinal multinomial response models use link functions other than the logit, such as the probit link and log-log link functions. The details of these different ordinal models can refer to Agresti (2010, Chapter 3-5).

Ordinal response models have been applied widely in applications. For example, recent applications include Lanfranchi et al. (2014), Cameron et al. (2014), Donneau et al. (2015), Bürkner and Vuorre (2018) and Ursino and Gasparini (2018). Before giving the model

we use, we review a few existing ordinal models. For a detailed review see Ananth and Kleinbaum (1997), Liu and Agresti (2005).

Proportional odds version of the cumulative logit models

One of the most popular models for ordinal variables is the *cumulative logit model*. This model attracted researchers' attention after the seminal article by McCullagh (1980). Suppose we have q ordered response categories. The model is specified:

$$\begin{aligned} \log\left[\frac{P(Y_i \le k)}{1 - P(Y_i \le k)}\right] &= \mu_k + \boldsymbol{\delta}^T \boldsymbol{x}_i \\ &= \mu_k + \delta_1 x_{i1} + \delta_2 x_{i2} + \cdots + \delta_p x_{ip}, \qquad k = 1, 2, \ldots, q-1 \end{aligned} \tag{1}$$

where Y_i stands for the outcome variable for the ith subject, p is number of covariates we may have, $\boldsymbol{\delta}$ is a column vector that contains all the parameters for the covariates, $\boldsymbol{x}_i$ are covariate values.

The parameters $\{\mu_1, \mu_2, \ldots, \mu_{q-1}\}$ are called cutpoints. The column vector $\boldsymbol{\delta} = (\delta_1, \delta_2, \ldots, \delta_p)^T$ contains the parameters which control the effects of the explanatory variables. The reason it is called the cumulative logit model is because we accumulate the probabilities of the first k response levels in the specification. The cutpoints $\{\mu_k\}$ must increase as k increases because $P(Y_i \le k)$ is increasing at fixed covariate values $\boldsymbol{x}_i$.

The model in equation (1) is also referred as the *proportional odds model*. However, Agresti (2010) pointed out that the name is not specific enough because there are other models that can have proportional odds form as well. This model is recommended to be referred to as the *proportional odds version of the cumulative logit model* (Agresti, 2010, section 3.3.1).

Stereotype models

The ordered stereotype model was introduced by Anderson (1984). Once again, suppose we have the ordered categorical variable Y which has q ordered levels, the ordered stereotype model is specified by

$$\log\left[\frac{P(Y_i = k)}{P(Y_i = 1)}\right] = \mu_k + \phi_k \boldsymbol{\delta}^T \boldsymbol{x}_i \qquad (0 = \phi_1 \le \phi_2 \le \cdots \le \phi_q = 1)$$
$$k = 2, \ldots, q,$$

where the monotone increasing constraint on the ϕ parameters must be included to ensure that the response variable Y is ordinal (Anderson, 1984). The baseline category is the first category, the parameters $\{\mu_2, \ldots, \mu_q\}$ are the intercepts (which are unconstrained), and $\{\phi_2, \ldots, \phi_{q-1}\}$ are the parameters which can be interpreted as the "score"for the categories of the response variable Y. In order to ensure identifiability, we set $\mu_1 = \phi_1 = 0$ and $\phi_q = 1$.

Our research focuses on the adjacent-categories logit model which will be introduced in section 3.

2.2 Clustering for ordinal data

Clustering is the unsupervised classification of patterns into groups. These groups are called clusters. Unsupervised classification means that we do not know anything about cluster membership of any observations, in other words, there are no predefined clusters (Bohte et al., 2002). Cluster analysis is useful in a variety of areas. In marketing, it has

been used to analyse customer behaviour with demographics. For example, when we have a large dataset about customers which includes the products the customers buy and also description information on each customer, say the age, nationality and the time they purchased the product. If we want to identify which customers are likely to buy particular products, by clustering the customers about their shopping behaviour, we will be able to get some information about what future customers will buy. Other applications include earth observation database record the land use types by clustering, city planning based on area cluster; also earthquake studies (Lin et al., 2007). Also, clustering is useful in many grouping, decision-making and machine-learning situations, including data mining and pattern classification.

