NAG®

GENSTAT

Newsletter

Issue No. 27

Editors

P W Lane
AFRC Institute of Arable Crops Research
Rothamsted Experimental Station
HARPENDEN
Hertfordshire
United Kingdom AL5 2JQ

G W Morgan
NAG Limited
Wilkinson House
Jordan Hill Road
OXFORD
United Kingdom OX2 8DR

Printed and produced by The Numerical Algorithms Group Limited

ISSN 0269-0764

Please note that the cover of this Newsletter has been adapted by kind permission of Oxford University Press, from the cover of the Genstat 5 Reference Manual.

NAG® *ORDER FORM*
Genstat Newsletter

The Genstat Newsletter is published twice a year. A two year subscription costs £16 and the price includes surface postage and packing, although airmail is available for an extra £4 per subscription. Renewal forms will be sent every two years with the last issue ordered. Back issues of the Newsletter are sold subject to availability.

To order the Genstat Newsletter, please complete this form and return it with your payment to:

The Services Group
NAG Ltd
Wilkinson House
Jordan Hill Road
OXFORD
United Kingdom OX2 8DR

Note: The representatives of Genstat sites which have annual licences or subscribe to the optional Personal Computer Support Service are automatically sent one copy of the latest issue at no extra charge.

Please supply four issues of the Genstat Newsletter (two year subscription), commencing with the current issue (No. 27), next issue (No. 28), or Issue No. ..... †.

[ ] copies @ £16 surface mail/@ £20 airmail † Total fee £ .....

Is this order to renew a previous subscription? Yes/No †

Please supply back copies of the Genstat Newsletter as follows:

[ ] copies of issues 1–6	@ £8	£	[ ] copies of issue 19	@ £4	£ .....
[ ] copies of issues 7–8	@ £8	£	[ ] copies of issue 20	@ £4	£ .....
[ ] copies of issues 9–10	@ £8	£	[ ] copies of issue 21	@ £4	£ .....
[ ] copies of issues 11–12	@ £8	£	[ ] copies of issue 22	@ £4	£ .....
[ ] copies of issues 13–14	@ £8	£	[ ] copies of issue 23	@ £4	£ .....
[ ] copies of issue 15	@ £4	£	[ ] copies of issue 24	@ £4	£ .....
[ ] copies of issue 16	@ £4	£	[ ] copies of issue 25	@ £4	£ .....
[ ] copies of issue 17	@ £4	£	[ ] copies of issue 26	@ £4	£ .....
[ ] copies of issue 18	@ £4	£	[ ] copies of issue 27	@ £4	£ .....

Please invoice / Total payment enclosed (pounds sterling) † £ .....

Payment must be received with the order form, otherwise a £5 invoicing surcharge will be added. Cheques (in pounds sterling) should be made payable to:
The Numerical Algorithms Group Ltd.

Name: .....

Address: .....

.....

Your Order Number (if any): .....

Signed: ..... Date: .....

Genstat Newsletter
Issue No. 27

NP2288

November 1991

Contents

Published Twice Yearly by
Rothamsted Experimental Station Statistics Department
and The Numerical Algorithms Group Ltd

Editorial

In Newsletter 26 we announced that Professor Vic Barnett had been appointed as the Head of the Statistics Department at Rothamsted; as promised, this edition starts with an article in which he gives his views on Genstat. The second article lists the new procedures in Procedure Library 2[2], which is, or soon will be, available for implementations of Release 2 of Genstat. For those writing procedures there is sometimes a need to refer to more information about a regression model than can be obtained using RKEEP; the third article gives details of the regression save structure which stores this information. The main part of this Newsletter consists of articles describing how Genstat can be adapted for particular applications. These are: fitting digit-preference models, fitting discrete-distribution models by maximum entropy, the provision of extra output from a principal components analysis, and minimizing a general function using the non-linear regression directives. Finally, there is the first of a series of articles based on talks given at the Seventh International Genstat Conference; this article describes techniques for dealing with collinear data.

Genstat News

A preliminary version of Release 3 of Genstat was demonstrated at the Genstat Conference. Some of the significant new features on display were: a new directive to estimate parameters in a wide range of distributions, the extension of the regression section to fit generalized additive models and ordinal response models, the combination of information in generally balanced designs, and the testing of fixed effects in analyses by residual likelihood (REML). Further details about Release 3 will be given in a subsequent edition of the Newsletter.

Implementation News

Since the last Newsletter, the implementations of Release 2 for SUN 4, HP 9000/800, IBM RS 6000 and Sequent Symmetry have become available; those for Data General MV and ICL 2900/3900 VME should be available shortly. Staff at NAG are working hard to finish the delayed implementation for IBM 386 PCs, which will include a split screen interface; this implementation should be available in the near future.

Genstat Courses

The next Genstat Introductory Course will take place in March or April in Southampton. We are sending out a short questionnaire to find out the types of courses that are best suited to your needs. We are looking at several possible arrangements and we would be grateful for your help in selecting the most suitable options. For details of Genstat Courses please contact Lesley Austen at NAG.

Genstat – A Tool for Professionals

V Barnett
AFRC Institute of Arable Crops Research
Rothamsted Experimental Station
Harpenden
Hertfordshire
United Kingdom AL5 2JQ

It is a great pleasure for me to have recently taken over the reins of the Biomathematics Division at IACR, including the Statistics Department at Rothamsted with its line of impressive heads of department who contributed so much to the development of the subject. Not least in this heritage is of course Genstat, which is so firmly placed within the international panoply of professional statistical software systems.

Genstat has come of age – it reached its 21st birthday this year (or 18th in 1988, depending on your national view of the age of maturity). As with any adult member of society, we have reason to expect much of it. Indeed much is planned – with constant refinement of the basic system, an impressive and ever-growing procedure library, further menu-driven options, enhancement of graphics facilities and intercommunicability with other systems, and timely developments into new hardware (and operating) environments. I shall certainly do all that I can to promote it and to keep it at the forefront of service as a fully professional system, providing a modern computational environment for the statistician and for the research worker across all the disciplines that use statistical methods.

The watchwords must continue to be **professionalism** and **accessibility**. Genstat has to seek to cover all the statistical needs of its various groups of users, but it must do so in a user-friendly manner, which does not deter those who do not see themselves as professional statisticians *per se*.

Genstat looked very different, of course, 21 years ago when John Nelder began to develop his impressive symbolic-algebraic approach to the categorisation and representation of experimental designs. When he implemented these in computer-algorithmic terms, a unique facility was born. The statistician, or statistically-able experimentalist, could at last instruct the computer in a few relatively simple statements to carry out the analysis of highly complicated and complex designed experiments. Initially, these could be crossed or nested but had to be of balanced and complete form. Progressively the frontiers were extended and the statistical analyst was liberated from the labour and tedium of the split-plot, or lattice design, or one sixty-fourth replicate of a 2^{14} experiment in 16 blocks. As someone who started work in a research station with a room full of young girls doing Yates' algorithm hour after hour on large sheets of paper, I am well placed to understand the computational revolution wrought by the Nelder approach and its implementation in the first form of the Genstat system.

Of course the system originally contained more than the analysis of designed experiments: the Manual of 1970 shows sections for regression, classification and multivariate analysis as well as analysis of variance. The system also rapidly expanded to other areas, and continues to do so. The developments of Genstat in the Statistics Department at Rothamsted continued an already long-established tradition of statistical computing which John Gower in these pages (in 1984) rightly attributed to the influence of Frank Yates in the early 1950s. John remarks of Genstat that, by 1984, it 'has changed enormously, by the addition of new statistical features, developments in the language, the provision of basic graphics, etc. Many of these extensions have been of an *ad hoc* nature, some opportunistic. The upshot is a language with many inconsistencies which make it difficult to learn, teach and remember – hence the Genstat 5 revision, whose specification is complete and whose implementation is in progress. The PROCEDURE facility in particular, with its similarity to the specification of directives, will make the new Genstat even easier to extend than is the current version and, hence, much increase the power of the language in providing new statistical facilities.'

