Homogeneity Analysis of Durant Bend Sherds

Jan de Leeuw and C. Roger Nance

Version February 24, 2018

Abstract

This paper is a non-technical and mostly graphical introduction to *Homogeneity Analysis*, also known as *Multiple Correspondence Analysis*. It is meant as an explanation and justification of a non-standard application of *Correspondence Analysis* to an example from archeology.

Contents

#Data

The data for this paper are 5888 pot sherds excavated in 1971 and 1975 from several sites in the vicinity of Durant Bend, Dalla County, Alabama. Each sherd was labeled by site and by the depth level in the excavation. In addition each sherd was classified usng three binary variables: design (check stamped vs plain), paste color (dark vs light), and thickness (thick vs thin). For the details we refer to C. Roger Nance (1976) and C. R. Nance and De Leeuw (2018). In the following table the sherds are aggregated over sites/depths.

```
##          CS Plain Dark Light Thin Thick
## Ds73S1  118   555  425   248  157   516
## Ds73S2  171   302  304   169  126   347
## Ds73S3   83   156  163    76   67   172
## Ds73S4   36    51   59    28   24    63
## Ds73S5+  25    50   46    29   19    56
## Ds73NAM 203   964  656   511  183   984
## Ds73NUM 196   389  342   243  100   485
## Ds73NLM 164   199  229   134   64   299
## Ds73NBM  74    85  102    57   27   132
## Ds791    11   292  170   133  125   178
## Ds792    17   247  163   101  116   148
## Ds793    24   163  100    87   62   125
## Ds794+    3    39   25    17   18    24
## Au1131    7    34   30    11   30    11
```

```
## Au1132   14   40  38   16  32   22
## Au1133+  15   18  25    8  22   11
## Ds98PZ   20  219 166   73 183   56
## DBUCC    11  110  36   85  85   36
## DBLCC     9   45  32   22  47    7
## DBBCC1   13   20  19   14  29    4
## DBBCC2   28   34  45   17  49   13
## DBBCC3   90   79 132   37 147   22
## DBBCC4   83   67 106   44 111   39
## DBBCC5+  66   45  81   30  81   30
## DS971     1   75  60   16  49   27
## DS972     7   67  47   27  51   23
## DS973+    9   45  43   11  38   16
```

#Homogeneity Analysis

The technique we will use to analyse the Durent Bend data is Homogeneity Analysis (Gifi (1990)), which is more widely known as *Multiple Correspondence Analysis* (Greenacre (1984), Greenacre and Blasius (2006)). We give a graphical introduction to Homogeneity Analysis, without using formulas.

Suppose we have m categorical *variables*, and that variable j has k_j *categories* $(j = 1, \cdots, m)$. The m variables *categorize* or *measure* n *objects*. Variable j partitions the set of n objects into k_j subsets, one subset for each category. Before we get to the analysis of the Durant Bend sherds, we will illustrate the main concepts of our approach with a small example in which three variables partition ten objects. The first two variables have three categories, the last one has two categories.

```
##    first second third
## 01     a      p     u
## 02     b      q     v
## 03     a      r     v
## 04     a      p     u
## 05     b      p     v
## 06     c      p     v
## 07     a      p     u
## 08     a      p     v
## 09     c      p     v
## 10     a      p     v
```

In Homogeneity Analysis we aim to make a *joint plot* of the objects and the categories of the variables. Joint plots are also known as *biplots* (Gower and Hand (1996)). In a joint plot both objects and categories are represented as points in a low-dimensional space, usually the plane, in such a way that the relations in the data are represented as precisely as possible in the plot. We will specify what we mean by “as precisely as possible” below. Homogeneity

Analysis is defined by defining a measure of the loss of information in a certain way, and then choosing the representation that minimizes that loss.

The n points in the plan, or more generally in p-dimensional space, representing the objects (sherds) are collected in a matrix of *object scores*. The k_j points representing the categories of variable j are in a matrix of *category quantifications* for variabe j. For each variable we can make a *graph plot* in which each of the n object scores is connected by a straight line to the quantification of the category that this object falls in. Thus there is one line departing from each object point, while the number of lines arriving at a category point is equal to the number of objects in the category. One graph plot has n lines, all graph plots together have $n \times m$ lines.

If all these lines have length zero, then all objects coincide with "their" categories for that variable, and thus we have reproduced the data exactly. If there is more than one variable, however, we cannot expect to have such a perfect representation, because objects which are together in a category for one variable may not be together for another variable.

Homogeneity Analysis is defined as the technique that produces a joint plot of objects and category quantification in such a way that the total length of all $n \times m$ lines in the m graph plots is a small as possible. Some qualifications are needed, however. We actually minimizes the sum of the squared length of the lines, for the same reason that we use the squares of the residuals in a regression analysis. It simplifies the mathematics and the computation to use squared distances. Secondly, we could trivially gain our objective of minimizing line length by collapsing all object scores and category quantifications in a single point, which makes our loss function equal to zero, but is useless in representing or reproducing the data. Thus we need some form of *normalization* to prevent this trivial solution from happening. In Homogeneity Analysis we require the columns of the object score matrix add up to zero, have sum of squares equal to one, and are uncorrelated.

Let's illustrate this with our small example. We start with a completely aribitrary *initial configuration*. The ten objects are placed at equal distances on a circle, and the categories for each of the variables are on the horizontal axis. This leads to the first three graph plots, which we have superimposed to get the fourth plot with $n \times m = 30$ lines.