Currently, a lot of clustering methods are only based on a distance matrix, which does not provide a likelihood-based estimation of the data. Major clustering approaches include partitioning algorithm, such as the K-means clustering algorithm (Jobson, 1992, Lewis et al., 2005, McCune et al., 2002) and K-Medoids; hierarchical clustering (Johnson, 1967, Kaufman and Rousseeuw, 1990); association analysis (Manly, 2005). However, the lack of a statistical likelihood underlines these methods makes statistical inference impossible. In our research, we use model-based clustering. This means that we have a distributional description for each component, and allow us to calculate the probabilities of cluster membership for each observation. Furthermore, we can make inference about the number of clusters.

2.3 Clustering via finite mixtures

The finite mixture model idea was introduced by Pearson (1894). It is assumed that data are from $C \geq 1$ groups. Each obeservation in data is a realization y from a finite mixture density,

$$f(y;\Omega) = \sum_{c=1}^{C} \kappa_c f_c(y;\theta_c).$$

Here, Ω contains all the unknown parameters in the finite mixtures, κ_c is the a priori probability that a data point belongs to group c, and θ_c is the vector of unknown parameters controlling the cth component density of the finite mixture $f_c(y;\theta_c)$. Notice that κ_c are all non-negative and

$$\sum_{c=1}^{C} \kappa_c = 1, \qquad 0 \leq \kappa_c \leq 1, \qquad c = 1, \ldots, C.$$

Consider a simple case, a dataset is assumed from two normal distribution with different means and variances. The dataset we have is the population data, which contain two subpopulations. We can estimate the parameters for these two normal distributions and estimate for each data point which component they most likely belong to. This is a simple finite mixture which assumes the population data is from two subpopulations. If we have more subpopulations, then we have a finite mixture model with more number of mixture components.

Clustering identifies how many subgroups a dataset has and associates each data point with a cluster, or possibly several clusters if its membership is unclear. Finite mixture models can be used to do clustering which treats the cluster membership as missing information. A literature review about how finite mixture model has been proposed can see Melnykov et al. (2010). In our case, with rows are subjects and columns are questions, we apply probabilistic models using finite mixtures and carry out fuzzy clustering for the rows or columns (one-way clustering) and both (two-way clustering). This approach enables us

to use likelihoods, which also makes model selection possible. Models can be fitted using the EM algorithm and compared through model selection methods such as the Akaike Information Criterion (AIC) or Bayesian Information Criterion (BIC).

Biclustering, two-way clustering and co-clustering all refer to simultaneous clustering of rows and columns. Biernacki et al. (2000), Pledger (2000) have proposed biclustering models for categorical data. Govaert and Nadif (2013) gives a thorough treatment of the subject. Recently, Fernández et al. (2016) have used the ordered stereotype model and Costilla et al. (2015) used proportional odds model and trend odds model to handle row, column and bi-clustered ordinal data, which have extended these models to ordinal responses. The purpose of our research is to extend these models by using adjacent-categories logit models.

The estimation of the parameters for the finite mixture model is through the Expectation-Maximization (EM) algorithm. It is an iterative method for finding maximum likelihood estimates of parameters in statistical models, where the model depends on unobserved latent variables. The EM algorithm was first proposed by Dempster et al. (1977). The formulation of the EM Algorithm is explained in McLachlan and Krishnan (2007, Section 1.5).

3. The Adjacent-Categories Logit Model

3.1 Data formation

Let $\boldsymbol{Y}$ be a $n \times m$ matrix where each cell y_{ij} is equal to any of the q ordinal categories, and:

$i = 1, \ldots, n$ (where the rows are different subjects);
$j = 1, \ldots, m$ (where the columns are different questions);
$k = 1, \ldots, q$ (ordinal categories).

3.2 Adjacent-categories logit model structure

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

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

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

3.2.1 Adjacent-categories logit model structure with clustering

In this section, we use the adjacent-categories logit model to model clustered data via a finite mixture model. First, we start with column clustering, then followed by row clustering and then biclustering.

