UCLA Department of Statistics
Statistical Consulting Center

Descriptive Multivariate Analysis

Jan de Leeuw

August 18, 2010

Outline

1. Data
   - Hartigan
   - Mammals
   - GALO
   - Senate
   - Roskam
   - Neumann
2. Coding
3. Algorithm
4. Stars and Rats
5. Constraints

The Data

Introduction

- We illustrate *Multiple Correspondence Analysis (MCA)* by first discussing some possible data sets (that happen to be in the `homals` package in R).

The Data

Introduction

- We illustrate *Multiple Correspondence Analysis (MCA)* by first discussing some possible data sets (that happen to be in the homals package in R).
- As we shall show (if we have enough time) this technique provides us with far-reaching generalizations of the two most popular multivariate techniques: *Principal Component Analysis (PCA)* and *Multiple Regression Analysis (MRA)*.

The Data

Introduction

- We illustrate *Multiple Correspondence Analysis (MCA)* by first discussing some possible data sets (that happen to be in the `homals` package in R).
- As we shall show (if we have enough time) this technique provides us with far-reaching generalizations of the two most popular multivariate techniques: *Principal Component Analysis (PCA)* and *Multiple Regression Analysis (MRA)*.
- Our explanation will emphasize geometry, and not use formulas.

The Data

Introduction

- We illustrate *Multiple Correspondence Analysis (MCA)* by first discussing some possible data sets (that happen to be in the `homals` package in R).
- As we shall show (if we have enough time) this technique provides us with far-reaching generalizations of the two most popular multivariate techniques: *Principal Component Analysis (PCA)* and *Multiple Regression Analysis (MRA)*.
- Our explanation will emphasize geometry, and not use formulas.
- The examples are somewhat smaller than typical multivariate data sets, but they show the variety of possibilities.

The Data

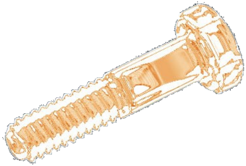
Introduction

- We illustrate *Multiple Correspondence Analysis (MCA)* by first discussing some possible data sets (that happen to be in the homals package in R).
- As we shall show (if we have enough time) this technique provides us with far-reaching generalizations of the two most popular multivariate techniques: *Principal Component Analysis (PCA)* and *Multiple Regression Analysis (MRA)*.
- Our explanation will emphasize geometry, and not use formulas.
- The examples are somewhat smaller than typical multivariate data sets, but they show the variety of possibilities.
- In R the data are in a data frame, of any size, consisting of numerical or character vectors, as well as ordered and unordered factors.

Hartigan's Hardware

What is it about ?

- A number of bolts, nails, screws, and tacks are classified according to a number of criteria.
- Taken from the book by John Hartigan on Cluster Analysis.
- `data(hartigan,package="homals")`



Hartigan's Hardware

Variables and Categories

Thread:	Y = Yes, N = No
Head:	F = Flat, C = Cup, O = Cone, R = Round, Y = Cylinder
Indentation:	N = None, T = Star, L = Slit
Bottom:	S = Sharp, F = Flat
Length:	(in half inches)
Brass:	Y = Yes, N = No

Hartigan’s Hardware

What are the data ?

	thread	head	indentation	bottom	length	brass
tack	N	F	N	S	1	N
nail1	N	F	N	S	4	N
nail2	N	F	N	S	2	N
nail3	N	F	N	F	2	N
nail4	N	F	N	S	2	N
nail5	N	F	N	S	2	N
nail6	N	C	N	S	5	N
nail7	N	C	N	S	3	N
nail8	N	C	N	S	3	N
screw1	Y	O	T	S	5	N
screw2	Y	R	L	S	4	N
screw3	Y	Y	L	S	4	N
screw4	Y	R	L	S	2	N
screw5	Y	Y	L	S	2	N
bolt1	Y	R	L	F	4	N
bolt2	Y	O	L	F	1	N
bolt3	Y	Y	L	F	1	N
bolt4	Y	Y	L	F	1	N
bolt5	Y	Y	L	F	1	N
bolt6	Y	Y	L	F	1	N
tack1	N	F	N	S	1	Y
tack2	N	F	N	S	1	Y
nailb	N	F	N	S	1	Y
screwb	Y	O	L	S	1	Y

Mammals Example

What is it about ?