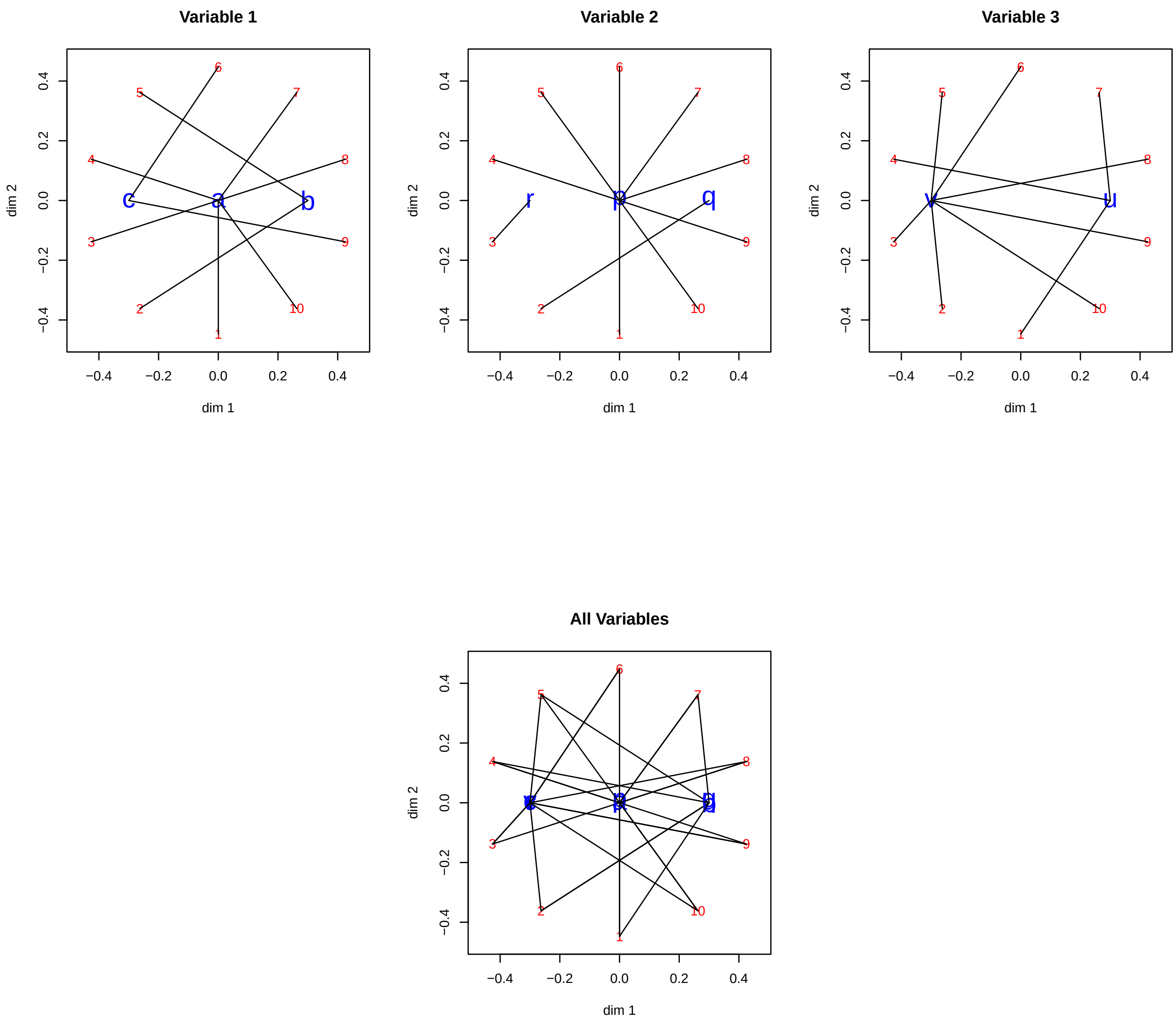

Figure 1: Graph Plots Initial Configuration, Small Example

For this arbitrary initial configuration the loss, i.e. the sum of squares of the line lengths, or the sum of the squred distances between objects and the categories they fall in, is equal to 8.1081098.

#Reciprocal Averaging

In Homogeneity Analysis we minimize loss by what is known as *reciprocal averaging* or *alternating least squares*. We alternate two substeps. The first substep improves the category quantifications for a given set of object scores, the second substep improves and normalizes the object scores for a given set of category quantifications, namely those we have just computed in the first substep. Taken together these two substeps are an *iteration*. So each iteration starts with object scores and category quantifications and uses its two substeps to improve both. Each of the two substeps decreases the loss functions, i.e. the total squared length of the lines in the graph plots.

The two substeps are both very simple. Let's look at the first one. We compute optimal category quantifications for given object scores by taking the averages (or *centroids*) of the objects scores in each of the categories. The corresponding graph plots are

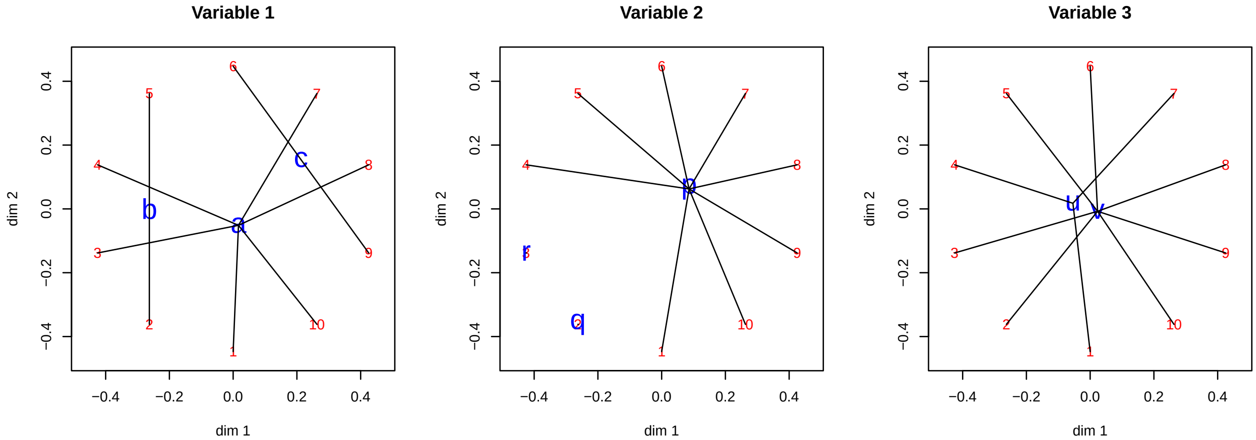

Figure 2: Graph Plots, Iteration 1, substep 1, Small Example

and the loss has decreased to 5.2016654. Note that we have not improved the object scores yet, so they are still their initial configuration, equally spaced on a circle. Also note, in variable 2 for instance, that category quantifications coincide with object scores, and thus contribute zero to the loss, if the object is the only observation in the category. In addition, because category quantifications are averages of objects points, they are in the convex hull of the object points, which means in this figure that they are within the circle. Averaging objects points makes the category quantifications move closer to the origin.

The second substep improves the object scores, while keeping the category quantifications in the locations we have just computed in the first substep. The second substaep has itself two substeps, say $2A$ and $2B$. In the substep $2A$ the score of an object for given category quantifications is computed as the average or centroid of the m category quantifications the object is in.

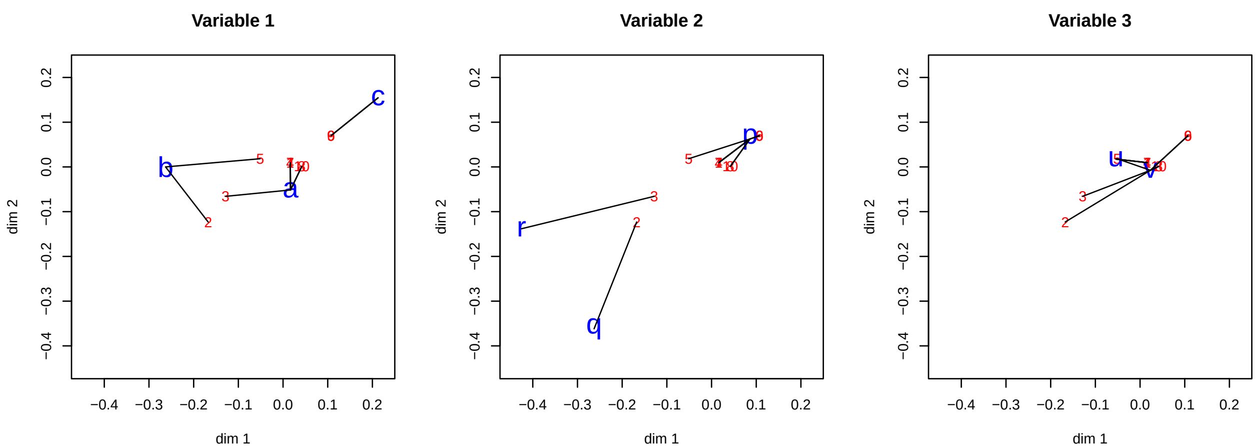

Figure 3: Graph Plots, Iteration 1, substep 2A, Small Example

The loss function is down all the way to 0.4852424. This is not a proper loss value, however, because the object scores are no longer centered, standardized, and uncorrelated, and that was a Homogeneity Analysis requirement. Substep 1 shrinks the object scores towards the origin by averaging, substep 2A takes the resulting category quantifications and shrinks them more by even more averaging. Thus in substep $2B$ we have to renormalize the object scores such that they are centered, standardized, and uncorrelated. This gives

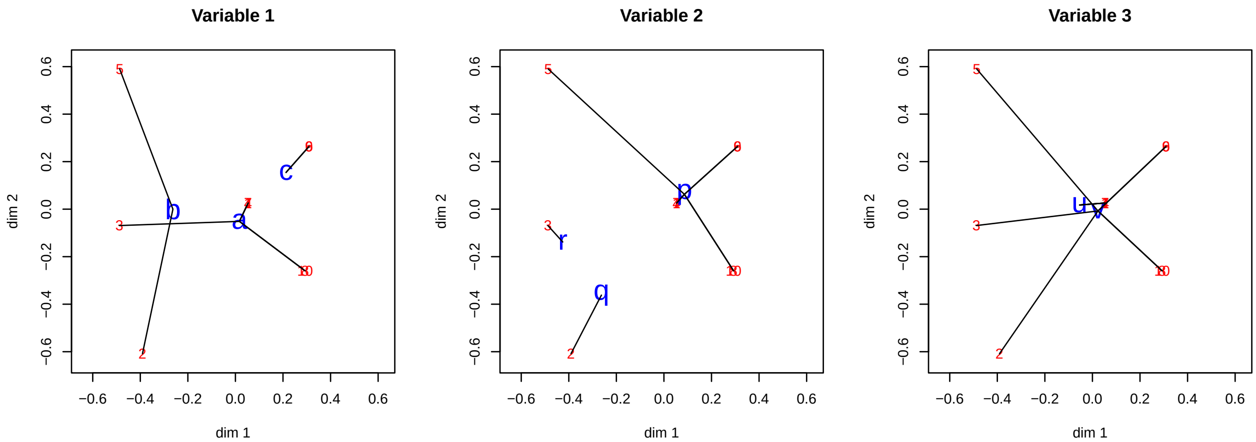

Figure 4: Graph Plots, Iteration 1, substep 2B, Small Example

Loss, which is now the proper loss for a normalized configuration, has decreased to 4.5538169.