Column clustering For column clustering, we focus on column clusters to find which questions are similar and cluster them into groups. Columns are assumed a priori to come from any of $c = 1, \ldots, C$ column groups with probabilities $\kappa_1, \ldots, \kappa_C$. 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 $1 \leq C < m$ 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:

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

where β_c is the cth column-cluster effect.

From equation 2, we have:

$$\theta_{ick} = P[Y_{ij} = k|j \in c] = \frac{\exp\left[\mu_k^* + (k-1)\beta_c\right]}{\sum_{\ell=1}^{q} \exp\left[\mu_\ell^* + (l-1)\beta_c\right]} \tag{3}$$
$$i = 1, \ldots, n, \quad c = 1, \ldots, C, \quad k = 1, \ldots, q,$$

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

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

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

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

The model parameters $\mathbf{\Omega}$ contain $(\boldsymbol{\mu}, \boldsymbol{\beta}, \boldsymbol{\kappa})$. The expression above is referred to as the incomplete data likelihood, given that the cluster memberships are unknown.

Estimation by using EM algorithm

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

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

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

$$\sum_{c=1}^{C} X_{jc} = 1, \quad j = 1, \ldots, m,$$

The complete data log-likelihood is:

$$\ell_c(\mathbf{\Omega}|\boldsymbol{Y}, \boldsymbol{X}) = \sum_{j=1}^{m} \sum_{c=1}^{C} X_{jc} \log(\kappa_c) + \sum_{i=1}^{n} \sum_{j=1}^{m} \sum_{k=1}^{q} \sum_{c=1}^{C} X_{jc} I(y_{ij} = k) \log(\theta_{ick})$$

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

1. **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:

$$\widehat{x}_{jc}^{(t)} = \frac{\widehat{\kappa}_c^{(t-1)} \prod_{i=1}^{n} \prod_{k=1}^{q} (\widehat{\theta}_{ick}^{(t-1)})^{I(y_{ij}=k)}}{\sum_{g=1}^{C} [\widehat{\kappa}_g^{(t-1)} \prod_{i=1}^{n} \prod_{k=1}^{q} (\widehat{\theta}_{ick}^{(t-1)})^{I(y_{ij}=k)}]} \tag{4}$$

2. **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 $Q^{(t)}$ numerically 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 the parameter 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.

The formulas for the bi-clustering is as below:

$$\log\left(\frac{P[Y_{ij} = k|i \in r, j \in c]}{P[Y_{ij} = k-1|i \in r, j \in c]}\right) = \mu_k + \alpha_r + \beta_c,$$

$$i = 1, \ldots, n, \quad j = 1, \ldots m, \quad c = 1, \ldots, C, \quad k = 2, \ldots, q,$$

4. Simulation Study

In this section, we use a simulated dataset to illustrate the performance of the model for column clustering.

In particular, we generate a dataset use the following set up:

- $n = 30$ number of rows
- $m = 60$ number of columns
- $C = 2$ number of column clusters
- $\boldsymbol{\kappa} = (0.5, 0.5)$
- $\boldsymbol{\mu} = (\mu_1, \mu_2, \mu_3) = (0, 0, 0)$ intercepts
- $q = 3$ number of ordinal response categories
- $\boldsymbol{\beta} = (\beta_1, \beta_2) = (-0.3, 0.3)$ column cluster effects

Table 1: True parameters used to generate the simulated dataset

κ_1	κ_2	μ_2	μ_3	β_1	β_2
0.5	0.5	0	0	-0.3	0.3

Table 2: Estimated parameters value from 5 different starting points, and the calculated data log-likelihood value

	Estimated parameters						Estimated log-likelihood
	$\hat{\kappa_1}$	$\hat{\kappa_2}$	$\hat{\mu_2}$	$\hat{\mu_3}$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\ell}$
1	0.56	0.44	0.04	-0.02	-0.28	0.28	-1959.98
2	0.44	0.56	0.04	-0.02	0.28	-0.28	-1959.98
3	0.44	0.56	0.04	-0.02	0.28	-0.28	-1959.98
4	0.44	0.56	0.04	-0.02	0.28	-0.28	-1959.98
5	0.02	0.98	-0.67	-0.67	-0.66	0.66	-1970.19

We estimate the model parameters using the EM algorithm as detailed in Section 3. To avoid the risk of converging to local maxima, we used five random starting points as the initial values for the EM algorithm.