- The objects (individuals) are 66 animal species.
- `data(mammals,package="homals")`

Mammals Example

The Variables

Eight variables describing the teeth of the animals.

- top incisors: (1) zero; (2) one; (3) two; (4) three or more.
- bottom incisors: (1) zero; (2) one; (3) two; (4) three; (5) four.
- top canine: (1) zero; (2) one.
- bottom canine: (1) zero; (2) one.
- top premolar (1) zero; (2) one; (3) two; (4) three; (5) four.
- bottom premolar: (1) zero; (2) one; (3) two; (4) three; (5) four.
- top molar: (1) zero, one or two; (2) more than two.
- bottom molar: (1) zero, one or two; (2) more than two.

GALO Example

What is it about ?

- The objects (individuals) are 1290 school children in the sixth grade of elementary school in the city of Groningen (Netherlands) in 1959.
- `data(galo,package="homals")`

GALO Example

The Variables

Gender: M/F.

IQ: The original range (60 to 144) has been categorized into 9 ordered categories.

SES: LoWC = Lower white collar; MidWC = Middle white collar; Prof = Professional, Managers; Shop = Shopkeepers; Skil = Schooled labor; Unsk = Unskilled labor.

Advice: Agr = Agricultural; Ext = Extended primary education; Gen = General; Grls = Secondary school for girls; Man = Manual, including housekeeping; None = No further education; Uni = Pre-University.

Schools: Schools are numbered from 1 to 37.

Senate Data

What is it about ?

- 2001 senate votes on 20 issues selected by Americans for Democratic Action. The votes selected cover a full spectrum of domestic, foreign, economic, military, environmental and social issues.
- 100 senators, 21 variables, the first variable is Republican vs Democrat, the remaining 20 are binary votes (with missing data).
- `data(senate,package="homals")`

Roskam Example

What is it about ?

In 1968 thirty-nine psychologists at the University of Nijmegen rank nine psychology research/teaching areas in order of preference.

Roskam Example

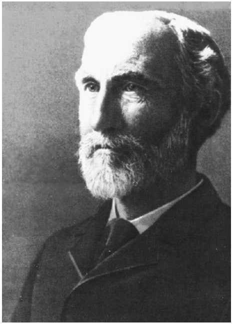
The Objects

SOC	Social Psychology
EDU	Educational and Developmental Psychology
CLI	Clinical Psychology
MAT	Mathematical Psychology and Psychological Statistics
EXP	Experimental Psychology
CUL	Cultural Psychology and Psychology of Religion
IND	Industrial Psychology
TST	Test Construction and Validation
PHY	Physiological and Animal Psychology

Neumann Example

What is it about ?

- Willard Gibbs discovered a theoretical formula connecting the density, the pressure, and the absolute temperature of a mixture of gases. He applied this formula and the estimated constants to 65 experiments of Neumann.
- A data frame with temperature, pressure, and density on 65 experiments.
- `data(neumann,package="homals")`



Outline

An Artifical Dataset

Small

	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

Variables, Objects, Categories

What characterizes this example ?

- In R terms, the data are a `data-frame`. Each variable is a `factor`.
- The n rows corresponds with *objects*, that are measured (or classified) by the m columns, which correspond with *variables*.
- Variable j maps the objects into a set of k_j *categories* (which R calls `levels`). We use $K = \sum_{j=1}^{m} k_j$ for the total number of categories.
- All variables have a finite number of categories, although it is possible that the number of categories of a variable is equal to the number of objects.

Data as Graphs

- We can think of the data as a *graph* on the $n + K$ objects and categories.
- An object and a category are *connected* or *adjacent* if the data place the object in the category.
- Each object is connected to m categories. Thus there are nm adjacencies in the graph (unless there are *missing data*).
- The graph is *bipartite*, connections only go from one group (objects) to another (categories).