Now that we have new category quantifications and new suitably normalized object scores we can start the next iteration, and again improve both in two substeps. Ultimately, after a certain number of iterations, there is no change any more from one iteration to another, and we have reached the optimal solution. In other words, there is convergence, and our Homogeneity Analysis is finished. Note that the renormalization in step 2B is necessary, because without it both object scores and category quantifications would become smaller and smaller, and converge to the origin. Of course the origin does have loss zero, but it is never a proper description of the data.

The optimal graph plots, after the iterations have converged, are

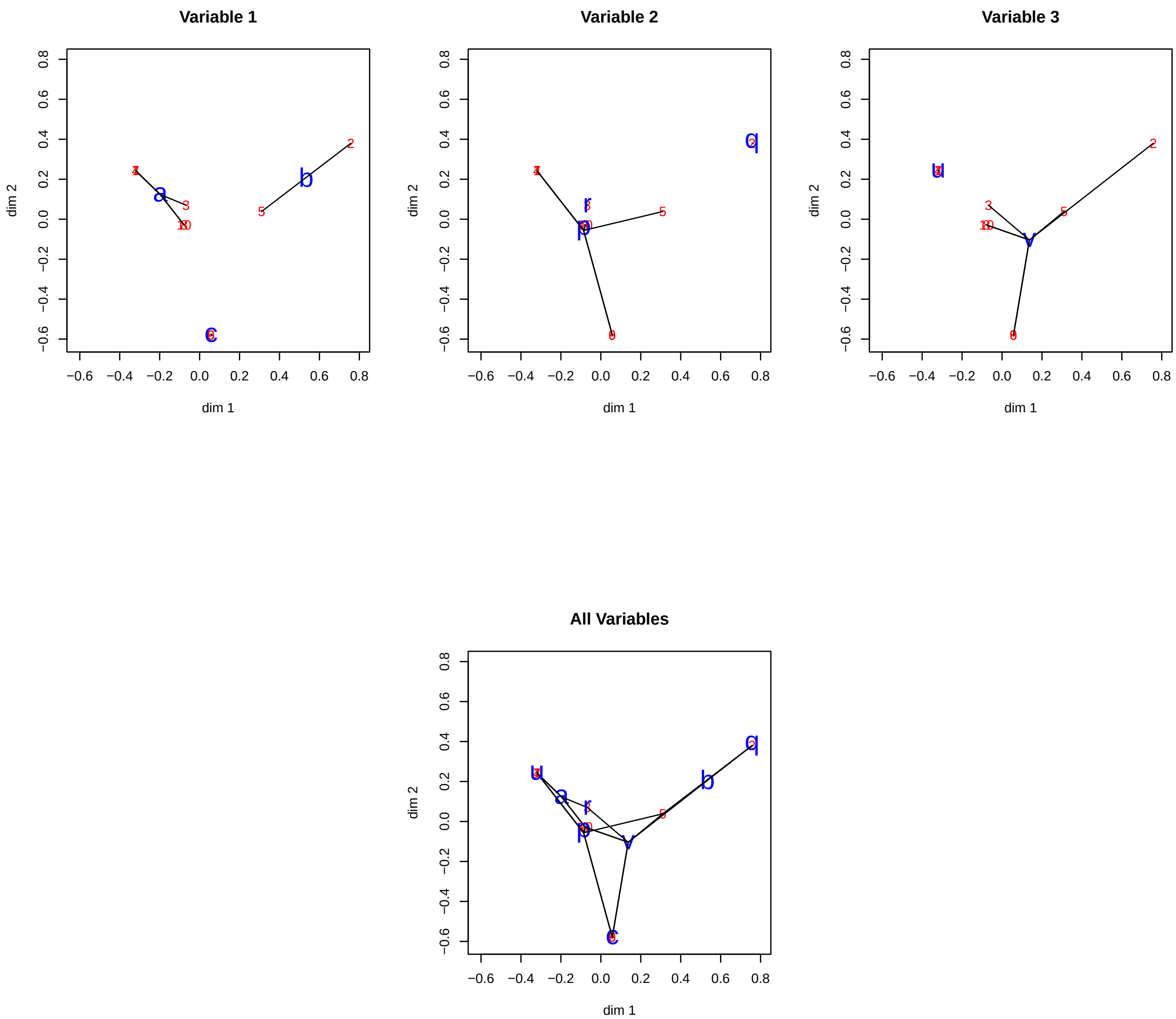

Figure 5: Graph Plots, Optimum Solution, Small Example

The minimum loss for these data is 2.8377218.

Because the object scores are in deviations from the mean, and the category quantifications are weighted means of object scores it follows that category quantifications are in deviations from the weighted mean, with weights equal to the marginals of the variable. Thus category quantifications for each variable are distributed around the origin.

In Homogeneity Analysis the graph plots for individual variables are often called *star plots*, because the optimal category quantification is the centroid of the scores of the objects in the category. Thus it is somewhere in the middle of a bunch of objects, which are connected to it by lines. Thus the subset of the graph for each category is a star graph, and the corresponding plot for the variable with these stars is a star plot. The top three plots in figure 5 are examples of such star plots. One could formulate the objective of Homeogeneity Analysis as finding normalized object scores in such a way that the stars (over all categories of all variables) are as small as possible. Or, in yet another formulation, we want to maximize the between-category variation and minimize the within-category variation.

Variables with a small star, which is necessarily close to the origin, have poor discrimination power. The average object score for each category is about the same. In general, categories with a large number of observations will have an average close to the average of all observations, and thus they will be close to the origin. And, conversely, categories with a small number of observations will tend to be relatively far from the origin.

#Specifics

##Passive, Supplementary, and Constraining Variables

Now, going back to the Durent Bend data, we do not have the values of all 5888 sherds on the three variables. The sherds are aggregated over various site/depth combinations and the original data cannot be recovered from the aggregated table. But the framework of Homogeneity Analysis can still be applied by using *equality restrictions* (Van Buuren and De Leeuw (1992)). The only thing added is that we require that sherds in the same site/depth get the same object score. Or, geometrically, all sherds in the same site/depth are mapped into the same point in the joint plot.

The Homogeneity Analysis loss function is still minimized in two steps. The first step, updating the category quantifications, is still the same as in Homogeneity Analysis without equality restrictions. The second step, which updates the object scores, now has three substeps instead of two. In substep $2A$ we compute the average of the category quantifications of the categories the sherd is in. In substep $2B$ we replace these tentative object scores for the sherds by the site/depth averages, and in substep $2C$ we normalize the object scores, making them centered, standardized, and uncorrelated.

The graph plots on unconstrained Homogeneity Analysis must now be replaced by plots of *valued graphs*. For any variable each site/depth point is now connected to all category points, and the edge connecting the object and category point has a value equal to the number of sherds in the category.

