Introducing Spaced Mosaic Plots[☆]

D. Fernandez[a,*], R. Arnold[a], S. Pledger[a]

[a]*School of Mathematics, Statistics and Operations Research, Victoria University of Wellington, PO Box 600, Wellington 6140, New Zealand.*

Abstract

Recent research has developed a group of likelihood-based finite mixture models for a data matrix with ordinal data, establishing likelihood-based multivariate methods which applies fuzzy clustering via finite mixtures to the ordered stereotype model. There are many visualisation tools which depict reduction of dimensionality in matrices of ordinal data. This technical report introduces the *spaced mosaic plot* which is one new graphical tool for ordinal data when the ordinal stereotype model is used. It takes advantage of the fitted score parameters to determine the spacing between two adjacent ordinal categories. We develop a function in **R** and its documentation is presented. Finally, the description of a *spaced mosaic plot* is shown.

Keywords: Braun-Blanquet scale, Cluster analysis, Dimension reduction, Likert scale, Mosaic plot, Ordinal data, Stereotype model, Score parameters.

1. Introduction

1.1. Ordinal Data and Clustering

An ordinal variable is one with a categorical data scale which describes order, and where the distinct levels of such a variable differ in degree of dissimilarity more than in quality (Agresti, 2010). Categorical data analysis methods were first developed in the 1960s and 1970s (Bock and Jones (1968); Snell (1964)), including loglinear models and logistic regression (see the review by Liu and Agresti (2005)). An increasing interest in ordinal data has since produced the articles by Goodman (1979) and McCullagh (1980) on loglinear modelling relating to ordinal odds ratios, and logit modelling of cumulative probabilities respectively. Recently, new ordinal data analysis methods have been introduced such as the proportional odds model version of the cumulative logit model, and the stereotype model with ordinal scores (Agresti, 2010, Chap. 3 and 4) from which new lines of research have developed. In particular, the stereotype model is a paired-category logit model which is an alternative when the fit of cumulative logits and adjacent-categories logit models in their proportional odds version is poor. Anderson (1984) proposed this model as nested between the

[☆]This document is a collaborative effort.

*Corresponding author

Email addresses: daniel.fernandez@msor.vuw.ac.nz (D. Fernandez), richard.arnold@msor.vuw.ac.nz (R. Arnold), shirley.pledger@vuw.ac.nz (S. Pledger)

adjacent-categories logit model and the standard baseline-category logits model (see the review by Agresti (2002, chapter 6)).

In the research literature, multiple algorithms and techniques have been developed which deal with the clustering of ordinal data such as hierarchical clustering (Johnson, 1967; Kaufman and Rousseeuw, 1990), association analysis (Manly, 2005) and partition optimization methods such as the k-means clustering algorithm (Jobson, 1992; Lewis et al., 2003; McCune and Grace, 2002). There has been research on cluster analysis for ordinal data based on latent class models (see Agresti and Lang (1993); Moustaki (2000); Vermunt (2001); DeSantis et al. (2008); Breen and Luijkx (2010) and the review by Agresti (2010, Section 10.1)). There are a number of clustering methods based on mathematical techniques such as distance metrics (Everitt et al., 2001), association indices (Wu et al. (2008); Chen et al. (2011)), matrix decomposition and eigenvalues (Quinn and Keough, 2002; Manly, 2005; Wu et al., 2007). A likelihood-based model approach using finite mixtures to define a fuzzy clustering where the components are based on the ordered stereotype model has been developed in Fernandez et al. (2014). All the analysis and visualisations shown in this report are based on this latter approach.

1.2. Data and Ordered Stereotype Model Including Row Clustering

For a set of m ordinal response variables each with q categories measured on a set of n units, the data can be represented by a $n \times m$ matrix Y where, for instance, the n rows represent the subjects of the study and the m columns are the different questions in a particular questionnaire. Although the number of categories might be different, we assume the same q for all such questions. If each answer is a selection from q ordered categories (e.g. strongly agree, agree, neutral, disagree, strongly disagree), then

$$y_{ij} \in \{1, \ldots, q\}, \qquad i = 1, \ldots, n, \qquad j = 1, \ldots, m.$$