The results are illustrated in Table 2. The estimated result from the starting points 2-4 reached the same incomplete data log-likelihood and the same estimates for all the parameters, which gives us confidence that we have found the global maximum of the likelihood. For result from the starting point one, it suggested the label switching problem of mixture models for clustering. $\hat{\kappa}_1$ and $\hat{\kappa_1}$ are swapped with each other, $\hat{\beta}_1$ and $\hat{\beta}_2$ are swapped with each other. This is an inevitable problem of mixture models. Even though we can separate the clusters, the labels for clusters might change in the end, more discussion about this is in Stephens (2000).

5. Future research

The next step of our research will be:

- Running simulation study on a large scale for the model. The simulation result illustrated in Section 4 is only for one small dataset. We will run more simulation studies and use heat maps to evaluate our proposed model on row/column clustering and biclustering.
- Using information criteria to select the right number of clusters. Model selection method such as AIC, BIC and CLC will be used to make the model selection.
- Evaluate and compare our model with other clustering models using real data.

References

Aggarwal, C. C. and Reddy, C. K. (2013). *Data clustering: algorithms and applications*. CRC press.

Agresti, A. (1999). Modelling ordered categorical data: recent advances and future challenges. *Statistics in medicine*, 18(17-18):2191–2207.

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

Agresti, A. (2013). *Categorical Data Analysis*. John Wiley & Sons.

Akaike, H. (1987). Factor analysis and aic. In *Selected Papers of Hirotugu Akaike*, pages 371–386. Springer.

Ananth, C. V. and Kleinbaum, D. G. (1997). Regression models for ordinal responses: a review of methods and applications. *International journal of epidemiology*, 26(6):1323–1333.

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

Berg, A., Meyer, R., and Yu, J. (2004). Deviance information criterion for comparing stochastic volatility models. *Journal of Business & Economic Statistics*, 22(1):107–120.

Biernacki, C., Celeux, G., and Govaert, G. (2000). Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE transactions on pattern analysis and machine intelligence*, 22(7):719–725.

Bohte, S. M., La Poutré, H., and Kok, J. N. (2002). Unsupervised clustering with spiking neurons by sparse temporal coding and multilayer rbf networks. *IEEE Transactions on neural networks*, 13(2):426–435.

Bürkner, P.-C. and Vuorre, M. (2018). Ordinal regression models in psychology: A tutorial. *PsyArXiv. September*, 15.

Cameron, I. M., Scott, N. W., Adler, M., and Reid, I. C. (2014). A comparison of three methods of assessing differential item functioning (dif) in the hospital anxiety depression scale: ordinal logistic regression, rasch analysis and the mantel chi-square procedure. *Quality of life research*, 23(10):2883–2888.

Chan, Y. H. (2005). Biostatistics 305. multinomial logistic regression. *Singapore medical journal*, 46(6):259.

Christensen, R. (2006). *Log-linear models and logistic regression*. Springer Science & Business Media.

Clark, M. and Randal, J. A. (2011). *A first course in applied statistics: with applications in Biology, business and the social sciences*. Pearson.

Costilla, R., Liu, I., and Arnold, R. (2015). A bayesian model-based approach to estimate clusters in repeated ordinal data. *JSM Proceedings, biometrics section*, pages 545–556.

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.

Donneau, A.-F., Mauer, M., Lambert, P., Lesaffre, E., and Albert, A. (2015). Testing the proportional odds assumption in multiply imputed ordinal longitudinal data. *Journal of Applied Statistics*, 42(10):2257–2279.

Fernández, D., Arnold, R., and Pledger, S. (2016). Mixture-based clustering for the ordered stereotype model. *Computational Statistics & Data Analysis*, 93:46–75.

Fernández, D., Liu, I., Costilla, R., and Gu, P. Y. (2019). Assigning scores for ordered categorical responses. *Journal of Applied Statistics*. In press.

Franke, T. M., Ho, T., and Christie, C. A. (2012). The chi-square test: Often used and more often misinterpreted. *American Journal of Evaluation*, 33(3):448–458.