As an example we use the GALO data, which has been used innumerable times before as a Homogeneity Analysis example, and is in the `Gifi` package in R ((**mair_deleeuw_17?**)). The GALO data can be used to show the difference between working with aggregated data (over sherds in the same site or students in the same school) and the raw data, which are actually unavailable for Durant Bend. Here is the description of the GALO data in the help file of the package.

```
galo Gifi   R Documentation
GALO dataset

Description

The objects (individuals) are 1290 school children in the sixth grade of elementary scho

Usage

galo
```

Format

`Data frame with the five variables Gender, IQ, Advice, SES and School. IQ (original rang`

`SES:`

`LoWC = Lower white collar; MidWC = Middle white collar; Prof = Professional, Managers; S`

`Advice:`

`Agr = Agricultural; Ext = Extended primary education; Gen = General; Grls = Secondary sc`

`References`

`Peschar, J.L. (1975). School, Milieu, Beroep. Groningen: Tjeek Willink.`

Note that IQ is measured by the GIT (Groningen Intelligence Test) and that Advice refers to the sixth grade teachers advice abut the most appropriate form of secondary educations for the students.

We first ignore the school variable, and only analyze the four variables Gender, IQ, Advice, and SES. Separate joint plots for the four variables, with both object scores and category quatifications, are in figure 6. We do not make star plots (by drawing the lines from the object points to the category points they are in) in this case, because 1290 lines in a plot just create a big black blob.

The joint plots show a curved one-dimensional solution with good students on the left and poor students on the right. Such curved solutions, sometimes called *horseshoes*, are a familiar outcome of Homogeneity Analysis when there is a dominant single dimension explaining the results (in this case student achievement). Both IQ and Advice differentiate students well (mainly because teachers rely on IQ scores in their advice), which means they will have the smallest stars. Girls tend to be better students than boys, and SES mainly contrasts the two extremes categories PROF and UNSK.

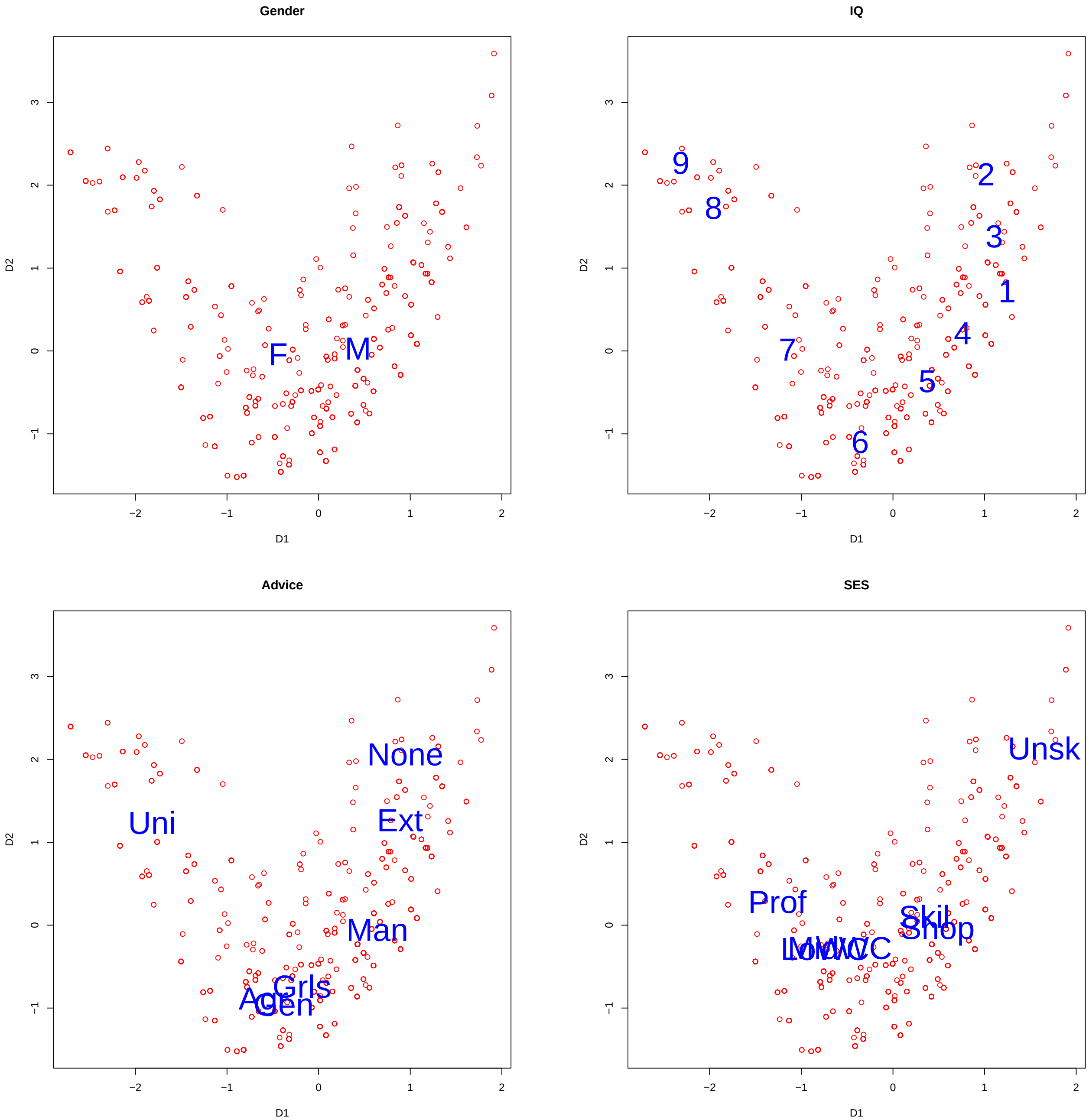

Figure 6: Joint Plots, GALO Example without School

Our first analysis does not use the School variable at all. In the terminology of Gifi School is a *passive variable*. There are a number of different ways in which we can incorporate School. The first is the obvious one: just repeat the Homogeneity Analysis over all five variables and include School. The joint plots of the first four variables are in figure 7.

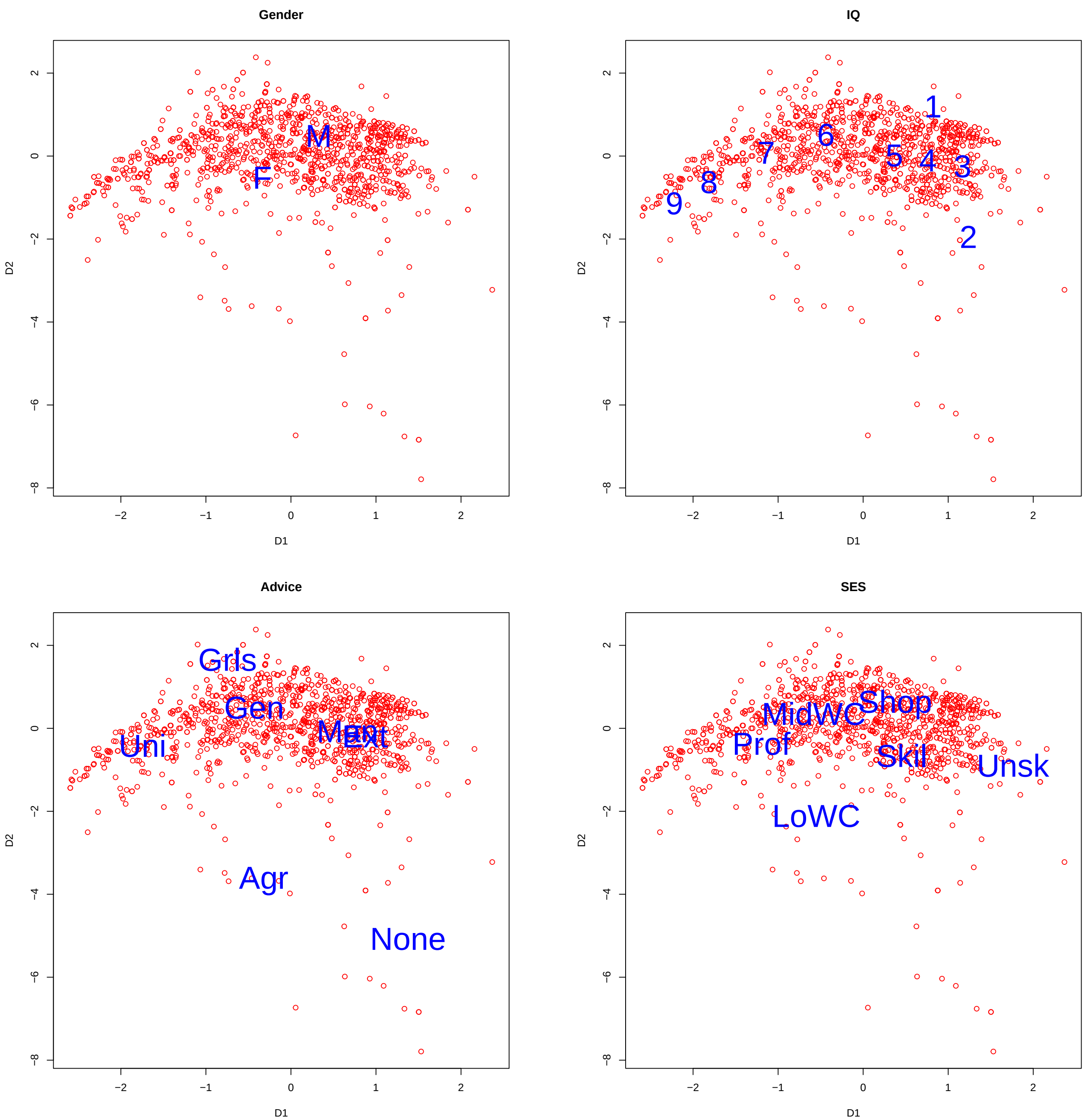

Figure 7: Joint Plots, GALO Example with School

The horseshoe pattern is still there, but it is less pronounced, mainly because of several outlying students at the bottom of the plot more or less defining the vertical dimension. This is due to including the School variable. The joint plot for the School variable is in figure 8. We see the outliers are in school 25, a small school with 11 low-IQ students, possibly some type of special education school. The figure also shows the star for school 25. The actual data for the eleven students are

```r
galo[galo[,5]=="25",]
```

```
##     gender IQ advice  SES School
## 898      F  3   None Skil     25
## 899      F  4    Agr Skil     25
## 900      F  5    Man Skil     25
## 901      F  5    Man Skil     25
## 902      F  5    Man Skil     25
## 903      F  5    Agr LoWC     25
## 904      F  4   None Skil     25
## 905      M  4   None Prof     25
## 906      F  4    Man LoWC     25
## 907      F  2   None Skil     25
## 908      F  3   None Skil     25
```

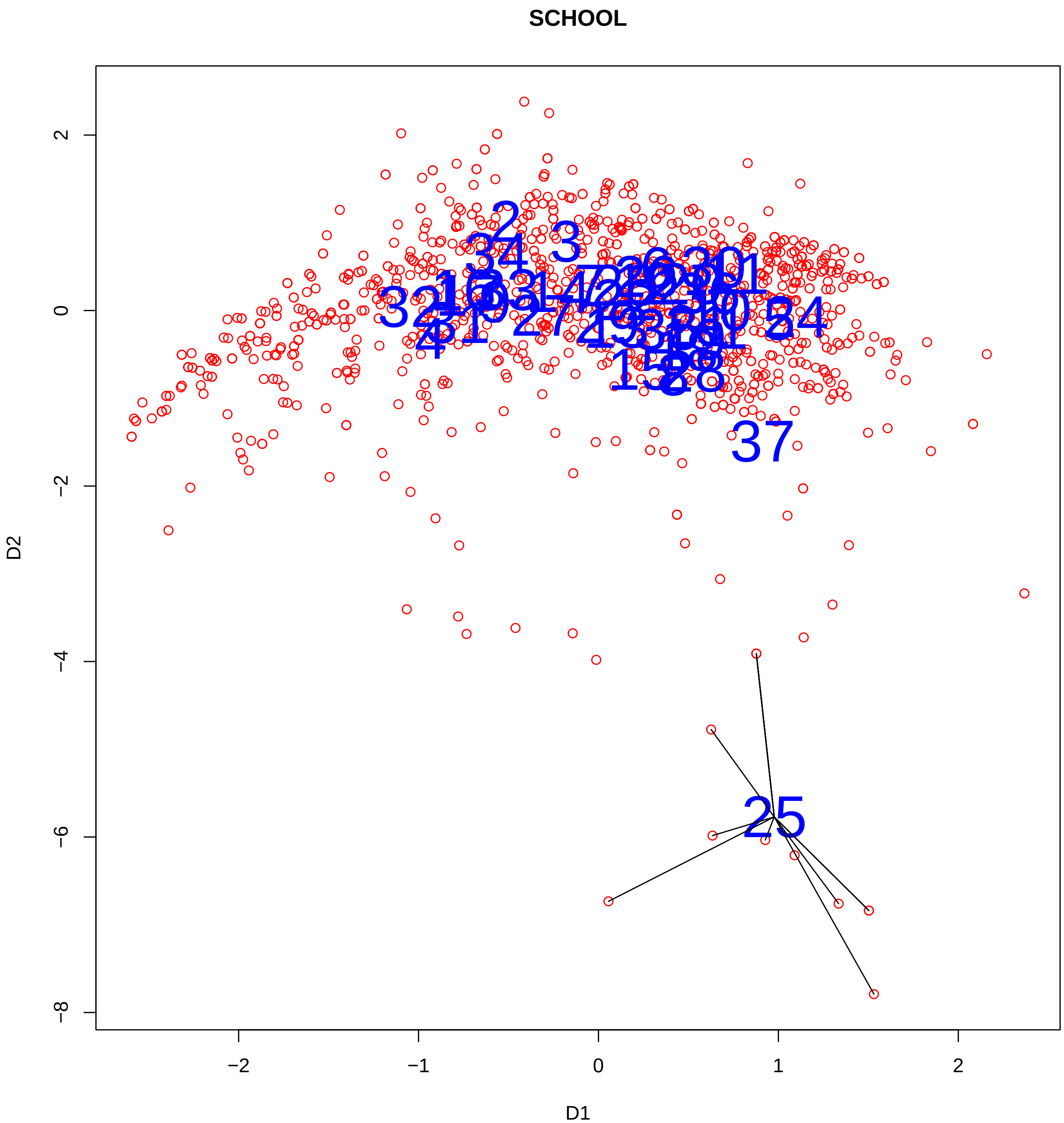

Figure 8: Joint Plot, GALO Example with School, School Variable

There is another, and perhaps more interesting, way to incorporate School in our analysis, by using it as what is commonly known as a *supplementary variable*. Such a supplementary variable does not actively enter into the Homogeneity Analysis, but after the analysis of the remaining variables we can compute category quantifications of the supplementary variable as centroids of object scores in the categories. Thus we can make star plots for the passive variables that have not been used in the analysis. This is done in figure 9. The horseshoe of object scores does not change from the one in figure 6, but by not including School in the analysis we do not give school 25 the opportunity to dominate the second dimension. It is still true that the same schools (5, 24, 25, 37) perform poorly, and the same schools (4, 17, 31, 32) perform well, but generaly the category quantifications are more well-behaved.