Genstat has indeed faced up to and solved the majority of these problems and shortcomings in the last six years in a most impressive way, and we are now poised for further refinements in terms of the new-generation prospects I have described above.

I have only one minor reservation ... a feeling that, in spite of sound credentials and a glowing future, our young adult may still not be quite as well known in the best circles as some of the other members of the 'smart set'. I shall do all I can to remedy this; I am sure that NAG will also do so. As dedicated and (dare I say) largely satisfied users, you can also play your part in introducing Genstat to colleagues. Will you? After all, why should we keep the good news to ourselves?

Genstat Procedure Library News

R W Payne
AFRC Institute of Arable Crops Research
Rothamsted Experimental Station
Harpenden, Hertfordshire
United Kingdom AL5 2JQ

In the first two revisions of the Procedure Library for Genstat 5 Release 2, 30 new procedures have been added. The index lines of the new procedures are as follows.

Release 2[1]

AKAIKEHISTOGRAM	prints histograms with improved definition of groups
BINOMIAL	calculates probabilities from the binomial distribution
BSUPDATE	creates or updates a backing-store subfile
DECIMALS	sets the number of decimals for a structure, using its round-off
DILUTION	calculates Most Probable Numbers from dilution series data
FITPARALLEL	carries out analysis of parallelism for non-linear functions
FITSCHNUTE	fits a general 4 parameter growth model to a non-decreasing Y-variable
FTEXT	forms a text structure from a variate
GINVERSE	calculates the generalized inverse of a matrix
INVNORMAL	calculates probabilities from the inverse normal distribution
LINDEPENDENCE	finds the linear relations associated with matrix singularities
LOGNORMAL	calculates probabilities from the lognormal distribution
MENU	initiates a menu system
PAIRTEST	performs *t*-tests for pairwise differences
POISSON	calculates probabilities from the Poisson distribution
RANK	produces ranks, from the values in a variate, allowing for ties
RPAIR	gives *t*-tests for all pairwise differences of means from regression/GLM
STUDENT	calculates probabilities from Student's *t*-distribution
VFUNCTION	calculates functions of variance components from a REML analysis
VPLOT	plots residuals from a REML analysis

Release 2[2]

ABIVARIATE	produces graphs and statistics for bivariate analysis of variance
AUDISPLAY	produces further output for an unbalanced design
AUNBALANCED	performs analysis of variance for unbalanced designs
LVARMODEL	analyses a field trial using the Linear Variance Neighbour model
MPOWER	forms integer powers of a square matrix
NORMTEST	performs tests of univariate and/or multivariate normality
PROBITANALYSIS	fits probit models allowing for natural mortality & immunity
SAMPLE	samples from a set of units, possibly stratified by factors
VTABLE	forms a variate and set of classifying factors from a table
WADLEY	fits models for Wadley's problem, allowing alternative links & errors

Fifteen of these are from British authors, thirteen are from the Netherlands, and two from New Zealand.

There are also minor amendments to help information or source code of several of the existing procedures. In particular, in Release 2[2], the procedures LIBINFORM and LIBMANUAL have been modified to allow procedures to be allocated to more than one module (for example to 'bioassay' as well as to 'glm', and so on). Releases 2[2] onwards will also have three new modules:

NEWHELP	to indicate the procedures whose help information has changed since the last release of the library
NEWSOURCE	to indicate those whose source code has been modified to improve efficiency or correct errors
NEW	to denote additions in the current release.

Thus, to find out the procedures whose help information has been updated, you can type:

```
LIBINFORM [CONTENTS] 'NEWHELP'
```

In addition, you can use procedure NOTICE to check on other Genstat news.

The Regression Save Structure

P W Lane
AFRC Institute of Arable Crops Research
Rothamsted Experimental Station
Harpenden
Hertfordshire
United Kingdom *AL5 2JQ*

The regression save structure (or RSAVE structure) is a system structure that stores information about a regression model. It is used implicitly by all the regression directives, which access information from it and store results into it. For example, the RKEEP directive is designed to extract many components of the structure, such as fitted values and the residual sum of squares, into simpler Genstat structures like variates and scalars. The intention is that most users of Genstat will be able to extract all the information they require by using RKEEP: therefore they need know nothing more about the RSAVE structure.

However, there is much more information stored in the RSAVE structure, which can be particularly useful to writers of procedures. It is possible to extract such information as the identifier of the response variate, or the code of the link function in a generalized linear model. For examples of use within a procedure, look at the source of the procedure RCHECK that is in the Procedure Library. The source can be displayed by the statements:

```
LIBEXAMPLE 'RCHECK'; SOURCE=tt
PRINT tt; JUSTIFICATION=left; SKIP=0
```

The constituent structures of the RSAVE structure are detailed below. Any of these can be accessed in a Genstat program just like the constituent structures of pointers. For example, to access the identifier of the response variate (or the first response variate if there are several), use the identifier `r[2][1][1]` in Release 2 of Genstat, where `r` is the identifier given to the RSAVE structure in the MODEL statement. Thus

```
MODEL [SAVE=r] Y=yield
PRINT r[2][1][1]
```

has the same effect as:

```
PRINT yield
```

If the RSAVE structure has not been explicitly named, it can be accessed by the GET directive:

```
GET [SPECIAL=s]
PRINT s['rsave'][2][1][1]
```

Some of the constituent structures that are system-defined are also non-standard, and cannot be used directly in most Genstat directives. However, the attributes and contents of all the structures can be inspected with the DUMP directive; for example,

```
DUMP [PRINT=attributes,values] r[4][1]
```

will show the attributes and values of a special long real structure. The PRINT directive will not display values of long real structures like this, but will display system structures that store integer values; for example,

```
PRINT r[1][3]
```

will show the values of a special integer structure.

The CALCULATE and EQUATE directives can be used with system structures that have real or integer values; for example,

```
SCALAR Tdf
CALCULATE Tdf = r[1][3]$[6]
```

will set up `Tdf` as a scalar with the total number of degrees of freedom.

The RSAVE structure is subject to variation between releases of Genstat. Usually, the only changes will be additions, and that is the case between Releases 1.3, 2.1 and 2.2.

[1] Common pointer

[1][1] Double precision variables

1. (Not used)

[1][2] Real variables (variate)

$[1]	RDISPR	dispersion parameter; from MODEL [DISPERSION]; * if unset, or 1 for Poisson, binomial, multinomial
$[2]	RPOWER	power for 'power' link function; from MODEL [EXPONENT]; * if unset, or –2 for default power link, –1 for reciprocal link, 0.5 for square root
$[3]	RCNVCR	convergence criterion; from RCYCLE [TOLERANCE]; 0.0004 if unset
$[4]	RDVRES	Residual deviance from GLM or nonlinear model; * if unset
$[5]	RORIGN	origin for curves; from FITCURVE [ORIGIN]; * if unset
$[6]	RSWPCR	criterion for aliasing; from TERMS [TOLERANCE]; if unset, 10*EPS for mixed precision implementations, 10000*EPS for double precision implementations

[1][3] Integer variables (special integer structure)

$[1]	KDIST	Count of distribution; from MODEL [DISTRIBUTION]; 1 Normal, 2 Poisson, 3 binomial, 4 gamma, 5 inverse Normal, 6 multinomial
$[2]	KLINK	Count of link function; from MODEL [LINK]; 1 identify, 2 log, 3 logit, 4 reciprocal, 5 power, 6 square root, 7 probit, 8 complementary log-log
$[3]	KLINKP	Recoded count of link function including special powers; as KLINK except for LINK 5: 5 negative integer, 6 positive inverse integer, 9 positive or negative real, 10 positive integer, 11 negative inverse integer
$[4]	MCYCLE	Maximum number of cycles, from RCYCLE [MAXCYCLE]; * if unset, interpreted later as 10 for GLM, 20 for nonlinear
$[5]	KMETH	Count of optimizing method, from RCYCLE [METHOD]; 1 Gauss-Newton, 2 Newton-Raphson, 3 Fletcher-Powell (3 introduced in Release 2)
$[6]	NDFTOT	Number of total degrees of freedom in maximal model
$[7]	NDFRES	Number of residual degrees of freedom in current model
$[8]	NTRCM	Number of terms in current model
$[9]	NALPR	Number of aliased parameters in current model
$[10]	NRWSS	Number of rows in working DSSP
$[11]	NTRSS	Number of terms in maximal model
$[12]	LRWPR	Length of row-parameter assignment array
$[13]	LTRBF	Length of term-buffer assignment array
$[14]	JCONST	Whether a constant term is estimated, from CONSTANT option; 0 no constant, 1 constant
$[15]	NENTRY	Number of entries in accumulated summary
$[16]	LACCUM	Length of accumualted summary
$[17]	NDFCH	Number of degrees of freedom associated with last change to model
$[18]	JLIK	Code for likelihood calculation; from MODEL [DISTRIBUTION]; 1 explicit optimization, 2 Normal no linear, 3 Normal with linear, 4 Poisson no linear, 5 Poisson with linear, 6 binomial, 7 multinomial, 9 gamma, 10 inverse Normal
$[19]	KCURVE	Count of curve-type, from FITCURVE [CURVE]; 1 exp, 2 dexp, 3 cexp, 4 lexp, 5 log, 6 glog, 7 gomp, 8 ldl, 9 qdl, 10 qdq
$[20]	JSENSE	Code for sense of curve; from FITCURVE [SENSE]; 1 right, 2 left
$[21]	NPARN	Number of nonlinear parameters

$[22]	NPARSE	Number of parameters for which standard errors have been calculated
$[23]	NPARTT	Number of parameters, linear and nonlinear
$[24]	JSEPNL	Whether to fit separate nonlinear parameters; from NONLINEAR option; 1 no, 2 yes
$[25]	JEXIT	Code for exit status; as accessed by RKEEP EXIT; 0 success, 1 limit on number of cycles, 2 out of bounds, 3 constant function, 4 failure to progress, 5 no s.e. due to singularity, 6 limiting form
$[26]	NGRID	Number of gridlines; from FITNONLINEAR [NGRIDLINES]; * if unset
$[27]	JRESID	Code for type of residuals; from MODEL [RMETHOD]; 0 none, 1 deviance, 2 Pearson

[2] Model pointer

- [2][1] Pointer to response variates (*y*-variates)
  Release 1: if there is 1 *y*-variate, [2][1] is its identifier
  Release 2: pointer always formed, even if only 1 *y*-variate
- [2][2] Variate of binomial totals; from MODEL NBIN
- [2][3] Variate of weights; from MODEL [WEIGHTS]
- [2][4] Variate of offsets; from MODEL [OFFSET]
- [2][5] Factor for grouping; from MODEL [GROUPS]
- [2][6] Variate of initial fitted values; from RCYCLE [FITTED]
- [2][7] Scalar to store function value; from MODEL [FUNCTION]

[3] Output pointer

- [3][1] Pointer to residual variates; named if in MODEL RESID
  Release 1: if 1 *y*-variate, [3][1] is identifier of residual variate
  Release 2: pointer always formed, even if only 1 *y*-variate
- [3][2] Pointer to fitted-values variates; named if in MODEL FITTED
  Release 1: if 1 *y*-variate, [3][2] is identifier of fitted-values variate
  Release 2: pointer always formed, even if only 1 *y*-variate
- [3][3] Variate of leverages; as accessed by RKEEP LEVER
- [3][4] Variate of accumulated model summaries
- [3][5] Pointer to gradient variates; as accessed by RKEEP GRAD
- [3][6] Variate of grid values; as accessed by RKEEP GRID
- [3][7] Linear predictor variate; as accessed by RKEEP LINEAR (Not in Release 1)
- [3][8] Iterative weights variate; as accessed by RKEEP ITER (Not in Release 1)
- [3][9] Variate storing working vector in GLM calculations (Not in Release 1)
- [3][10] Text storing parameter labels (Not in Release 1)

[4] Working DSSP (special pointer structure)

- [4][1] Double-precision working matrix (special long real structure)
- [4][2] Double-precision group means for response variate
- [4][ ] Double precision group means, one for each term

[5] GLM pointer (all sub-structures are special integer structures)

- [5][1] Positions of *y*-variates in variate set (special integer structure)
- [5][2] Positions of *y*-variates in maximal model (special integer structure)
- [5][3] Positions of current model terms in maximal model (special integer structure)
- [5][4] Indicators of working matrix rows: 0 unswept, 1 swept (special integer structure)
- [5][5] Row-parameter assignment array for maximal model (special integer structure)

[5][6] First and last rows for maximal model (special integer structure)

[5][7] Row numbers of *y*-variates in SSPM (special integer structure) (Not in Release 1)

[6] Linear model pointer

[6][1] Variate of original diagonal values of SSPM

[6][2] Variate of original means of SSPM

[6][3] Variate of latest diagonal values of working matrix for *y*-variates

[7] Cycle pointer

[7][1] Pointer of parameters; from `RCYCLE PARAMETER`

[7][2] Variate of lower bounds; from `RCYCLE LOWER`

[7][3] Variate of upper bounds; from `RCYCLE UPPER`

[7][4] Variate of initial steps; from `RCYCLE STEP`

[7][5] Variate of initial values; from `RCYCLE INITIAL`

[8] Nonlinear pointer

[8][1] Current parameter values (special long real structure)

[8][2] Working steplengths (special long real structure)

[8][3] Inverse matrix (special long real structure)

[8][4] *X*-variate for `FITCURVE` or pointer to parameter names for `FITNONLINEAR`

[8][5] Factor for `FITCURVE`

[8][6] Compiled-code structure for calculations from `FITNONLINEAR [CALCULATION]`

Fitting Digit-Preference Models to Fecundability Data

M S Ridout
Horticulture Research International
East Malling
West Malling
Kent
United Kingdom *ME19 6BJ*

1. Introduction

Fecundability is defined as 'the monthly probability of conception in the absence of contraception, outside the gestation period and the temporary sterile period following the termination of a pregnancy' (Jain [1]). Fecundability can be studied by observing the number of cycles needed to achieve pregnancy by couples who are actively trying to conceive. Such data are often collected retrospectively, during or after pregnancy. One problem with retrospective data is that the number of cycles may be misreported; in particular digit-preference often occurs, with unexpectedly high frequencies of couples reporting that pregnancy occurred in the 6th, 12th, 18th, ... month/cycle. Similar patterns of digit-preference occur in other types of retrospectively collected duration data, for example duration of breast-feeding (Diamond *et al* [2] or of unemployment Torelli and Trivellato [3]).

Ridout and Morgan [4] fitted simple digit-preference models to two sets of fecundability data using the NAG Library optimization routine E04UCF. This note shows how these models, and some that are more complex, can be fitted using the `FITNONLINEAR` command in Genstat.

The techniques developed may be useful in other situations where a model is to be fitted to several data sets and one wants to be able to constrain parameters to be equal for some of the data sets.

2. Data

Let X denote the number of cycles to conception and suppose that the data consist of the numbers of couples reporting values of X = 1,2,...,12, together with the number reporting values of $X > 12$. This grouping of values greater than 12 is common in fecundability work, because medical intervention often occurs if conception is not achieved within one year. For reasons that will become clear later, the frequencies of X = 1,2,...,12 are input in a variate of length 17, whose last 5 values are zero. The variates must be called `Freq[1...NSets]` where `NSets` is the number of sets of data being studied. Tail frequencies (i.e. of values $X > 12$) are similarly entered in scalars `TailFreq[1...NSets]`.

The four data sets used in [4] (from two studies, each comparing two types of individual) are entered as follows:

```
UNIT [17]
CALCULATE NSets = 4
READ [SERIAL=yes] Freq[1...NSets]
  29  16  17   4   3   9  4  5  1  1 1   3  0 0 0 0 0 :  Smokers
 198 107  55  38  18  22  7  9  5  3 6   6  0 0 0 0 0 :  Non-smokers
 383 267 209  86  49 122 23 30 14 11 2  43  0 0 0 0 0 :  Pill users
1674 790 480 206 108 263 54 56 21 33 8 130  0 0 0 0 0 :  Non-pill users
SCALAR TailFreq[1...NSets]; VALUE=7,12,35,191
CALCULATE TotFreq[1...NSets] = SUM(Freq[]) + TailFreq[]
```

The last line evaluates the total number of couples in each data set.

3. Models

The models are specified in two parts. The first part specifies the distribution of X in the absence of misreporting, whilst the second part specifies the way in which misreporting occurs. The two parts are combined to give a model that can be fitted to observed data.

3.1. Distribution of X in the Absence of Misreporting

We assume that each couple has fecundability p that remains constant over time. The number of cycles to conception for that couple is then geometric with parameter p. However, fecundability

varies between couples and we assume that p has a beta distribution with mean μ. Then X has a beta-geometric distribution with

$$Pr(X=k) = \frac{\mu \prod_{i=1}^{k-1} [1-\mu+(i-1)\theta]}{\prod_{i=1}^{k} [1+(i-1)\theta]}$$

where $\prod_{i=1}^{0} = 0$ and θ is a parameter related to the variance of the beta distribution ($\theta = 0$ gives the ordinary geometric distribution).

Weinberg and Gladen [5] reviewed the use of the beta-geometric distribution in fecundability work and showed how the distribution can be fitted as a binomial generalized linear model, albeit with a non-standard (reciprocal) link function. This could be implemented in Genstat, but it is easier, and probably more efficient, to use FITNONLINEAR. The real advantage of the generalized linear model formulation is that one can introduce into the model covariates that may affect fecundability (e.g. type of previous contraceptive use). But unfortunately there seems to be no simple way of modifying this approach to incorporate digit-preference.

3.2. Digit-Preference Models

A simple digit-preference model is as follows:

(i) When the true value of X is 6 or 12, this is always reported correctly.

(ii) When the true value of X is 1,2,3,4 or 5, this is either misreported as 6, with probability ϕ_1^{6-X}, or is reported correctly.

(iii) When the true value of X is 7,8,9,10 or 11, this is either misreported as 6, with probability ϕ_2^{X-6}, or misreported as 12, with probability ϕ_3^{12-X}, or it is reported correctly.

(iv) When the true value of X is 13,14,15,16 or 17, this is either misreported as 12, with probability ϕ_4^{X-12}, or is reported correctly.

Thus, the true number of cycles may, if it is not already a multiple of 6 cycles, be misreported forwards or backwards to the nearest multiple of 6 cycles. The different ϕ-parameters allow one to model, for example, a greater likelihood of misreporting to 12 cycles than to 6 cycles and/or a greater likelihood of misreporting backwards than forwards. Ridout and Morgan [4] considered models with $\phi_1 = \phi_2 = \alpha$ and $\phi_3 = \phi_4 = \beta$ (their model 2), and also with $\alpha = \beta$, i.e. $\phi_1 = \phi_2 = \phi_3 = \phi_4$ (their model 3).

4. Fitting the Models

4.1. Preliminaries

Before fitting any models there are some commands that need to be executed. These are stored in a file INIT.INP and executed using an INPUT command:

```
OPEN 'init.inp'; CHANNEL=2
INPUT [PRINT=*] 2
CLOSE 2
```

A listing of INIT.INP is given in the Appendix.

4.2. Specifying the Model

Models can be fitted to some or all of the datasets. A variate VSet is declared to indicate which datasets are to be included. For example,

```
VARIATE VSet; VALUES=!(1,3)
```

indicates datasets 1 and 3.

The INIT.INP file defines separate parameters μ and θ for each data set, called Mu[1...NSets] and Theta[1...NSets]. Variates MSet and TSet must be declared to specify which of these parameters are involved in the current model. Since we are using datasets 1 and 3, we could use

```
VARIATE MSet,TSet; VALUES=!(1,3),!(1,3)
```

This will result in separate values of μ and θ being fitted to the two datasets.

However, the values of MSet and TSet need not be distinct, and this allows one to specify models in which some parameters are constrained to be equal for some data sets. For example

```
VARIATE MSet,TSet; VALUES=!(1,3),!(1,1)
```

specifies a model in which there is a separate parameter μ for each data set, but a single common parameter θ (which will be labelled Theta[1] in the output).

Similarly, for example,

```
VARIATE VSet,MSet,TSet; VALUES=!(1,2,4),!(1,1,4),!(1,2,1)
```

indicates a model involving data sets 1,2 and 4 with parameters

Mu[1] – data sets 1 and 2
Mu[4] – data set 4
Theta[1] – data sets 1 and 4
Theta[2] – data set 2

After the model is fitted, variates VSet, MSet and TSet are automatically deleted and must be redefined for the next model.

A similar device is used to specify digit-preference parameters. It is assumed that the same digit-preference parameters apply to all data sets included in the fit. There are nominally 4 parameters called Phi[1...4]. In the model the parameters are referred to as Phi[One], Phi[Two], Phi[Three] and Phi[Four] and the values of the scalars One, Two, Three and Four must be assigned values before fitting the model. By assigning the same value more than once, one can impose various constraints, for example

```
SCALAR One,Two,Three,Four; VALUE=1,2,3,4
```
gives four distinct parameters. These will be labelled Phi[1,2,3,4] in the output.

```
SCALAR One,Two,Three,Four; VALUE=1,1,3,3
```
gives $\phi_1 = \phi_2$ and $\phi_3 = \phi_4$. The two parameters will be labelled Phi[1,3] in the output.

```
SCALAR One,Two,Three,Four; VALUE=1,1,1,1
```
gives $\phi_1 = \phi_2 = \phi_3 = \phi_4$. The single parameter will be labelled Phi[1] in the output.

Unlike the variates VSet, MSet and TSet, the scalars One, Two, Three and Four do not need to be redefined after each successive fit (unless, of course, different constraints on the ϕ-parameters are wanted).

The last thing that needs to be done before fitting the model is to use the RCYCLE command to give bounds and initial estimates for the parameters. The bounds required are as follows:

	Parameter	Lower bound	Upper bound
All	μ-parameters	0.000001	0.999999
All	θ-parameters	0	*
All	ϕ-parameters	0	0.75

Strictly, the only constraint on the ϕ-parameters is that the total probability that any true value is misreported should not exceed one. Constraining all ϕ-parameters to be less than 0.75 is convenient and ensures that this constraint is not violated, but it is unnecessarily stringent in that large values of some ϕ-parameters are acceptable provided that other ϕ-parameters are sufficiently small. Therefore, if an estimated ϕ-parameter is equal to its upper bound, the model should be refitted with the upper bound for this parameter increased slightly, but with reduced upper bounds for other parameters. This problem has not so far arisen in practice.

Suggested initial parameter estimates are as follows:

all μ-parameters : 0.35
all θ-parameters : 0.1
all ϕ-parameters : 0.2

For the datasets and models considered here, convergence is usually achieved within 10 iterations from these starting values, though of course if better initial estimates are available these should be used instead.

The commands for fitting the model are stored in a file FITMODEL.INP and can be (repeatedly) accessed with an INPUT command:

```
OPEN 'fitmodel.inp'; CH=2
INPUT [PRINT=*] 2
CLOSE 2
```

A listing of FITMODEL.INP is given in the Appendix.

5. Examples

5.1. Smokers Versus Non-smokers

For both of these data sets, Ridout and Morgan [4] found that the model with $\phi_1 = \phi_2 = \phi_3 = \phi_4$ (their model 3) fitted almost as well as the model with $\phi_1 = \phi_2 \neq \phi_3 = \phi_4$ (their model 2). Both models fitted better than a model with no misreporting. The following code fits their model 3 to each data set in turn:

```
SCALAR One,Two,Three,Four; VALUE=1,1,1,1

VARIATE VSet,MSet,TSet; VALUE=!(1)
RCYCLE Mu[1],Theta[1],Phi[1]; INITIAL = 0.33,0.1,0.2; \
       LOWER = 0.000001,0,0.000001; UPPER = 0.999999,*,0.75
OPEN 'fitmodel.inp'; CH=2    : INPUT [PRINT=*] 2     : CLOSE 2

VARIATE VSet,MSet,TSet; VALUE=!(2)
RCYCLE Mu[2],Theta[2],Phi[1]; INITIAL = 0.33,0.1,0.2; \
       LOWER = 0.000001,0,0.000001; UPPER = 0.999999,*,0.75
OPEN 'fitmodel.inp'; CH=2    : INPUT [PRINT=*] 2     : CLOSE 2
```

The output consists of

(a) Minus two times the maximized log-likelihood function.

(b) Parameter estimates and standard errors.

(c) Correlation matrix of parameter estimates.

(d) Table of observed frequencies and two sets of fitted frequencies. The fitted frequencies in the first column are those that would be expected, given the estimated values of μ and θ, in the absence of digit-preference. The second column gives modified frequencies based on the estimated digit-preference parameters.

(e) Two χ^2 statistics, comparing the two sets of fitted frequencies to the observed frequencies.

When there is more than one data set, items (d) and (e) are given for each data set in turn.

The estimates of the parameter ϕ_1 are quite similar for the two data sets and Ridout and Morgan suggested that a common value of ϕ_1 could be adopted for smokers and non-smokers. This model can be fitted with the commands:

```
VARIATE VSet,MSet,TSet; VALUE=!(1,2)
SCALAR One,Two,Three,Four; VALUE=1,1,1,1
RCYCLE Mu[1,2],Theta[1,2],Phi[1]; \
       INITIAL = 0.33,0.33,0.1,0.1,0.2; \
       LOWER = 0.000001,0.000001,0,0,0.000001; \
       UPPER = 0.999999,0.999999,*,*,0.75

OPEN 'fitmodel.inp'; CH=2
INPUT [PRINT=*] 2
```

Minus two times the maximized log-likelihood for this model is 2207.8. The corresponding values for fitting a separate value of ϕ_1 to the two data sets are 431.3 and 1775.4. The log-likelihood-ratio statistic for testing the hypothesis that ϕ_1 is the same for both data sets is therefore

$$2207.8-(431.3+1775.4))=1.1$$

which, with 1 degrees of freedom, is clearly not significant.

5.2. Pill-users Versus Users of Other Contraceptives

The other data set studied in [4] compared fecundability of couples where the woman had previously used the pill with fecundability of couples who had used other types of contraceptive. This data set shows much stronger digit-preference and the following modified digit preference model was considered.

(i) When the true value of X is 1,2,3,6 or 12, this is always reported correctly.

(ii) When the true value of X is 4 or 5, this is either misreported as 3, with probability ϕ_5^{X-3}, or misreported as 6, with probability ϕ_1^{6-X}, or it is reported correctly.

(iii) As in Section 3.2.

(iv) As in Section 3.2.

Only minor changes are required to fit this model. First, the following commands should be added to `INIT.INP`, immediately before the `RETURN` statement:

```
VARIATE [NVALUES=17] To3[1...NSets],V3a,Wt3a
READ [SERIAL=yes; PRINT=*] V3a,Wt3a,Wt6b
 0  0  0  1  2  0  0  0  0  0  0  0   0  0  0  0  0 :   V3a
 0  0  0  1  1  0  0  0  0  0  0  0   0  0  0  0  0 :   Wt3a
 0  0  0  1  1  0  0  0  0  0  0  0   0  0  0  0  0 :   Wt6b
```

Secondly, in the file `FITMODEL.INP`, expressions are modified as follows:

(i) Replace

```
EXPRESSION model[1...6]; VALUE= \
```

by

```
EXPRESSION model[1...7]; VALUE= \
```

(ii) Replace

```
!E( Prob[#VSet] = TrueProb[#VSet] - \
                  To6[#VSet]  + SUM(To6[#VSet])  * (MCycle==6) - \
                  To12[#VSet] + SUM(To12[#VSet]) * (MCycle==12) ), \
```

by

```
!E( To3[#VSet]  = TrueProb[#VSet] *  Wt3a * Phi[Five] **  V3a), \
!E( Prob[#VSet] = TrueProb[#VSet] - \
                  To3[#VSet]  + SUM(To3[#VSet])  * (MCycle==3) - \
                  To6[#VSet]  + SUM(To6[#VSet])  * (MCycle==6) - \
                  To12[#VSet] + SUM(To12[#VSet]) * (MCycle==12) ), \
```

Lastly, before fitting the model, the scalar `Five` must be set, in addition to the scalars `One`, `Two`, `Three` and `Four`.

Model fitting is then straightforward, for example the following Genstat code fits a model to the two data sets, with common digit-preference parameters but separate values of the parameters μ and θ.

```
VARIATE VSet,MSet,TSet; VALUE=!(3,4)
SCALAR One,Two,Three,Four,Five; VALUE=1,1,3,3,5
RCYCLE Mu[3,4],Theta[3,4],Phi[1,3,5]; \
       INITIAL = 2(0.33, 0.1), 3(0.2); \
       LOWER = 2(0.000001, 0), 0.000001, 0, 0; \
       UPPER = 2(0.999999, *), 3(0.75)

OPEN 'fitmodel.inp'; CH=2
INPUT [PRINT=*] 2
```

6. References

[1] Jain, A.K.
Fecundability and its relation to age in a sample of Taiwanese women.
Population Studies, 23, pp. 69-85, 1969.

[2] Diamond, I.D., McDonald, J.W. and Shah, I.H.
Proportional hazards models for current status data: application to the study of differentials in age at weaning in Pakistan.
Demography, 23, pp. 607-620, 1986.

[3] Torelli, N. and Trivellato, U.
Youth unemployment duration from the Italian Labour Force Survey: accuracy issues and modelling attempts.
European Economic Review, 33, pp. 407-415, 1989.

[4] Ridout, M.S. and Morgan, B.J.T.
Modelling digit-preference in fecundability studies.
Biometrics, (submitted).

[5] Weinberg, C.R. and Gladen, B.C.
The beta-geometric distribution applied to comparative fecundability studies.
Biometrics, 42, pp. 547-560, 1986.

7. Appendix

Listing of file `INIT.INP`

```
"-------------------------------------------
 File INIT.INP
 Define and initialize various data-structures
 used in fitting fecundability models
 -------------------------------------------"
SCALAR Mu[1...NSets],Theta[1...NSets],LL[1...NSets]; VALUE=1
VARIATE [NVALUES=17] To6[1...NSets],To12[1...NSets], \
        TrueProb[1...NSets],Prob[1...NSets], \
        MCycle,WtProb,V6a,Wt6a,V6b,Wt6b,V12a,Wt12a,V12b,Wt12b
READ [SERIAL=yes; PRINT=*] \
     MCycle,WtProb,V6a,Wt6a,V6b,Wt6b,V12a,Wt12a,V12b,Wt12b
 1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 :   MCycle
 1  1  1  1  1  1  1  1  1  1  1  1  0  0  0  0  0 :   WtProb
 0  0  0  0  0  0  1  2  3  4  5  0  0  0  0  0  0 :   V6a
 0  0  0  0  0  0  1  1  1  1  1  0  0  0  0  0  0 :   Wt6a
 5  4  3  2  1  0  1  2  3  4  5  0  0  0  0  0  0 :   V6b
 1  1  1  1  1  0  0  0  0  0  0  0  0  0  0  0  0 :   Wt6b
 0  0  0  0  0  0  0  0  0  0  0  0  1  2  3  4  5 :   V12a
 0  0  0  0  0  0  0  0  0  0  0  0  1  1  1  1  1 :   Wt12a
 0  0  0  0  0  0  5  4  3  2  1  0  0  0  0  0  0 :   V12b
 0  0  0  0  0  0  1  1  1  1  1  0  0  0  0  0  0 :   Wt12b
CALCULATE MCycle_1 = MCycle - 1
&         Multiplier = MCycle-2+(MCycle.eq.1)
MODEL [FUNCTION=LogLik]
RETURN
```

Listing of file `FITMODEL.INP`

```
"--------------------------------
 File FITMODEL.INP
 Fits various fecundability models
 --------------------------------"
EXPRESSION model[1...6]; VALUE= \
   !E( TrueProb[#VSet] = Mu[#MSet]/(1-Mu[#MSet]) * EXP( CUM( \
                         LOG(1-Mu[#MSet] + Multiplier*Theta[#TSet] ) - \
                         LOG(1 + MCycle_1*Theta[#TSet]) ) ) ), \
   !E( To6[#VSet]  =  TrueProb[#VSet] * (Wt6a * Phi[One] **  V6a + \
                         Wt6b * Phi[Two] **  V6b )  ), \
   !E( To12[#VSet] =  TrueProb[#VSet] * (Wt12a * Phi[Three] **  V12a + \
                         Wt12b * Phi[Four] **  V12b )  ), \
   !E( Prob[#VSet] = TrueProb[#VSet] - \
                     To6[#VSet]  + SUM(To6[#VSet])  * (MCycle==6) - \
                     To12[#VSet] + SUM(To12[#VSet]) * (MCycle==12) ), \
   !E( LL[#VSet] = -(SUM(Freq[#VSet] * LOG(Prob[#VSet])) + \
                    TailFreq[#VSet] * LOG(1-SUM(WtProb*Prob[#VSet])) ) ), \
   !E( LogLik = 2 * VSUM(!P(LL[#VSet])) )
FITNONLINEAR [PRINT=summary,estimate,monitoring,corr; \
              CALCULATION=model]
```

```
  "Display fitted values and chi-squared statistic"
CALCULATE FitFreq1[#VSet] = TotFreq[#VSet] * WtProb * TrueProb[#VSet]
&         FitFreq2[#VSet] = TotFreq[#VSet] * WtProb * Prob[#VSet]
&         FitTail1[#VSet] = TotFreq[#VSet] * \
                            (1-SUM(WtProb*TrueProb[#VSet]))
&         FitTail2[#VSet] = TotFreq[#VSet] * \
                            (1-SUM(WtProb*Prob[#VSet]))
&         ChiSq1[#VSet] = SUM( (Freq[#VSet] - FitFreq1[#VSet]) ** 2 / \
                           FitFreq1[#VSet] ) + \
                           (TailFreq[#VSet] - FitTail1[#VSet]) ** 2 / \
                           FitTail1[#VSet]
&         ChiSq2[#VSet] = SUM( (Freq[#VSet] - FitFreq2[#VSet]) ** 2 / \
                           FitFreq2[#VSet] ) + \
                           (TailFreq[#VSet] - FitTail2[#VSet]) ** 2 / \
                           FitTail2[#VSet]
FOR Observed=Freq[#VSet]; Fitted1=FitFreq1[#VSet]; \
    Fitted2=FitFreq2[#VSet]; T1=TailFreq[#VSet]; \
    T2=FitTail1[#VSet]; T3=FitTail2[#VSet]; \
    X1=ChiSq1[#VSet]; X2=ChiSq2[#VSet]
    RESTRICT MCycle,Observed,Fitted1,Fitted2; MCycle<=12
    PRINT MCycle,Observed,Fitted1,Fitted2; DECIMALS=0,0,1,1
    PRINT [IPRINT=*] '> 12',T1,T2,T3; D=0,0,1,1
    PRINT            ' ',' ',X1,X2; DECIMALS=0,0,2,2
    RESTRICT MCycle,Observed,Fitted1,Fitted2
ENDFOR
DELETE [REDEFINE=yes] MSet,VSet,TSet

RETURN
```

Extra Output for Principal Components Analysis

A Bar-Hen
Institute of Agricultural Research
B.P. 2123, Yaoundé, Cameroon.

G McLaren
ODA Biometrician IRA
c/o FCO(Yaoundé)
King Charles Street
London
United Kingdom *SW1A 2AH*

1. Introduction

The interpretation of a Principal Components Analysis often seeks to explain structure amongst variables and the individuals on which they are measured. The Genstat PCP directive provides the basic calculations of the eigenstructure of the covariance or correlation matrix but there are other derived statistics more useful for the interpretation. The PCA procedure uses the PCP directive and then computes measures of the contribution of each variable and each individual to selected principal axes. These statistics can be printed, saved or both. Graphs of correlations between variables and axes and of the projection of individuals on the basic planes of the principal component space can also be printed.

The data for the procedure consists of a set of variates, specified in a pointer given by the DATA parameter of PCA. Input from an SSPM structure is not supported because information about the individuals is required. Calculations, printing and plotting are controlled by options and storage of results by further parameters. These are explained in full in the procedure listing.

2. Description of Methods used in Procedure

2.1. Basic Computations

Given the basic data matrix X of n individuals by p variables, the basic eigenstructure of the covariance or correlation matrix, depending on the METHOD option, is calculated by the PCP directive. If the $p \times p$ covariance or correlation matrix is C, and V is the $p \times k$ matrix of eigenvectors corresponding to the k eigenvalues in the diagonal matrix D, we have the relationship: $CV = VD$. The number of components computed, k, is controlled by the NROOTS option of PCA. The values of D, the percentage of variation accounted for by each component, a test of the smallest roots, the eigenvectors or loadings, V, and the principal component scores, ZV, (where Z is the centralised or standardised data matrix derived from X depending on the METHOD option), can all be printed, saved or both, with options and parameters exactly as for the PCP directive.

2.2. Contribution of Variables

The correlation between the n observations on the ith original variable and the n scores of the jth principal component is given by the (i,j)th element of the matrix $S^{-\frac{1}{2}}VD^{\frac{1}{2}}$ where $S^{-\frac{1}{2}}$ is a diagonal matrix of reciprocals of the sample standard errors for each variable and the $D^{\frac{1}{2}}$ is the diagonal matrix of square roots of the k eigenvalues.

These correlations are used to assess the strength and direction of the linear relationships between the variables and the principal components. The squares of these coefficients measure the proportion of the variability of each variable accounted for by each principal component. The sum of these squared coefficients over the set of components for each variable is the squared multiple correlation between the variable and the k principal components.

2.3. Contribution of Individuals

The correlation between the p measurements on the ith individual and the p coefficients of the jth principal component are given by the (i,j)th element of the matrix $Q^{\frac{1}{2}}ZV$. Z is either the centralised or standardised $n \times p$ data matrix, depending on the value of option METHOD, and Q is a diagonal matrix of entries $1/\sqrt{q_i}$ where q_i is the sum of squared values in the ith row of Z. ZV is the matrix of principal component scores and is calculated by the PCP directive.

These correlations indicate how well each individual is represented on each principal axis. The sum of squared coefficients across the k axes gives the squared multiple correlation between the ith individual and the k axes and so is a measure of the quality of representation of each point by the k axes. Small multiple correlations indicate individuals that are not well represented by the k chosen axes.

2.4. Graphical Representation of Structure of Individuals

Graphs of the projections of all individuals on the plane of each pair of chosen principal axes can be printed according to the PRINT option. These plots are scatter plots of the principal component scores and are labelled with case numbers of the individuals. Coincident points are listed below the plot.

The main reason for viewing these graphs, and the scores from which they are produced, is to determine which individuals contribute most to each axis and whether any axis is almost entirely due to outlying individuals. The scores should be Normally distributed with zero mean and variance equal to the corresponding eigenvalue. Outlying individuals are easily detected by looking at those with large absolute scores, and outlying groups are easily seen on the graphs.

However, even when there appear to be no outlying groups it is possible that individuals are poorly represented on certain axes. An interpretation of the structure of the individuals which ignores the quality of the representation on the axes is likely to detect spurious relationships. The quality of representation of an individual on an axis is essentially determined by the angle between the axis and the vector to the point. If this angle is small then the point is well represented on the axis, if it is large, near a right angle, it is poorly represented. The correlations computed as the contribution of individuals are the cosines of these angles. The sum of squares of these cosines for any group of axes gives the multiple squared correlation coefficient for that individual and those axes. Therefore individuals that have small squared multiple correlations with the two axes forming a particular plot are poorly represented and should only be included in interpretations of structure with care.

2.5. Graphical Representation of the Correlations of Variables

Using the magnitude of coefficients of the eigenvectors to interpret the principal axes in terms of the variates suffers the same disadvantage as the interpretation of individuals in terms of the magnitude of their scores; that is, poorly represented variables can have large coefficients. It is necessary to consider the magnitude of coefficients only for those variables that are well represented; that is, highly correlated with the principal axes.

The correlations between the variables and the principal axes produced for the contribution of variables can be plotted for each pair of chosen axes according to the PRINT option. These plots indicate the quality of representation of each variable by the axes. The further a point lies from the origin of the plot the better represented is the corresponding variable. Since the points are correlations, all points lie within the unit circle. The interpretation of the axes depends on the closeness of the points corresponding to well represented variables to the axes. Variables plotted far from the origin but close to one axis indicate a practical interpretation of the axis. Those that lie between two axes indicate a contribution to the ordinations in both directions.

3. Example

In a study of the morphology of Cacao pods, 10 different measurements were made on pods from 33 different sources. The objective was to see if differences in morphology could be described by few variables, and whether these could be interpreted. The data plus four derived variates and the Genstat job to perform the principal components analysis are given in Appendix 1.

Because of the varying units and magnitudes of variables, the analysis is perfomed on the standardised data; that is, on the correlation matrix, with METHOD=CORRELATION. A preliminary run with NROOTS set to 4 shows that the fourth eigenvalue is only 0.803 and accounts for 6% of the variability in the sample, and furthermore no variable or individual was particularly well represented by the fourth axis. It was therefore decided to adopt three axes. Results are shown in Appendix 2.

3.1. Structure of the Variables

The first three axes account for 85% of the variability in the data. Looking at the contribution of variables and the circle of correlations of variables with the first two axes, we see that the masses (MS, MC and MB) and the dimensions of the pods (LG, WD, and LW) are very well represented on the first axis while the thickness of cortex (TX and TN) and the ratio of bean mass to pod mass (PC) are well represented on the second axis. The first axis can therefore be interpreted as a measure of overall size of pod. These measures of size are all correlated in the same sense with this axis. The second axis appears to measure the internal structure, thickness of cortex and amount of beans; these are correlated in opposite senses with the axes. There are no well represented variables on the third axis, but the best represented are the ratio of cortex thickness (XN) and the number of wilted beans (NW). The count of normal beans (NN) is best represented on the first axis with the measures of pod size. The bean size (MN) is not well represented on any of the first three axes which account for only 50% of the variability in this measurement.

3.2. Structure of the Individuals

Considering the above interpretation of the variables we can look at the principal component scores, the contributions of individuals and the projections on the principal axes planes to check that the structure of the individuals accords with the variables. Given that the scores are central variates with variance estimated by the corresponding eigenvalues, we can see that individuals 8 and 24 have abnormally large scores on the first axis. One can drop these individuals from the analysis to see if the axes are very sensitive to them. In fact the axes change very little; the large values are just extremes of the general axis trend. Similarly the large scores of variables 13, 21 and 24 on axis two represent pods with a small proportion of beans (PC) and thicker than average cortex (TX and TN) again following the general trend of the axis and not overly influential. There is a group of individuals, 3, 12, 28, 30 and 31 at the centre of each plot which are poorly represented.

