

Agglomerative Clustering of Rankings Data, with an Application to Prison Rodeo Events

Christopher Zorn[†]

July 3, 2003

Abstract

This paper considers the problem of assessing item similarity on the basis of rankings data, that is, data on ordinal outcomes. I discuss a modification to the standard dissimilarity measure used in agglomerative clustering which addresses the ordinal nature of ranking data. I then apply this alternative to cluster nine events comprising the Angola, Louisiana prison rodeo.

Keywords: Cluster analysis, ordinal data, classification, rankings.

[†]Department of Political Science, Emory University, Atlanta, GA 30322. E-mail address: czorn@emory.edu

1 Introduction

Cluster analysis provides a useful tool for deductively evaluating similarities and dissimilarities among observations. Here I discuss a modification to standard agglomerative clustering methods due to Kaufman and Rousseeuw (1990) for use with ordinal data, including rankings. The modification transforms ordinal data to the interval level prior to calculating the dissimilarity matrix. I demonstrate this modification through a cluster analysis of data on nine events held at the four Angola, Louisiana prison rodeos during October, 2002.

2 Agglomerative Clustering of Rankings Data

Consider data on K variables X_1, X_2, \dots, X_k with N objects $i = 1, 2, \dots, N$ for each variable. Suppose we are interested in making use of the information in the K variables to evaluate the similarity of the N observations in our data. That is, we wish to know how similar or dissimilar each of our N items are from one another. A common approach is to use the method of agglomerative clustering (e.g. Hartigan 1975; Kaufman and Rousseeuw 1990). In agglomerative clustering, the two clusters are determined by iteratively combining objects which are more “alike.”

Models for agglomerative clustering rely on a *dissimilarity matrix*, typically defined as:

$$\mathbf{D} = \begin{pmatrix} 0 & & & & & \\ d_{21} & 0 & & & & \\ d_{31} & d_{32} & 0 & & & \\ \vdots & \vdots & \vdots & \ddots & & \\ d_{N1} & d_{N2} & \dots & \dots & 0 & \end{pmatrix} \quad (1)$$

where \mathbf{D} is an $\frac{N(N+1)}{2}$ lower-half matrix with the various d_{ij} corresponding to covariate-based measures of dissimilarity between the two objects i and j . In the case of interval- or ratio-level variables \mathbf{X} , the elements of \mathbf{D} are typically defined in terms of a “distance measure”; two of the most common are Euclidean distance, defined as:

$$d_{ij} = \sqrt{\sum_{k=1}^K (X_{ik} - X_{jk})^2} \quad (2)$$

and “city-block” (or “Manhattan”) distance:

$$d_{ij} = \sum_{k=1}^K |X_{ik} - X_{jk}| \quad (3)$$

Because (2) and (3) depend strongly on the scale of the variables, common practice is to standardize the variables (typically using their mean absolute deviation) prior to calculating dissimilarities.

Once \mathbf{D} is known, agglomerative clustering occurs by iteratively combining those pairs of clusters with the lowest levels of dissimilarity, and then recalculating the \mathbf{D} matrix. Intercluster dissimilarities are most often calculated according to the group-average method, whereby the dissimilarity between clusters A and B is:

$$d_{AB} = \frac{1}{|A||B|} \sum_{i \in A, j \in B} d_{ij} \quad (4)$$

and d_{ij} are calculated as above.

Calculation of d_{ij} according to (2) or (3) assumes (at least) interval-level measurement of the \mathbf{X} s. In some circumstances, however, data on \mathbf{X} may be measured at the ordinal level only. For example, much consumer preference data takes the form of ordinal ratings (e.g., on a 0–10 scale) or rankings of products. Under such circumstances, the validity of the interval/ratio assumption can be severely tested.

To address this issue of ordinality, Kaufman and Rousseeuw (1990) suggest the following two-step modification to the calculation of d_{ij} . For an ordinal variable X_k with, at most M_k ordered categories, first replace X_{ik} with its rank $r_{ik} \in \{1, 2, \dots, M_k\}$. Then, transform r_{ik} to the unit interval by calculating:

$$Z_{ik} = \frac{r_{ik} - 1}{M_k - 1} \quad (5)$$

Equations (2) or (3) can then be applied, replacing X_{ik} with Z_{ik} . This transformation retains the (ranking) information present in the ordinal variables, but ameliorates some of the problems that otherwise attend clustering of ordinal measures.