Govaert, G. and Nadif, M. (2013). *Co-clustering*. Wiley-IEEE Press.

Han, J., Pei, J., and Kamber, M. (2011). *Data mining: concepts and techniques*. Elsevier.

Hand, D. J. (2007). Principles of data mining. *Drug safety*, 30(7):621–622.

Jain, A. K., Murty, M. N., and Flynn, P. J. (1999). Data clustering: a review. *ACM computing surveys (CSUR)*, 31(3):264–323.

Jobson, J. (1992). Applied multivariate data analysis, categorical and multivariate methods.

Johnson, S. C. (1967). Hierarchical clustering schemes. *Psychometrika*, 32(3):241–254.

Kaufman, L. and Rousseeuw, P. J. (1990). Finding groups in data. an introduction to cluster analysis. *Wiley Series in Probability and Mathematical Statistics. Applied Probability and Statistics, New York: Wiley, 1990*, 1.

Kwak, C. and Clayton-Matthews, A. (2002). Multinomial logistic regression. *Nursing research*, 51(6):404–410.

Lanfranchi, M., Giannetto, C., and Zirilli, A. (2014). Analysis of demand determinants of high quality food products through the application of the cumulative proportional odds model. *Applied mathematical sciences*, 8(65-68):3297–3305.

Lewis, S., Foltynie, T., Blackwell, A., Robbins, T., Owen, A., and Barker, R. (2005). Heterogeneity of parkinson's disease in the early clinical stages using a data driven approach. *Journal of Neurology, Neurosurgery & Psychiatry*, 76(3):343–348.

Lin, G., Shearer, P. M., and Hauksson, E. (2007). Applying a three-dimensional velocity model, waveform cross correlation, and cluster analysis to locate southern california seismicity from 1981 to 2005. *Journal of Geophysical Research: Solid Earth*, 112(B12).

Liu, I. and Agresti, A. (2005). The anysis of ordered categorical data: An overview and a survey of recent developments. *Test*, 14(1):1–73.

Manly, B. F. (2005). Multivariate statistical methods a primer.

McCullagh, P. (1980). Regression models for ordinal data. *Journal of the royal statistical society. Series B (Methodological)*, pages 109–142.

McCune, B., Grace, J. B., and Urban, D. L. (2002). *Analysis of ecological communities*, volume 28. MjM software design Gleneden Beach, OR.

McHugh, M. L. (2013). The chi-square test of independence. *Biochemia medica: Biochemia medica*, 23(2):143–149.

McLachlan, G. and Krishnan, T. (2007). *The EM algorithm and extensions*, volume 382. John Wiley & Sons.

Melnykov, V., Maitra, R., et al. (2010). Finite mixture models and model-based clustering. *Statistics Surveys*, 4:80–116.

Murtagh, F. (1983). A survey of recent advances in hierarchical clustering algorithms. *The Computer Journal*, 26(4):354–359.

Murtagh, F. and Contreras, P. (2012). Algorithms for hierarchical clustering: an overview. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 2(1):86–97.

Pearson, K. (1894). Contributions to the mathematical theory of evolution. *Philosophical Transactions of the Royal Society of London. A*, 185:71–110.

Pledger, S. (2000). Unified maximum likelihood estimates for closed capture–recapture models using mixtures. *Biometrics*, 56(2):434–442.

Pledger, S. and Arnold, R. (2014). Multivariate methods using mixtures: Correspondence analysis, scaling and pattern-detection. *Computational Statistics & Data Analysis*, 71:241–261.

Schwarz, G. et al. (1978). Estimating the dimension of a model. *The annals of statistics*, 6(2):461–464.

Stephens, M. (2000). Dealing with label switching in mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62(4):795–809.

Strehl, A., Gupta, G. K., and Ghosh, J. (1999). Distance based clustering of association rules. In *Proceedings ANNIE*, volume 9, pages 759–764.

Ursino, M. and Gasparini, M. (2018). A new parsimonious model for ordinal longitudinal data with application to subjective evaluations of a gastrointestinal disease. *Statistical methods in medical research*, 27(5):1376–1393.

Witten, I. H., Frank, E., Hall, M. A., and Pal, C. J. (2016). *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann.