Lecture Notes in Quantitative Biology

Multivariate Analysis -- Combining Variables

Chapter 20.4 (from 30 November and 2 December 1993)

Revised 26 November 1997

ReCap
EDA - Combining variables
Multivariate analysis
  Relation to GLM
  Strategy and tactics
  Examples

on chalk board

Handout. Mvar.ref

Add more examples that focus on interpreting axes.

ReCap. EDA is a combination of graphical and formal analysis with the goal of discovering the "best" model.

EDA and inference. The inference from the model to a larger population is much looser than in formal statistical inference. EDA is iterative. It uses a screening criterion rather than a significance level α

Execution. Elements of good quantitative analysis still apply.
- Define all quantities that are used
- Identify response and explanatory variables.
- Decide whether to undertake exploratory or confirmatory analysis, state reasons, use screening criterion or significance level, as appropriate.
- Box and arrow diagrams useful.

Correlation measures the relation of two variables. There are two response variables, which are related to a single (fabricated) explanatory variable. We can extend this to more than two response variables.

Today: Multivariate Analysis

Multivariate techniques combine variables, simplification to bring out pattern. Procedure is
- Define quantities.
- Identify response and explanatory variables.
- State rationale for exploratory analysis.
- Reduce set of response variables to unobserved variables (factors).
- Quantify the fit of data to model.
- Interpret the axes.

EDA Combining variables with Multivariate analysis.

Biological systems are complex. Understanding requires skilful summarization, when data sets consist of many measurements on many quantities. Examples:
Wisconsin trees.
Skeletal measurements on many bones.

Verbal models are usually the first step in simplification, but these are difficult to develop for complex data sets with many variables.

Graphs are also limited. It is difficult to display more then 3 variables in one graph. Graphs of 2 or 3 quantities can be used to visualize the information, but the number of such graphs becomes too many for data sets with many quantities. For 10 quantities, the number of XY plots will be $10!/(2!\ 8!) = 45$.

Formal models are a potentially powerful way of summarizing information. But these are often so far removed from the data that it becomes difficult to pick out pattern.

An effective solution is to combine graphical methods with summary statistics.

Multivariate analysis is a combination of logic, numerical summarization, and graphical display to extract pattern from a large number of quantities, measured across multiple cases. The basic technique is to combine quantities to create new quantities. The new quantities are weighted combinations of the original set of quantities. We have already seen a simple example. Correlation analysis combines two quantities to create a new (unobserved) quantity.

A large number of statistical techniques, most of them exploratory in aim, go under the name of "multivariate analysis." Nearly all have three things in common. There are two or more response variables, there are usually no explanatory variables, and they use eigen analysis to compute new axes, which are then interpreted. Examples of multivariate analysis are

Principal components analysis
Principal coordinate analysis
Factor analysis
Canonical correlation
Multidimensional scaling (uses ranks)

Multivariate techniques based on eigen analysis are sometimes called canonical analyses (not the same as canonical correlation)

Relation to GLM

Multivariate methods are related to the linear modelling techniques already learned. To show the relation, we need to distinguish among sets of variables. Boldface type designates a matrix consisting of columns (variables) and rows (cases). Three different kinds of sets

Y a set of observed response variables.
X a set of observed explanatory variables.
F a set of unobserved explanatory variables.
| separate response | from explanatory variables.

GLM - regression [Y | **X**] one response
case . .. one or more explanatory on ratio scale
case . ..
case . ..

GLM - ANOVA [Y | **X**] one response
case . .. one or more explanatory on nominal scale
case . ..
case . ..

GLM - ANCOVA [Y | **X**] one response
case . .. one or more explanatory on ratio scale
case . .. one or more explanatory on nominal scale
case . ..

GLM - MANOVA [**Y** | **X**] two or more response variables
case .. .. one or more explanatory on nominal scale
case .. ..
case .. ..

GLM - MANCOVA [**Y** | **X**] two or more response variables
case .. .. one or more explanatory on nominal scale
case .. .. one or more explanatory on ratio scale
case .. ..