Data as Adjacency Matrices

Small Example

This is the off-diagonal part of the adjacency matrix of the graph. The diagonal parts are zero, because the graph is bipartite.

	a	b	c	p	q	r	u	v
01	1	0	0	1	0	0	1	0
02	0	1	0	0	1	0	0	1
03	1	0	0	0	0	1	0	1
04	1	0	0	1	0	0	1	0
05	0	1	0	1	0	0	0	1
06	0	0	1	1	0	0	0	1
07	1	0	0	1	0	0	1	0
08	1	0	0	1	0	0	0	1
09	0	0	1	1	0	0	0	1
10	1	0	0	1	0	0	0	1

Defining the Graphplot

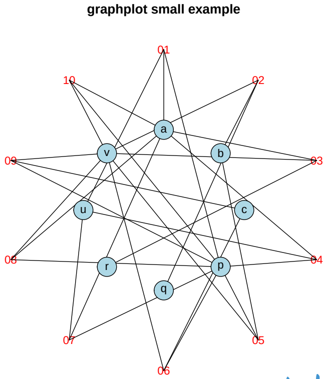
A Picture Is Worth A Thousand Numbers

- Let us now engage in *Graph Drawing*.
- Suppose X is a configuration of the n objects drawn in $\mathbb{R}^p$ and suppose Y is a configuration of the $K = \sum k_j$ categories in $\mathbb{R}^p$.
- We can connect (by a line) each object point x_i to the m category points y_ℓ that object i is in. This is the *graphplot*.
- The graphplot has $n + K$ points and nm lines, with m lines connected to each object (unless there are missing data).
- Note that the graphplot contains all information in the data, we can reproduce the data exactly from the plot (with some patience).

Graphplot

With Arbitrary Locations

We construct a graphplot by (arbitrarily) putting the $n = 10$ object scores as red names equally spaced on a circle and the $K = 8$ category quantifications as lightblue circles equally spaced on a smaller concentric circle.



Homogeneity

Qualitative discussion

- We say a graphplot is *homogeneous* if the lines are short.
- Short means "relatively short". There are a total of nK possible lines in the graph, given that it must be bipartite. We have homogeneity if the actual nm lines generated by the data are short compared to the nK possible lines.
- But of course the location so far in $\mathbb{R}^p$, in our example the plane with $p = 2$, has been completely arbitrary.
- And thus we may want to look for the graphplot that is as homogeneous as possible by moving the object and category points around.

Outline

Loss Function

- In MCA the loss function we use is the sum of squares of the lengths of all the *nm* lines in the graphplot. Other choices are possible, of course, but using the sum of squares leads to easy computations and also allows us to tie MCA to principal component analysis and multiple regression.

Loss Function

- In MCA the loss function we use is the sum of squares of the lengths of all the *nm* lines in the graphplot. Other choices are possible, of course, but using the sum of squares leads to easy computations and also allows us to tie MCA to principal component analysis and multiple regression.
- But simply minimizing the sum of squares will not do, because we can just collapse all points into a single point. Some form of normalization is needed. We normalize the objects scores by requiring that they are centered, standardized, and uncorrelated.

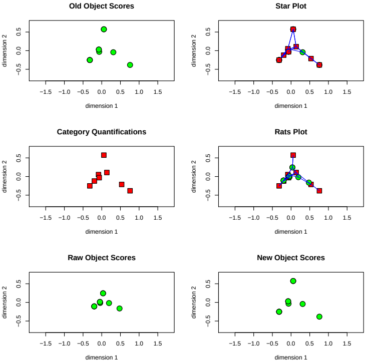
Alternating Least Squares

Or: Reciprocal Averaging

- Start with iteration $k = 0$ and some arbitrary (normalized) object scores.
- Then do
  1. Compute category centroids as the averages of the scores of objects in the category.
  2. Compute raw object scores as the averages of the quantifications of the categories the object is in.
  3. Normalize the raw object scores.
  4. If there is no change in iteration k then stop, otherwise $k \leftarrow k + 1$ and repeat.