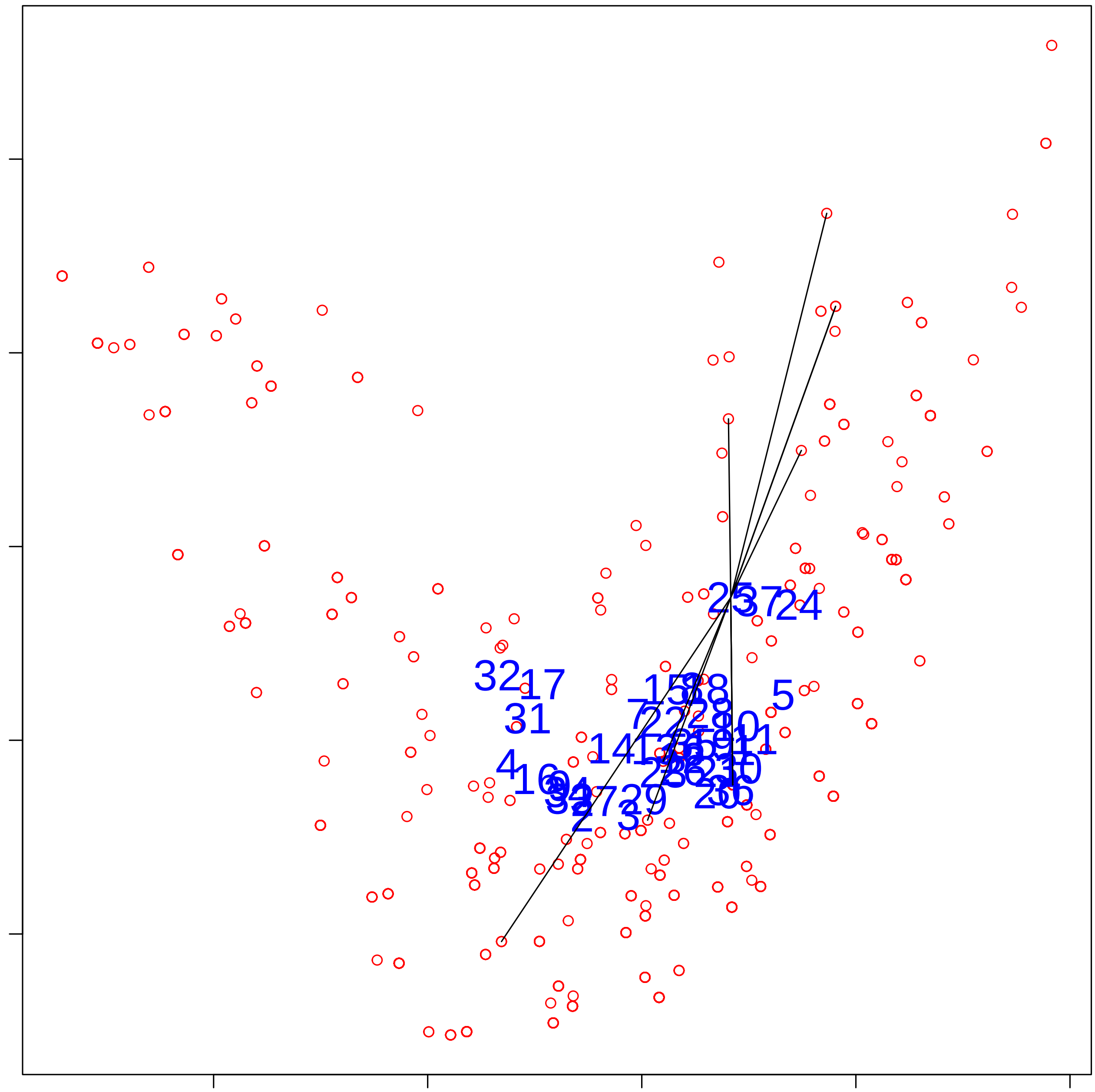

Figure 9: Object Score Plot, School Supplementary, GALO Example

We now repeat the analysis, requiring that students in the same school get the same object score. We treat school as a *constraining variable*, performing a Homogeneity Analysis with equality restrictions on the object scores (Van Buuren and De Leeuw (1992)). In terms of the joint plot we require the stars for the school variable to collapse into single points. Computationally this is easiest to do using the the R package `anacor` (De Leeuw and Mair (2009)). In this constrained Homogeneity Analysis each school gets an object score, and these scores are plotted in figure 10. Not surprisingly school 25 is now even more of an outlier, but otherwise schools are dispersed pretty much in the same way as before. For the GALO example this constrained analysis throws away useful information and gives a result which is inferior to the supplementary variable approach. For the archeological data we do not ignore within-site information, because the data are aggregated over sherds in the same site to begin with.

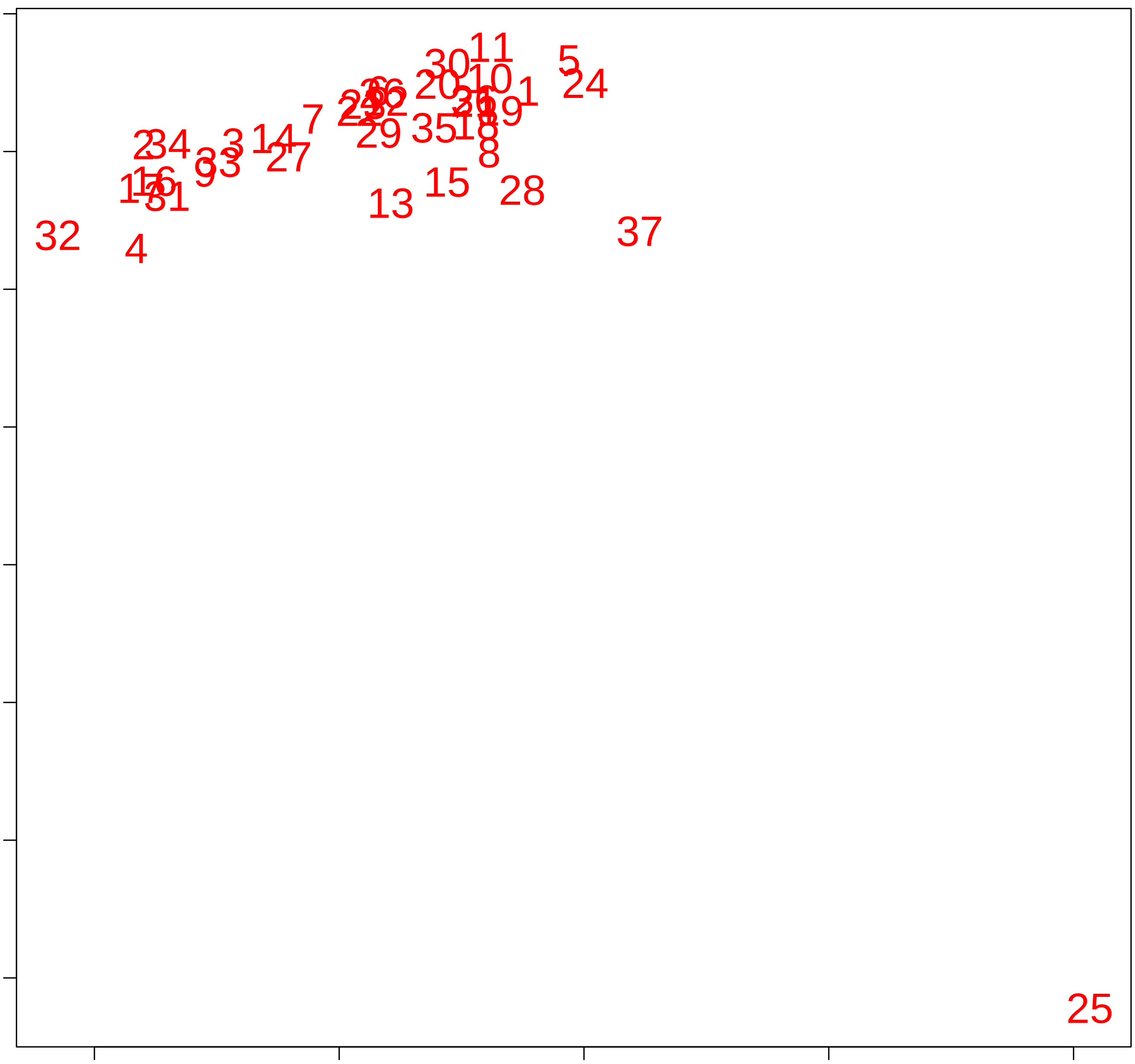

Figure 10: Object Score Plot, School Constraints, GALO Example

##Binary Variables

Besides aggregation, another proerty of the Durant Bend data is that the three variables describing the sherds are binary (CS/Plain, Dark/Light, Thin/Thick). This implies some special properties of the Homogeneity Analysis.