The ordered stereotype model (Anderson, 1984) for the probability that y_{ij} takes the category k and including row clustering is characterized by the following log odds

$$\log\left(\frac{P\left[y_{ij} = k \mid i \in r\right]}{P\left[y_{ij} = 1 \mid i \in r\right]}\right) = \mu_k + \phi_k(\alpha_r + \beta_j + \gamma_{rj}),$$
$$k = 2, \ldots, q, \qquad r = 1, \ldots, R, \qquad j = 1, \ldots, m.$$

where $R \leq n$ is the number of row groups, $\{\alpha_1, \ldots, \alpha_R\}$ and $\{\beta_1, \ldots, \beta_m\}$ as the sets of parameters quantifying the main effects of the R row clusters and m columns respectively, and the set $\{\gamma_{r1}, \ldots, \gamma_{Rm}\}$ are the associations between the different row groups and columns.

The inclusion of the following monotone increasing constraint

$$0 = \phi_1 \leq \phi_2 \leq \cdots \leq \phi_q = 1 \tag{1}$$

ensures the variable response Y is ordinal (see Anderson (1984)).

1.3. Mosaic Plots

There are a number of visualisation tools which can depict reduction of dimensionality in matrices of ordinal data such as multidimensional scaling

and correspondence analysis plots (see e.g. Manly (2005); Quinn and Keough (2002)). In this technical report, we introduce a new graphical tool for ordinal data based on mosaic plots. The mosaic plot was developed by Hartigan and Kleiner (1981) and refined by Friendly (1991). It is a graphical method for visualizing data from two qualitative variables which gives an overview of the data, makes it possible to recognize relationships and show the cross-sectional distribution of different variables. In our case, we consider the ordinal response variable and the number of fitted clusters in the data as those two qualitative variables. For instance, an ordinal data matrix following a four-category Likert scale ("Disagree","No Opinion","Agree","Strongly Agree") and with three row clusters is depicted as a mosaic plot in Figure 1. The mosaic plot is divided in 3 horizontal bands over the y-axis (one for each row cluster) and 4 vertical bands over the x-axis (one for each ordinal response category). The areas represent the frequencies as explained in Section 2.1.

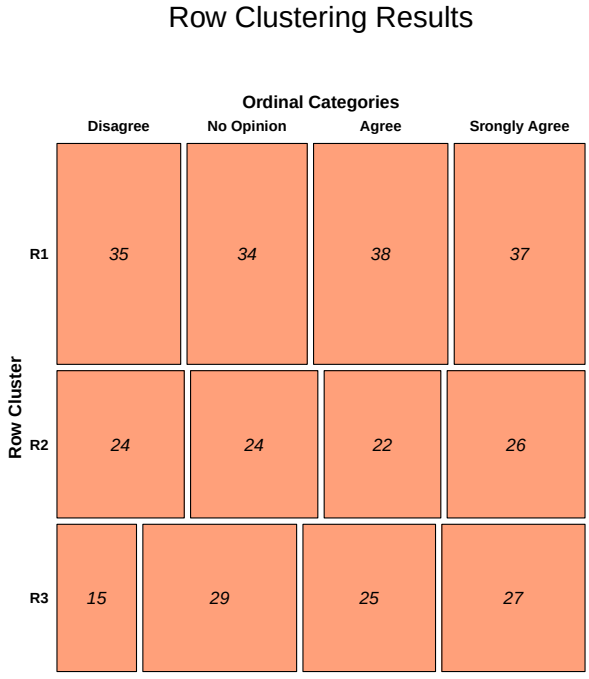

Figure 1: Mosaic plot including row clustering structure $R = 3$ and 4 ordinal categories.

One improvement we can incorporate in a mosaic plot due to the use of the ordinal stereotype model is the estimation of score parameters $\{\phi_k\}$. Those parameters determine the space between two adjacent ordinal categories based on the data (see Anderson (1984); Agresti (2010) for more detail). For instance, the space between "Disagree" and "No Opinion" can be higher than the space between "Agree" and "Strongly Agree". The inclusion of space within a regular mosaic plot generates an upgraded graph with more information which we called the *spaced mosaic plot*. This report is structured as follows. Section 2 describes this new visualisation tool and Section 3 presents the documentation of a **R** function (R Development Core Team (2010)) we develop to generate *spaced*

mosaic plots.

2. Spaced Mosaic Plots

2.1. Description

We use an ordinal real data set from community ecology as an example to illustrate the *spaced mosaic plots* in the case of clustering the rows. The data set is regarding the distribution of 77 different angiosperms along 30 different sites. The study was carried out at Bola Heights in Royal National Park, about 37 km south-west of Sydney and 200 meters above sea level (see Tozer and Bradstock (2002) for more detail). The goal of this vegetation survey data is to group species observations to derive community types. The 2310 ordinal observations consist in level of angiosperm species presence at each site in combination with the percentage of coverage within the site. Thus, the ordinal scale follows a modified Braun-Blanquet scale (Westhoff and van der Maarel E., 1978) as follows:

$$\begin{cases} 0 & \text{no data recorded} \\ 1 & \text{one/a few individuals and less than 5\% cover} \\ 2 & \text{uncommon and less than 5\% cover} \\ 3 & \text{common/very abundant and less than 5\% cover} \\ & \text{or coverage higher than 5\%.} \end{cases}$$

After fitting a complete set of models and comparing them by using the Akaike information criteria (AIC, Akaike (1973)), the selected model was the stereotype model version including row clustering with $R = 4$ row groups. Figures 2-4 show the results for this example. Firstly, Figure 2 depicts the raw data without including row clustering, Figure 3 depicts the data including row clustering structure and Figure 4 depicts the data including both row clustering structure and fitted spacing between ordinal categories. A comprehensive description for each Figure is as follows:

- Figure 2 shows the overall distribution of ordinal responses over all the cells, ignoring rows and columns. Thus, area is equivalent to frequency. The ordinal category 0 response is most common by far, and ordinal category 3 the least.

- Figure 3 shows the clustering in the rows, putting each row into one of four clusters according to the distribution of ordinal responses across the columns of the original data matrix. This divides the plot into four horizontal bands, one for each row group. The height of each band is proportional to the number of rows in the group. Therefore, we can see that row groups 1 and 4 ($R1$ and $R4$) are the largest, much larger than row groups 2 and 3 ($R2$ and $R3$). Within each row group we represent the frequencies of the four ordinal responses by the area of each block. Members of row group 4 show a strong preference for ordinal response categories 0 and 1, and rarely respond at category 2 or 3. Contrast this with row group 2, which has 50% of its responses at ordinal category 3. Note that this diagram does not in any way show the ordering across the columns – it is simply a pooling of frequencies of all of the responses for individuals in the same row group.

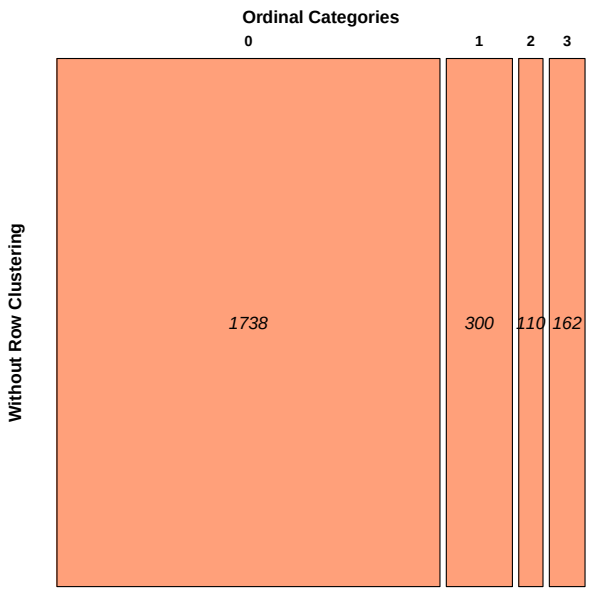

Figure 2: Mosaic plot without spacing or row clustering.

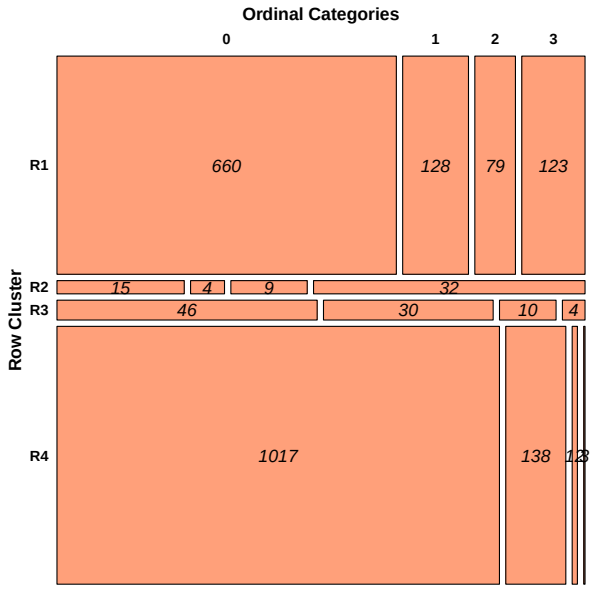

Figure 3: Mosaic plot including row clustering structure $R = 4$.

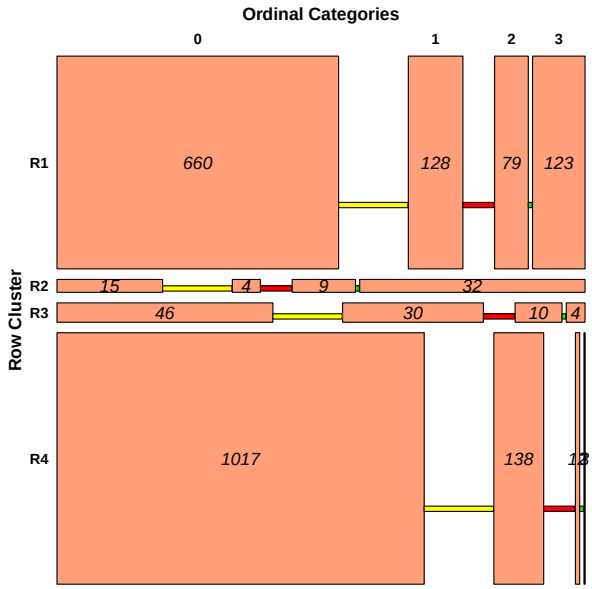

Figure 4: Mosaic plot with spacing for the row clustering model $R = 4$.

- Figure 4 takes the bands and blocks from Figure 3, but separates them out to indicate the numerical spacing between the response categories that the model has identified. Since each ordinal response category is associated with a score parameter ϕ_k $(k = 0, \ldots, 3)$ the spacing between these ϕ_k values shows us how similar or different adjacent categories are. In this model the fitted score parameters are $\phi_0 = 0$, $\phi_1 = 0.66$, $\phi_2 = 0.96$ and $\phi_3 = 1$ (the end points being fixed at 0 and 1). The distance between category 0 and category 1 (0.66) is much greater than that between categories 1 and 2 (0.30) or categories 2 and 3 (0.04). In each row group band, we have inserted space and a different color block proportional to these differences between two adjacent ordinal categories. For instance, a yellow block of the same width has been inserted between categories 0 and 1 in each band. Note that these blocks do not line up vertically with each other between bands due to the differing counts at category 0, nevertheless the color blocks are the same width. In so doing, we can immediately see that categories 2 and 3 are close to each other, without needing to refer to the numerical values of ϕ_k. Inspection of Figure 4 might lead us to conclude that categories 2 and 3 are so similar that these two groups might just as well be collapsed into a single group.

2.2. Outlining Space Mosaic Plots

The main features from a *spaced mosaic plot* are:

- x-axis represents the ordinal categories in the data and y-axis represents the row clustering obtained by our methodology. The data frequency of each combination in terms of ordinal category and row cluster is shown in each box.

- The more width a specific box has, the higher the proportion of data allocated in the related ordinal category.
- The more height a specific box has, the higher the proportion of data classified in that particular row group. For example, the bottom left box corresponding to the row cluster 4 (R4) and the ordinal category 0 is the widest and the highest because it contains 1017 combinations of species-samples over 2310 (44%). None of the other boxes have higher frequencies.
- Each box area is proportional to the frequency in the corresponding row group and ordinal category. For instance, the box located on the top right depicts the proportional number of species (angiosperms) allocated in the first cluster ($R1$) and with Braun-Blanquet scale 3.
- The spacing between two levels of the ordinal categories (x-axis) is dictated by the data. It represents the proximity of two adjacent ordinal categories. Determining the distance among ordinal categories is a key advantage of the stereotype model in comparison with other similar methods.

The *spaced mosaic plots* allow us to see the at once the relative sizes of the row groups, the relative frequencies of the different response categories within each row group and the differences between the levels of the response categories.

3. R function

In this section we describe the R function to fit a spaced mosaic plot. This function will be included within a R package. In the meantime, you can e-mail the corresponding author (D. Fernandez - `daniel.fernandez@msor.vuw.ac.nz`) to obtain this function.

The description of the R function we have developed is:

`spaced.mosaic.plot` *Draw spaced mosaic plots for clustering ordinal data*

Description

The function `spaced.mosaic.plot` computes a spaced mosaic plot of a given ordinal data, clustering structure and fitted score parameters.

Usage

```
spaced.mosaic.plot(y.mat, phi, R, ClusterRowY, labels=NA)
```

Arguments

`y.mat`	a matrix object containing the ordinal dataset.
`phi`	a vector with the fitted score parameters ($\{\phi_k\}$) from the ordinal stereotype model.
`R`	an integer specifying the number of row clusters.
`ClusterRowY`	a vector with the allocated cluster allocated to each row.
`labels`	(optional) a 4 dimension list where:

`categ`	contains the labels for the ordinal categories.
`cluster`	contains the labels for the clusters.
`row`	contains the labels for the data rows.
`col`	contains the labels for the data columns.

Value

The function return a data frequency table with `R` rows and one column for each category. In addition, three `pdf` files are generated in the working directory with the overall distribution (MosaicPlot_withoutRowCluster.pdf), the row clustering structure (MosaicPlot_R=*R*.pdf) and the inclusion of the space between adjacent ordinal categories (MosaicPlot_SPACING_R=*R*.pdf).

Author(s)

Daniel Fernandez

See Also

mosaicplot

Example

```
library(grid)
library(vcd)
#Score parameters
phi <- c(0,0.5,0.7,1)

#Generation of simulated data
q <- length(phi)
n <- 28
m <- 12
R <- 3
labels <- list(categ=c("Disagree", "No Opinion", "Agree","Strongly Agree"),
               cluster=paste("R",seq(1,R,1),sep=""),
               row=paste("r",seq(1,n,1),sep=""),
               col=paste("c",seq(1,m,1),sep=""))
y.mat <- matrix(NA,n,m)
for(i in 1:n) for (j in 1:m) y.mat[i,j] <- sample(1:q,1,replace=TRUE)

ClusterRowY <- array(NA,n)
for (i in 1:n) ClusterRowY[i] <- sample(1:R,1,replace=TRUE)
rownames(ClusterRowY) <- labels$row

#Generate spaced mosaic plot
spaced.mosaic.plot(y.mat, phi, R, ClusterRowY, labels)
```

```
  Col
Row  Disagree No Opinion Agree Strongly Agree
  R1       31         41    29             31
  R2       34         32    27             27
  R3       21         21    21             21
```

Acknowledgments

The authors are sincerely grateful to Mark Tozer for permission to use the angiosperm data set.

References

Agresti, A., 2002. Categorical Data Analysis. Wiley Series in Probability and Statistics. 2nd ed., Wiley-Interscience.

Agresti, A., 2010. Analysis of Ordinal Categorical Data. Wiley Series in Probability and Statistics. 2nd ed., Wiley.

Agresti, A., Lang, J.B., 1993. Quasi-symmetric latent class models, with application to rater agreement. Biometrics 49, 131–139.

Akaike, H., 1973. Information theory and an extension of the maximum likelihood principle, in: Petrov, B.N., Csaki, F. (Eds.), 2nd International Symposium on Information Theory, pp. 267–281.

Anderson, J.A., 1984. Regression and ordered categorical variables. J. R. Statist. Soc. B 46, 1–30.

Bock, R., Jones, L., 1968. The measurement and prediction of judgment and choice. Holden-Day series in psychology, Holden-Day.