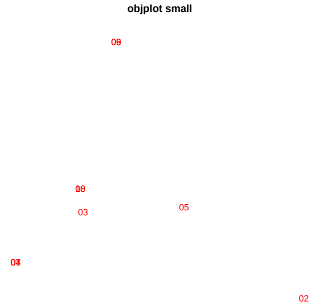
Small Example

The Four Steps



Small Example

Iterations

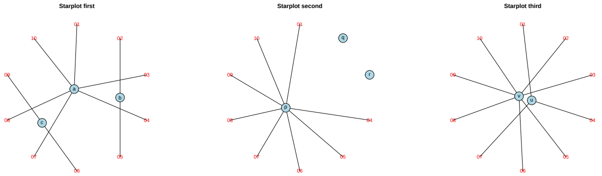


Outline

Making Starplots

Small Example

Taking the category quantifications equal to the category centroids gives the smallest within category variance, i.e. the smallest sum of squared line-lengths, for given object scores. The corresponding graph subplots are called *starplots* because the category graphs are stars. For the circular object scores we make one for each variable.



Between and Within

Geometry of Object Normalization

- The sum of squares of the n line lengths for a given variable are the *discrimination measures*.
- We maximize the sum of between-category variances, and minimize the sum of the within-category variances, while keeping the sum of total variances equal to the identity.
- Of course this is easy to do if there is only one variable. Just let all object points coincide with the category points they are in. So the MCA solution for $m > 1$ is a (least squares) *compromise*.

MCA

Smallest stars

- MCA can now be formulated as finding the normalized object scores such that the $\sum_{j=1}^{m} k_j$ stars are as small as possible.
- It is the basic technique implemented in the R package `homals`, although the package can do much more.
- It was first described by Guttman in 1941, then rediscovered by, among others, Burt [1950], Hayashi [1952], Benzécri (see Cordier, 1964), and De Leeuw [1968, 1973].

MCA

Points to Remember

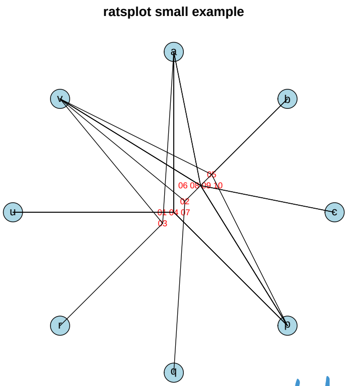
Remember that:

- the size of the stars is measured by the square of the length of the lines connecting object points and category points;
- coordinates of object points are *object scores*, coordinates of category points are *category quantifications*;
- the configuration of object points is orthonormal, i.e. dimensions are normalized and uncorrelated.

Ratsplots

Small Example

We compute the raw object score as the centroid of the quantifications of the m categories it is in, and then connect it to these category quantifications. We call this dual of the Starplot the *Ratsplot*. Here is the overlay of 10 Ratsplots.



Outline

Constraints

MCA Extensions

- So far, we have only discussed MCA, which seems rather limited.

Constraints

MCA Extensions

- So far, we have only discussed MCA, which seems rather limited.
- But it turns out that we can connect MCA to classical multivariate analysis by various constraints on the category quantifications.

Constraints

MCA Extensions

- So far, we have only discussed MCA, which seems rather limited.
- But it turns out that we can connect MCA to classical multivariate analysis by various constraints on the category quantifications.
- We first discuss *rank constraints*, which will allow us to incorporate principal component analysis.

Constraints

MCA Extensions

- So far, we have only discussed MCA, which seems rather limited.
- But it turns out that we can connect MCA to classical multivariate analysis by various constraints on the category quantifications.
- We first discuss *rank constraints*, which will allow us to incorporate principal component analysis.
- And then *additivity constraints*, which will allow us to incorporate multiple regression, canonical analysis, and discriminant analysis.

Constraints

MCA Extensions

- So far, we have only discussed MCA, which seems rather limited.
- But it turns out that we can connect MCA to classical multivariate analysis by various constraints on the category quantifications.
- We first discuss *rank constraints*, which will allow us to incorporate principal component analysis.
- And then *additivity constraints*, which will allow us to incorporate multiple regression, canonical analysis, and discriminant analysis.
- All this can be done with the R package `homals`.