We have seen that category quantifications are in deviations from the weighted mean, with the weights equal to the marginal frequencies of the variable. If a variable has only two categories, and our Homogeneity Analysis has two dimensions, that means that the two category quantifications for a variable are on a line through the origin. The direction of the line is determined by the marginals of the variables. What Homogeneity Analysis gives us is how far away from the origin the category quantifications are placed on the line to get the smallest stars.

We have said very little so far about the number of dimensions we choose for our Homogeneity Analysis. The default is to choose two, because two-dimensional joint and graph plots are the easiest to look at. The maximum number of dimensions in Homogeneity Analysis, i.e. the number of dimensions that are needed to represent all variation in the data, is equal to the total number of categories minus the number of variables. In the GALO example (without School) that is $2+9+6+7-4=20$ but in the Durant Bend example it is $2+2+2-3=3$. Only three dimensions will capture all variation.

#Analysis Durant Bend Data

The Durant Bend analysis is an aggregated Homogeneity Analysis of three binary variables, requiring equal object scores for all sherds in the same site/depth. The joint plot is in figure 11. For a discussion of these results we refer to the companion paper by (**nance_deleeuw_18?**). There is not much variation in the category quantifications of the three variables (the lines are rather short). In particular, the averages for light sherds and for dark sherds are very close, indicating not much discriminatiry power for that variable.

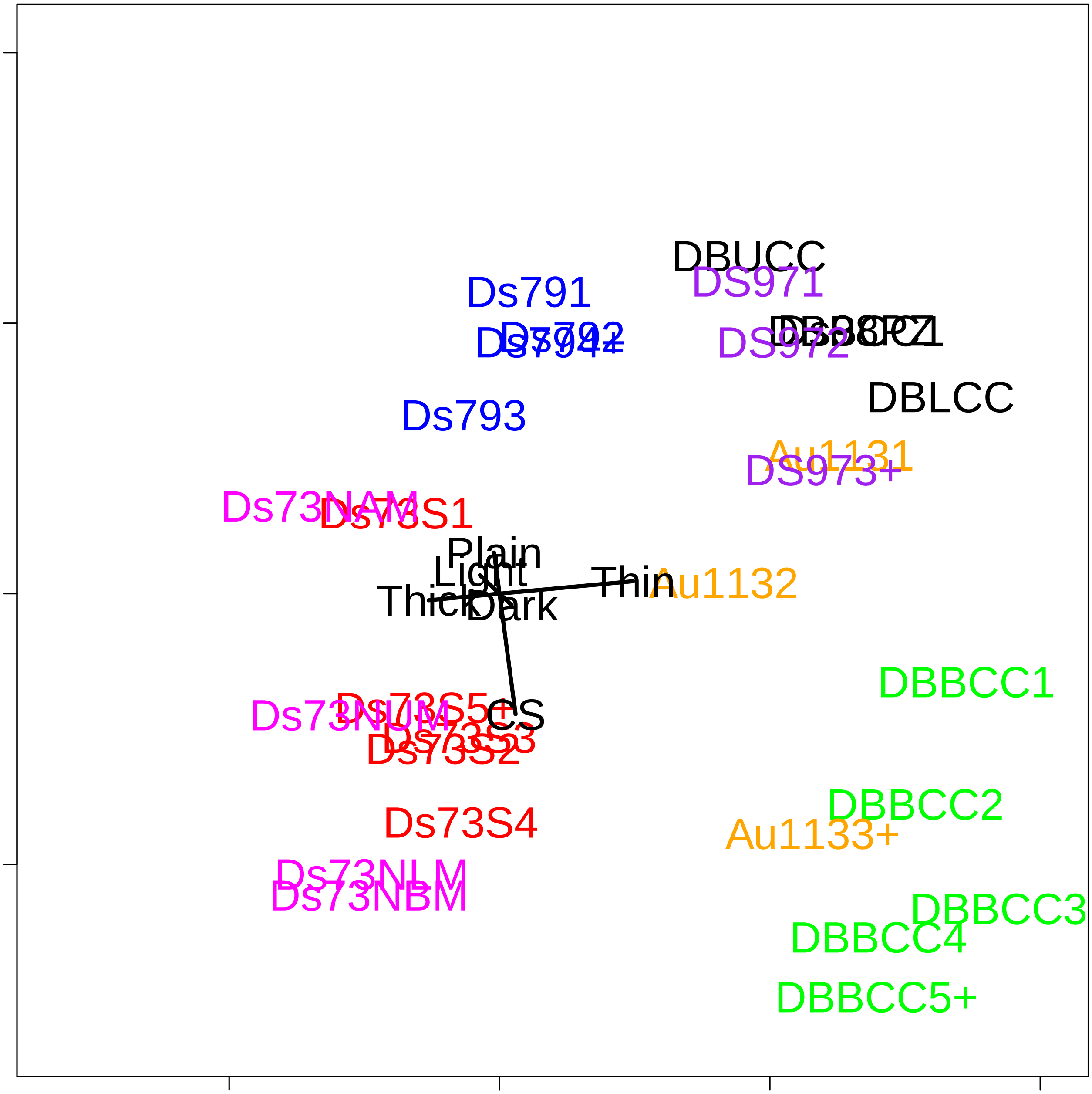

Figure 11: Joint Plot, Durant Bend, Two Dimensions

We have to realize, of course, that there are only three dimensions available to describe our data. This makes it interesting to look at the three-dimensional solution, specifically at light vs dark sherds. We more or less expect each variable to define a dimension, indicating relatively low correlations between the three variables, and consequently not much difference between sites. The summary of the three dimensional Homogeneity Analysis from `anacor` is

```
##
## CA fit:
##
## Total chi-square value: 2230.319
```

```
## Sum of eigenvalues (total inertia): 0.126
## Eigenvalues (principal inertias):
## 0.079 0.042 0.005
##
## Benzecri RMSE columns:  3.454901e-20
##
## Chi-square decomposition:
##                Chisq Proportion Cumulative Proportion
## Dimension 1 1391.515      0.624                 0.624
## Dimension 2  744.667      0.334                 0.958
## Dimension 3   94.136      0.042                 1.000
## Dimension 4    0.000      0.000                 1.000
## Dimension 5    0.000      0.000                 1.000
```

The two three-dimensional scatterplots, one for categoiry quantifications and one for sites, are in figures 14 and 15. We see that the third dimension indeed separates the light from the dark.