3 An Application to the Categorization of Prison Rodeo Events

I next consider an application of the model discussed above. Every Sunday in October, the Louisiana State Penitentiary at Angola conducts its prison rodeo, at which inmates compete in standard and non-standard rodeo events.¹ All together, the rodeo comprises nine events, which take place in the following order (“Angola Rodeo Home Page” 2002):

- **Bust Out.** All six chutes open simultaneously, releasing six angry bulls, with temporarily attached inmate cowboys. The last man to remain on the bull wins the event.
- **Bareback Riding.** Riders are expected to keep one hand in the air, and must stay on the horse for eight seconds to qualify.
- **Wild Horse Racing.** Six wild horses are simultaneously released into the arena with short ropes dragging behind them. Three-man teams attempt to grab the ropes and hold the horse long enough for a team member to mount. The first team to cross the finish line while still on top of the horse is the winner.
- **Bull-Dogging.** The animal is placed in a chute, with two cowboys positioned just outside the chute. Their job is to wrestle the animal to the ground as quickly as possible.
- **Buddy Pick-Up.** This event requires one man on a horse (riding bareback) to navigate the length of the arena, pick up another inmate who is standing on a barrel, and race back to the finish line.
- **Wild Cow Milking.** Teams of inmate cowboys chase the animals around the arena trying to extract a little milk. The first team to bring milk to the judge wins the prize.
- **Bull Riding.** Inexperienced inmates sit on top of a 2,000 pound Brahma bull. To be eligible for the coveted “All-Around Cowboy” title, a contestant must successfully complete the ride (6 seconds).

¹For insightful perspectives on the Angola prison rodeo, see Bergner (1998) and Gill (2002).

- **Convict Poker.** Four inmate cowboys sit at a table in the middle of the arena playing a friendly game of poker. Suddenly, a wild bull is released with the sole purpose of unseating the poker players. The last man remaining seated is the winner.
- **Guts & Glory.** A poker chip is tied to the meanest, toughest Brahma bull available. The object is to get close enough to the bull in order to snatch the chit.

The substantive question of interest is the extent to which certain events are “like” others. One might expect these events to cluster in any number of ways: by livestock (i.e., broncos vs. bulls), chronologically (by the order in which they occur), or by the level of skill or “toughness” required to compete in the event.

To address this issue, I consider data on 69 inmates, each of which participated in at least one of the four Angola prison rodeos held in October 2002. For each inmate, I record their rank (1st, 2nd, or 3rd) in each event, including team events.² Inmates failing to place in the event score a zero, those placing third a one, etc. Note that, because of the transformation in equation (5), arbitrarily coding non-placing entrants as zero has no effect on the final results. I use the data on each of the 276 measurements (69 inmates \times four rodeos) as covariates in the analysis of the nine events. That is, data on inmate rankings are used to assess the extent to which events “cluster”.

The dendrogram of the agglomerative cluster analysis, using a Euclidean distance measure and the group–average linkages described in (4), is presented in Figure 1.³ The agglomerative coefficient (Rousseeuw 1986) for these data is 0.38, suggesting that, overall, few clear clusters exist among the Anglola data. Examining the dendrogram itself, the clear conclusion is that “toughness matters”; risky events which require high levels of bravery and/or tolerance for physical abuse tend to cluster together. A clear grouping emerges among the bull–riding, cowboy poker, “Guts & Glory” and “Bust–Out” events, all of which require both high risk and (at least potentially) punishing physical abuse. Conversely, relatively safe events such as Wild Cow Milking are among the last to be incorporated. At the same time, we see little indication that the organizing principle is (e.g.) chronology.

²Ties are recorded as placing 1.5, 2.5 and 3.5 for first, second and third, respectively.

³Analysis was performed using the *agnes* routine in *S–Plus 6.1* (MathSoft 2001).

4 Conclusion

Agglomerative clustering techniques present a powerful tool for deductive data analysis. Such tools, however, require attention to issues of measurement, particularly in cases where substantive results may be altered as a result of variations in levels of covariate measurement. This paper has illustrated a modification to standard approaches for agglomerative cluster analysis which accounts for ordinally-measured covariates, such as might be found in studies of consumer preferences. Future work might consider the extent to which such modifications alter substantive findings under more controlled conditions (e.g., in simulations) as well as addressing other measurement issues surrounding the application of cluster analysis techniques.

5 References

Angola Rodeo Home Page, <http://www.angolarodeo.com/>. Visited May 9, 2003.

BERGNER, D. (1998), *God of the Rodeo: the Search for Hope, Faith and a Six-Second Ride in Louisiana's Angola Prison* (New York: Crown Publishers).

GILL, J. (2002), *Bayesian Methods: A Social and Behavioral Sciences Approach* (Boca Raton: Chapman & Hall).

HARTIGAN, J.A. (1975), *Clustering Algorithms* (New York: John Wiley & Sons, Inc.).

KAUFMAN, L. and ROUSSEEUW, P.J. (1990), *Finding Groups in Data: An Introduction to Cluster Analysis* (New York: John Wiley & Sons, Inc.).

MATHSOFT. (2001), *S-Plus Guide to Statistics* (Seattle, WA: MathSoft).

ROUSSEEUW, P. J. (1986), A Visual Display for Hierarchical Classification, in E. DIDAY, Y. ESCOUFIER, L. LEBART, J. PAGES, Y. SCHEKTMAN and R. TOMASSONE (eds.) *Data Analysis and Informatics*, vol. 4 (Amsterdam: North-Holland), 743-48.

Figure 1: Dendrogram for the 2022 Angola Prison Rodeo Data