Relation to GLM (continued)

Eigen analysis is used to compute new or unobserved explanatory variables. These are computed, and then added to the matrix.

Correlation		[**Y** | F]	two or more response variables
	case	.. .	one unobserved explanatory variable F
	case	.. .	
	case	. .	

Discriminant		[**Y** | X F]	two (rarely more) response variables
analysis	case	.. 0 .	one observed explanatory variable X
	case	.. 0 .	one unobserved explanatory variable
	case	.. 0 .	
	case	.. 1 .	
	case	.. 1 .	

F is constructed so as to predict X from measurements of **Y**

Principal components		[**Y** | **F**]	three or more response variables
analysis	case	.. ..	two or more unobserved explantory
	case	.. ..	
	case	.. ..	

Factor analysis		[**Y** | **F**]	three or more response variables
	case	.. ..	two or more unobserved explantory
	case	.. ..	
	case	.. ..	

Same as principal components, except uses different set of rules to construct **F**actor matrix.

Relation to GLM (continued)

Canonical correlation	[$\mathbf{YA}_1$ | $\mathbf{FA}_1$]	three or more response variables
case	.. ..	two or more unobserved explantory
case	.. ..	
case	.. ..	
	[$\mathbf{YB}_1$ | $\mathbf{FB}_1$]	three or more response variables
case	.. ..	two or more unobserved explantory
case	.. ..	
case	.. ..	
	[$\mathbf{FA}_1$ $\mathbf{FB}_1$ | $\mathbf{F}_2$]	

Two sets of factors are correlated with each other via $\mathbf{F}_2$

Cluster analysis	[	Y_1	Y_2	Y_3	Y_4	Y_5	Y_6 ..]	many response variables.
	case	.	.	.	.	.		
	case	.	.	.	.	.		
	case	.	.	.	.	.		
	case	.	.	.	.	.		

Draw dendritic pattern to the left, connecting cases in hierarchical fashion.
No explanatory variables. Response variables grouped according to similarity across variables.

The following diagrams from Van de Geer show the relation between the multivariate methods described in that book.

Draw Figures 9.1 thru 9.8 of Van de Geer p90 on board.
to label these figures, use
Y_1 ... Y_6 and F_1 ... F_4

Fig	Van de Geer
L28a	9.1
L28b	9.2
L28c	9.3
L28d	9.4
L28e	9.5
L28f	9.6
L28g	9.7
L28h	9.8

Multivariate Analysis. Strategy

1. Define quantities
2. Separate response and explanatory (box and arrow diagrams help)
3. State reason for exploratory approach (rather then confirmatory
4. Decide on technique (specific procedures, screening criteria, etc)
5. Execute the analysis.
   a. Form matrix of variables (columns) and cases (rows)
      symbols **Y** = observed response variables
      X = observed explanatory variables
      not known in canonical analysis. Instead use:
      F = unobserved explanatory variables
      b = set of parameters relating **Y** to **F**
      E = residual, or unique, or unexplained variance.
   b. Reduce the variables (Y) to smaller number of factors F
      (unobserved variables)
   c. Examine degree of reduction (variance explained in statistical sense)
6. Interpret the axes in terms of
   -variables Y
   -cases

Tactics. Explained in detail in texts

See list of refence books in Handout Mvar.ref

5b. Reduce the variables **Y** to fewer axes **F**. Variety of techniques available

Canonical methods-use linear algebra to obtain "best" axes according to minimization criteria. There are several
   least dispersion from axis
   capture the most variance

Several names for Canonical Analysis, each referring to diffent tactics.
   Principle coordinates (includes "Factor analysis")
   Principle components

simple correlation is a special case of Canonical Analysis

5c. Degree of reduction. Several criteria, many related to concept of explained variance.
   How much by first axis, by second axis ?
   Rate of increase in explained variance by adding new axis