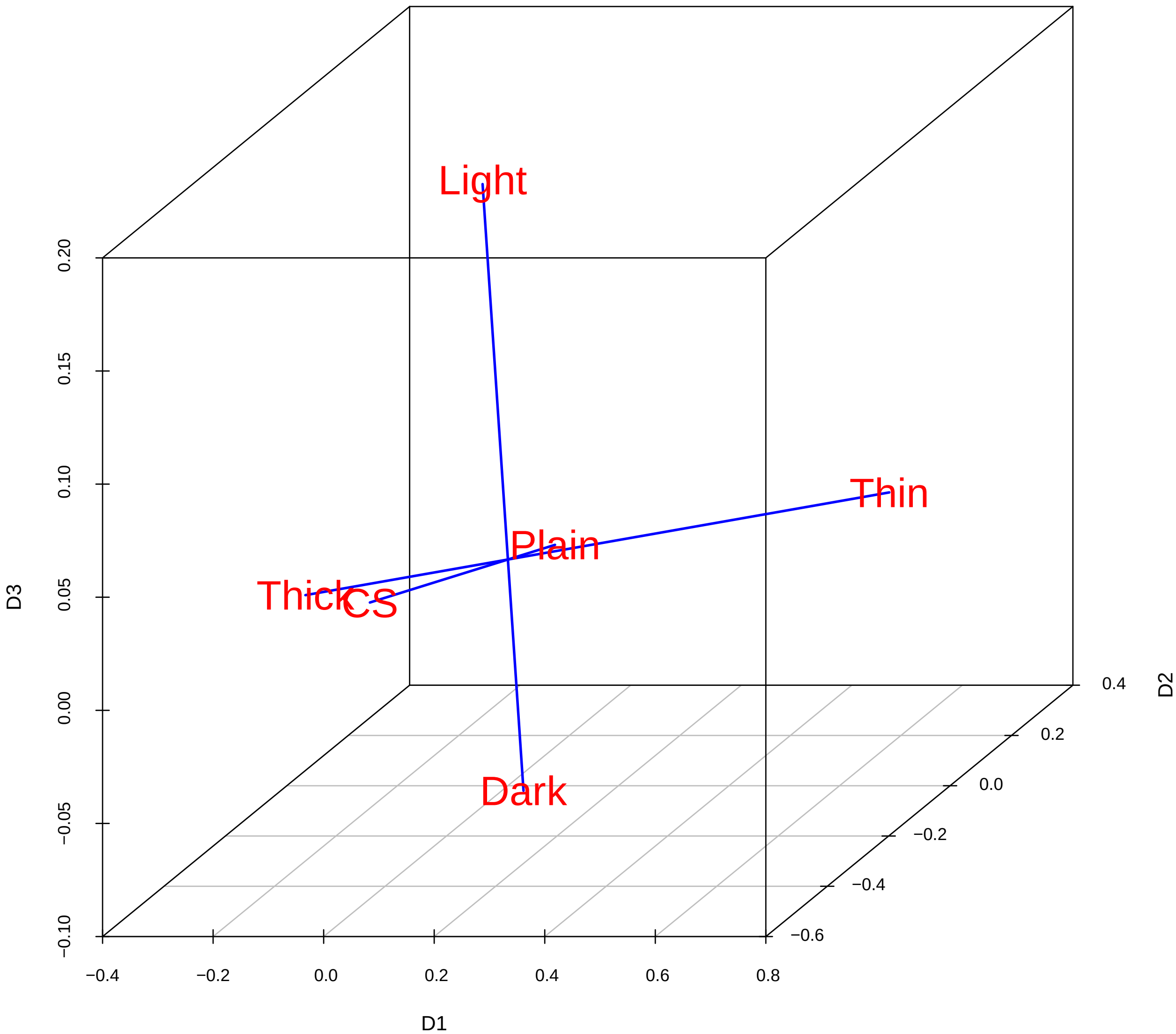

Figure 12: Category Quantifiations, Durant Bend, Three Dimensions

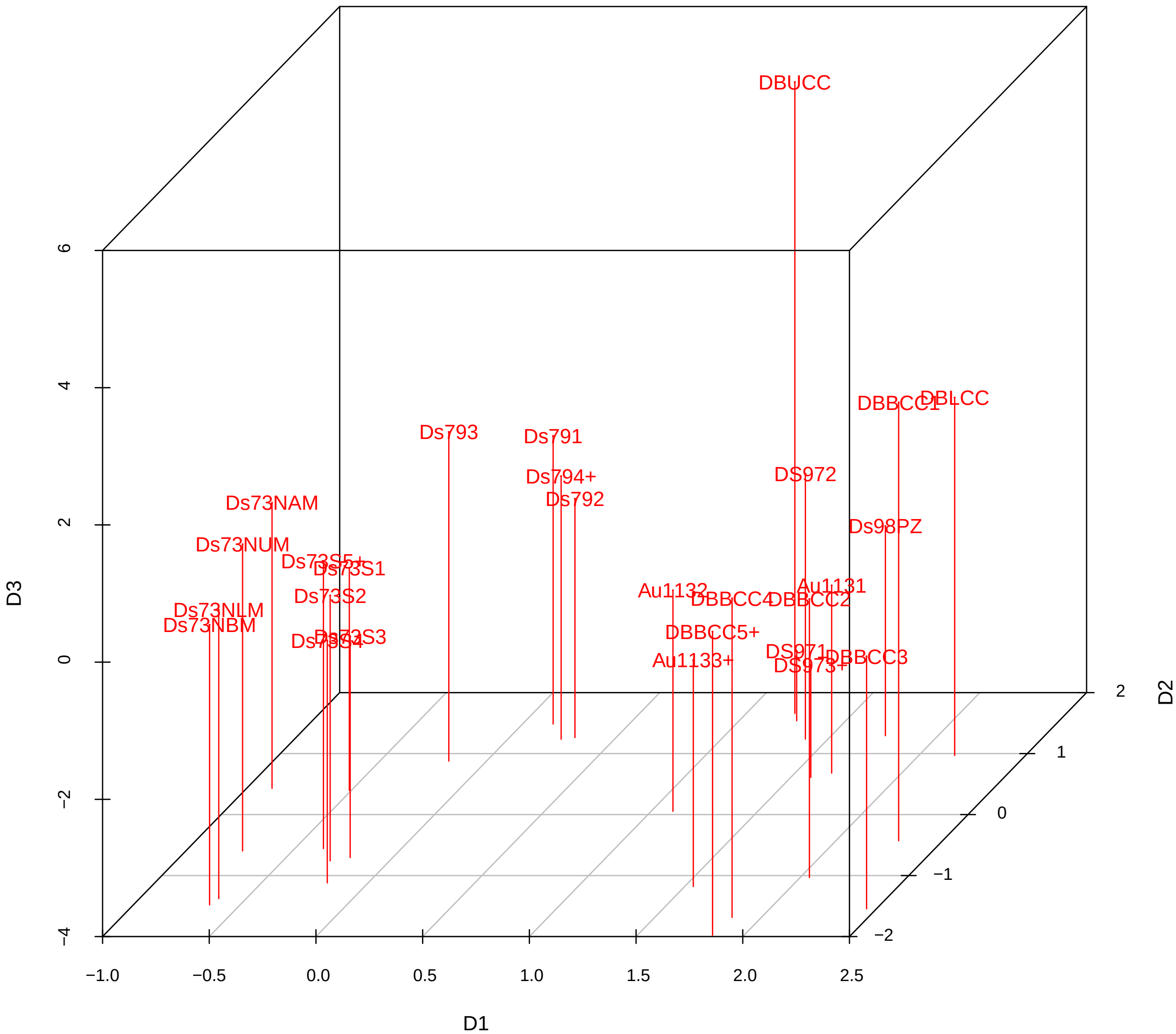

Figure 13: Object Scores, Durant Bend, Three Dimensions

#References

De Leeuw, J., and P. Mair. 2009. "Simple and Canonical Correspondence Analysis Using the R Package Anacor." *Journal of Statistical Software* 31 (5): 1–18.

Gifi, A. 1990. *Nonlinear Multivariate Analysis*. New York, N.Y.: Wiley.

Gower, J. C., and D. J. Hand. 1996. *Biplots*. Monographs on Statistics and Applied Probability 54. Chapman & Hall.

Greenacre, M. J. 1984. *Theory and Applications of Correspondence Analysis*. New York, New York: Academic Press.

Greenacre, M. J., and J. Blasius, eds. 2006. *Multiple Correspondence Analysis and Related Methods*. Chapman; Hall.

Nance, C. R., and J. De Leeuw. 2018. "Statistical Pespectives on Woodland Cultures in Central Alabama ." 2018.

Nance, C. Roger. 1976. "The Archeological Sequence at Durant Bend, Dallas County, Alabama." Special Publication 2. Orange Beach: Alabama Archeological Society.

Van Buuren, S., and J. De Leeuw. 1992. "Equality Constraints in Multiple Correspondence Analysis." *Multivariate Behavioral Research* 27: 567–83.