Breen, R., Luijkx, R., 2010. Assessing proportionality in the proportional odds model for ordinal logistic regression. Sociological Methods and Research 39, 3–24.

Chen, L.C., Yu, P.S., Tseng, V.S., 2011. A weighted fuzzy-based biclustering method for gene expression data. International Journal of Data Mining and Bioinformatics 5, 89–109.

DeSantis, S.M., Houseman, E.A., Coull, B. A. ad Stemmer-Rachamimov, A., Betensky, R.A., 2008. A penalized latent class model for ordinal data. Biostatistics 9, 249–262.

Everitt, B.S., Leese, M., Landau, S., 2001. Cluster Analysis. 4th ed., Hodder Arnold Publication.

Fernandez, D., Arnold, R., Pledger, S., 2014. Fuzzy clustering for the ordered stereotype model via finite mixtures. Computational Statistics and Data Analysis. *Under review.*

Friendly, M., 1991. Mosaic Displays for Multiway Contingency Tables. Technical Report 195. New York University Department of Psychology Reports.

Goodman, L.A., 1979. Simple models for the analysis of association in cross-classifications having ordered categories. Journal of the American Statistical Association 74, 537–552.

Hartigan, J.A., Kleiner, B., 1981. Mosaics for contingency tables. Proceedings of the 13th Symposium on the Interface between Computer Sciencies and Statistics , 268–273.

Jobson, J.D., 1992. Applied Multivariate Data Analysis: Categorical and Multivariate Methods. Springer Texts in Statistics, Springer.

Johnson, S.C., 1967. Hierarchical clustering schemes. Psychometrika 2, 241–254.

Kaufman, L., Rousseeuw, P.J., 1990. Finding Groups in Data an Introduction to Cluster Analysis. Wiley, New York.

Lewis, S.J.G., Foltynie, T., Blackwell, A.D., Robbins, T.W., Owen, A.M., Barker, R.A., 2003. Heterogeneity of parkinson's disease in the early clinical stages using a data driven approach. Journal of Neurology, Neurosurgery and Psychiatry 76, 343–348.

Liu, I., Agresti, A., 2005. The analysis of ordered categorical data: An overview and a survey of recent developments. TEST: An Official Journal of the Spanish Society of Statistics and Operations Research 14, 1–73.

Manly, B.F.J., 2005. Multivariate Statistical Methods: a Primer. Chapman & Hall/CRC Press, Boca Raton, FL.

McCullagh, P., 1980. Regression models for ordinal data. Journal of the Royal Statistical Society 42, 109–142.

McCune, B., Grace, J.B., 2002. Analysis of Ecological Communities. volume 28. MjM Software Design.

Moustaki, I., 2000. A latent variable model for ordinal variables. Applied Psychological Measurement , 211–233.

Quinn, G.P., Keough, M.J., 2002. Experimental Design and Data Analysis for Biologists. Cambridge University Press.

R Development Core Team, 2010. R: A Language and Environment for Statistical Computing. Vienna, Austria. URL: `http://www.R-project.org`. ISBN 3-900051-07-0.

Snell, E.J., 1964. A scaling procedure for ordered categorical data. Biometrics 20, 592–607.

Tozer, M.G., Bradstock, R.A., 2002. Fire-mediated effects of overstorey on plant species diversity and abundance in an eastern australian heath. Plant Ecology , 213–223.

Vermunt, J.K., 2001. The use of restricted latent class models for defining and testing nonparametric and parametric item response theory models. Applied Psychological Measurement 25, 283–294.

Westhoff, V., van der Maarel E., 1978. The braun-blanquet approach, in: H., W.R. (Ed.), Classification of Plant Communities. Junk, The Hague, pp. 287–328.

Wu, H.M., Tzeng, S., Chen, C.H., 2007. Matrix visualization. Handbook of Data Visualization , 681–708.

Wu, X., Kumar, V., Quinlan, J.R., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G.J., Ng, A., Liu, B., Yu, P.S., Zhou, Z.H., Steinbach, M., Hand, D.J., Steinberg, D., 2008. Top 10 algorithms in data mining. Knowledge and Information Systems 14, 1–37.