6. Interpret axes. Rotations and other methods to bring out pattern

Tips on execution
cross-validation of packages.
cross validation against known cases-is this package
doing what I think it is doing ?
matrix check sometimes possible, computations can be written
as a series of matrix equations. NTSYS of Rohlf etc is
an interesting hybrid of black box and matrix approach
concordance of symbols
Plethora of confusing notation.
Eg van de Geer:
X = Y F' + E X=observed
Y=scores or hypothetical variable
F'=loading (parameters)
better to introduce new symbol for unexplained, call this F
try to adhere to following notational conventions:
Y = response variable, observed
X = explanatory variable, observed
F = explanatory variable, unobserved
α ß etc = parameters.
E = residual, or unique, or unexplained variance

Leads to problems in carrying through an analysis.
Is this the same thing that Smith (1950) did ?
Solutions are to
develop concordance of symbols,
develop equivalences in naming (find synonymies and subsets)
1. Reduce dimensions [?] [Y] = [F]
"Principle Component Analysis"
2. Find set of composite variables.
"Principle Coordinate Analysis"
includes "factor analysis"

cross-verification through computation.
Compare results from different packages for example.
Compare results to known or textbook examples.

Work through several examples, using steps 1-6 listed above.
Focus on interpretation of axes, rather than execution.

I. Data from intertidal zone on
Wave energy E, Env. temperature T, Food Intake I,
Growth rate G, and per capita fecundity r_o

Assign r_o to **Y** matrix, others to **X**, do multiple regression

Re-do as path analysis. Emphasizing use of logical relations
to simplify the diagram.

II. Morphometric data on bones.

Set up as factor analysis.
First factor is size, second factor is shape.

III. Data on abundance in of 8 species of tree in 10 plots.

Set up as factor analysis

IV. Data on abundance of seals hauled out on rocks at 3 location
with measurements of tide stage, wind speed,
air temperature, sky cover, wave intensity, and
disturbance at each location.

Set up as canonical correlation, show loadings on
first two canonical variates, with interpretation.

V. Amoco Cadiz, effects of oil spill. From Clarke 1993

Fig L28i = Fig 3 in Clarke

A good example, axes are readily interpreted.
No need to interpret factors in terms of loadings, as is usually the case for
multivariate analysis.

1. **Define quantities.**
2. **Identify explanatory and response variables.**
   Box and arrow diagram
3. **Rationale for exploratory analysis.**
4. **Criterion and Procedure**
5. **Execution**
6. **Interpret axes**. Report results

Steps. (from 1990). Puts too much emphasis on execution,
but not enough detail for execution of analyses.

1. Data
2. Label variables, using mnemonic symbols.
3. Assign variables to a matrix
   Y = observed response variable
   X = observed explanatory variable
   F = unobserved explanatory variable
4. Draw box and arrow diagram of relation of variables.
5. Correlation matrix of variance/covariance matrix of **Y**
   Typically based on columns called variables.
   But in some cases the rows can be considered
   variables, and then the matrix can be transposed
   to analyze rows as variables
   (former columns are now cases) Horrible jargon Q R
6. Extract **F** matrix (and parameters lambda) from **Y** matrix
   Constraints and assumptions needed here.
   Commonest rules are
   that **F** variables explain maximum variance.
   that **F** variables are not correlated with each other.
   Show this graphically by drawing cloud of point on plane,
   fitting line, rotating to explain max variance, then
   fitting second line at right angles,
   this lie to explain max residuals
7. Reduce dimensions of **F** according to some rule. For example
   the first 2 dimensions extracted often explain most of the
   variance in a data set.
8. Plot variables or cases against **F**. Several methods available
   here. Common ones are to
   plot column of **Y** as score on new axes **F**
   plot each row as a correlation with new axes **F**
9. Interpret reduced set of variables **F**.
   Label points.
   Interpret axes **F** relative to columns, relative to rows.
   Look for separation, examine identity of end points on **F**
   Look for groupings and interpret.
   Look for patterns such as arches. These may need to
   be extracted. Seasonal trends, for example produce
   arches because they are not monotonic