4. The Procedure PCA

Editors' Note. The values of the options PRINT and METHOD need to be in UPPER CASE. This procedure has been modified to work with Release 2. For use with Release 1.3 see the comment in the section on 'compute contribution of individuals'.

```
PROCEDURE 'PCA':"EXTRA OUTPUT FOR PRINCIPAL COMPONENTS ANALYSIS"

OPTION NAME='PRINT','SMALLEST','METHOD','NROOTS','NROWS',\
                'NCOLUMNS','WEIGHTS'; MODE= 3(t),3(v),p;\
                DEFAULT=' ','NO','S',2,21,61,*
PARAMETER NAME='DATA','LRV','SSPM','SCORES','RESIDUALS',\
                'VARCORR','INDIVCORR'; MODE=7(p)

" Description
This procedure provides a principal component analysis, with
extra options for the contribution of variables, contribution of
individuals and plots of correlations between variables and axes
and projections of individuals on the chosen axes. The data for
the procedure consist of a set of variates, specified in a
pointer given by the DATA parameter. Correlations between
individuals or variables and the axes can be printed using
the PRINT option or saved with the VARCORR and INDIVCORR
parameters. The projections of individuals onto the principal axes
are labelled by case numbers on the graphical output and those of
the variables by the first two letters of the variable names. A
list of coincident points follows each graph.

Action with RESTRICT
The input data may be restricted. The analysis is based only on the
units retained by the restriction.

OPTIONS
PRINT   = text      What to output:LOADINGS,ROOTS,
                    RESIDUALS,SCORES,TESTS as for PCP directive,
                    VARCORR for the correlation between variables and
```

```
                        the principal axes,
                        INDIVCORR for the correlation between individuals
                        and the principal axes,
                        GRAPH to plot the projection of and correlations
                        variables on the principal axes;
                        default * i.e. no output.
NROOTS   = scalar       Number of latent roots to compute,
                        as for PCP directive; default 2
SMALLEST= string        Whether to print smallest instead of the largest,
                        as for PCP directive; default no.
METHOD   = string       Whether to use SSCP matrix of correlations,
                        as for PCP directive; default SSPM
WEIGHTS = vector        weightings for the units; default *
                        i.e. all units have weight one
NROWS    = scalar       number of rows in the graph frame; default 21
NCOLUMNS= scalar        number of rows in the graph frame; default 61

PARAMETERS
DATA     = pointer of variates forming the data matrix, unlike PCP
           an SSPM structure is not permitted for input.
LRV      = lrv to store the eigen structure as for PCP directive
SSPM     = sspm to store SSCP or correlation matrix as for PCP
SCORES   = matrix to store scores as for PCP directive
RESIDUAL= matrix to store residuals as for PCP directive
VARCORR = matrix to store the contribution of variables.
INDIVCOR= matrix to store the contribution of individuals."

                                "set the environment"
GET [ENVIRONMENT=ENV]: SET [DIAGNOSTIC=F;CASE=I]
                                 "set parameters"
GETATTRIBUTE [ATTRIBUTE=type] DATA;NB
IF NB['type'].NE.14
   PRINT '**********DATA MUST BE A POINTER*************';\
   JUSTIFICATION=LEFT
   EXIT [CONTROL=PROCEDURE]
ENDIF
GETATTRIBUTE [ATTRIBUTE=NVALUES] DATA[1],DATA;NB,NB1
IF NROOTS.NI.!(2...#NB1[])
   PRINT '**********ILLEGAL NUMBER OF ROOTS*************';\
   JUSTIFICATION=LEFT
   EXIT [CONTROL=PROCEDURE]
ENDIF
IF UNSET(LRV)
   ASSIGN DLrv ; POINTER=LRV
   LRV [ROWS=DATA;COLUMNS=#NROOTS] LRV
ENDIF
IF UNSET(WEIGHTS)
   ASSIGN DWeights; POINTER=WEIGHTS: CALC WEIGHTS=!(#NB[](1))
ENDIF
IF UNSET(SSPM): ASSIGN DSspm; POINTER=SSPM: ENDIF
SSPM [TERMS=DATA[]] SSPM: FSSPM [WEIGHTS=WEIGHTS;PRINT=*] SSPM
IF UNSET(SCORES)
   ASSIGN DScores ; POINTER=SCORES
   MATRIX [ROWS=#NB[];COLUMNS=#NROOTS] SCORES
ENDIF
IF UNSET(RESIDUALS)
   ASSIGN DResiduals; POINTER=RESIDUALS
   MATRIX [ROWS=#NB[];COLUMNS=1] RESIDUALS
ENDIF
IF SMALLEST.NI.!T(NO,N,YES,YE,Y,no)
PRINT '*********ILLEGAL VALUE FOR SMALLEST OPTION***********';\
    JUSTIFICATION=LEFT
    EXIT [CONTROL=PROCEDURE]
ENDIF
                               "separate values of print options"
SCALAR [VALUE=0] OPT[1...3]
FOR J= !T(VARCORR,VARCOR,VARCO,VARC,VAR,VA,V),\
       !T(INDIVIDU,INDIVID,INDIVI,INDIV,INDI,IND,IN,I),\
       !T(GRAPH,GRAP,GRA,GR,G);K=1...3
    IF MAX(PRINT.IN.J): CALC OPT[K]=1: ENDIF
    RESTRICT PRINT;PRINT.NI.J
```

```
    EXIT .NOT.NVALUES(PRINT)
ENDFOR
IF NVALUES(PRINT)
PRINT [CH=IMP;IPRINT=*] PRINT;JUSTIFICATION=LEFT;SKIP=0
RESTRICT PRINT;PRINT.NI.!T(ROOTS,ROOT,ROO,RO,R,\
         LOADINGS,LOADING,LOADIN,LOADI,LOAD,LOA,LO,L,\
         SCORES,SCORE,SCOR,SCO,SC,S,\
         RESIDUALS,RESIDUAL,RESIDUA,RESIDU,RESID,RESI,RES,RE,\
         TESTS,TEST,TES,TE,T,' ')
ENDIF
IF NVALUES(PRINT)
    PRINT '*********ILLEGAL VALUE IN PRINT OPTION*********';\
    JUSTIFICATION=LEFT
    EXIT [CONTROL=PROCEDURE]
ENDIF


                          "compute the pcp"
PCP[PRINT=#IMP;METHOD=#METHOD;SMALLEST=#SMALLEST;NROOTS=#NROOTS]\
     SSPM;LRV=LRV;SCORES=SCORES;RESIDUALS=RESIDUALS


                "compute the contribution of variables"
IF .NOT.UNSET(VARCORR).OR.OPT[1].OR.OPT[3]
   IF UNSET(VARCORR): ASSIGN DVARCORR; POINTER=VARCORR: ENDIF
   IF METHOD.NI.!T('SSPM','SSP','SS','S')
      CALC A[1...#NB1[]]=#NB1[](1)
   ELSE
       CALC A[1...#NB1[]]=1/SQRT(#SSPM['Sums']$[1...#NB1[]])
   ENDIF
   CALC D=SQRT(LRV['Roots'])
   DIAGONALMATRIX [ROWS=#NROOTS] DIAG ; VALUES=D
   & [ROWS=#NB1[]] STAND; VALUES=!(A[])
   CALC VARCORR=PRODUCT(LRV['Vectors'];DIAG)
   & VARCORR=PRODUCT(STAND;VARCORR)
   IF OPT[1]
     PRINT [CH=TT;SQUASH=Y;IPRINT=*;SERIAL=Y] 1...#NROOTS,'SUM';\
     JUSTIFICATION=LEFT
     MATRIX [ROWS=DATA;COLUMNS=!T(#TT)] PRTCORR
     CALC PRTCORR$[*;1...#NROOTS]=VARCORR$[*;1...#NROOTS]
     & PRTCORR$[*;' SUM']=PRODUCT(VARCORR**2;!(#NROOTS(1)))*100
     PRINT [SERIAL=Y;SQUASH=Y] 'CONTRIBUTION OF VARIABLES',\
'(Correlation between variables and principal axes and multiple',\
'squared correlation coefficient for each variable on chosen axes)'
     PRINT [IPRINT=*] PRTCORR;DECIMALS=4
   ENDIF
ENDIF


                      "compute contribution of individuals"
IF .NOT.UNSET(INDIVCORR).OR.OPT[2]
 IF UNSET(INDIVCORR): ASSIGN DINDIVCORR;POINTER=INDIVCORR: ENDIF
 IF METHOD.NI.!T('SSPM','SSP','SS','S')
   " For Release 1 use the following line "
 " CALC Z[1...#NB1[]]=((DATA[]-MEAN(DATA[]))/SQRT(VAR(DATA[])))**2"
   " For Release 2 use the following lines"
   CALC Z[1...#NB1[]]=(DATA[]-MEAN(DATA[]))**2
   CALC Z[1...#NB1[]]=Z[]/SUM(Z[])
 ELSE
   CALC Z[1...#NB1[]]=(DATA[]-MEAN(DATA[]))**2
 ENDIF
   CALC P=1/SQRT(VSUMS(Z))
   DIAGONALMATRIXRIX [ROWS=#NB[]] YDIAG ;VALUES=P
   CALC INDIVCORR=PRODUCT(YDIAG;SCORES)
   IF OPT[2]
     PRINT [CH=TT;SQUASH=Y;IPRINT=*;SERIAL=Y] 1...#NROOTS,'SUM';\
     JUSTIFICATION=LEFT
     MATRIX [ROWS=#NB[];COLUMNS=!T(#TT)] PRTINDIV
     CALC PRTINDIV$[*;1...#NROOTS]=INDIVCORR$[*;1...#NROOTS]
     & PRTINDIV$[*;!T(' SUM')]=PRODUCT(INDIVCORR**2*100;\
     !(#NROOTS(1)))
```

```
      PRINT [SERIAL=Y;SQUASH=Y] 'CONTRIBUTION OF INDIVIDUALS',\
  '(Correlation between individuals and principal axes and multiple',\
  'squared correlation coefficient for each individual on chosen axes)'
      PRINT [IPRINT=*] PRTINDIV;DECIMALS=4
    ENDIF
  ENDIF



                          "do graphs"
  IF OPT[3]
   PRINT [CH=TE;SQUASH=Y;IPRINT=*] DATA;JUSTIFICATION=LEFT;SKIP=0
   FACTOR [LABELS=TE;VALUES=1...#NB1[]] VLABELS
   FACTOR [LEVELS=#NB[];VALUES=1...#NB[]] ILABELS
   CALC comp[1...#NROOTS]=SCORES$[*;1...#NROOTS]
   & dim[1...#NROOTS]=VARCORR$[*;1...#NROOTS]
   CALC TEMP=NROOTS-1
   FOR I=1...TEMP
      CALC TEMP1=I+1
      FOR J=TEMP1...NROOTS
          PRINT[CH=TIT] 'PROJECTION OF INDIVIDUALS ON PRINCIPAL AXES '\
           ,I,'AND',J;DECIMALS=0;FIELDWIDTH=46,2,3,2
          GRAPH [TITLE=TIT;EQUAL=SCALE;NROWS=#NROWS;NCOLUMNS=#NCOLUMNS]\
           comp[J];comp[I]; SYMBOLS=ILABELS
          PRINT[CH=TIT] 'PROJECTION OF VARIABLES ON PRINCIPAL AXES '\
           ,I,'AND',J;DECIMALS=0;FIELDWIDTH=46,2,3,2
          GRAPH [TITLE=TIT;NROWS=#NROWS;NCOLUMNS=#NCOLUMNS;XLOWER=-1;\
          XUPPER=1;YLOWER=-1;YUPPER=1] dim[J];dim[I];SYMBOLS=VLABELS
      ENDFOR
   ENDFOR
  ENDIF
                          "reset the environment"
  SET [DIAGNOSTIC=#ENV['diagnostic'];CASE=#ENV['case']]
  ENDPROCEDURE
```

5. Appendix 1 – Example

```
      UNITS [33]
      VARIATE MS;EXTRA='MASS OF POD (GMS)'
            & LG;EXTRA='LENGTH OF POD (CMS)'
            & WD;EXTRA='WIDTH OF POD (CMS)'
            & TX;EXTRA='MAXIMUM THICKNESS OF CORTEX (CMS)'
            & TN;EXTRA='MINIMUM THICKNESS OF CORTEX (CMS)'
            & MC;EXTRA='MASS OF CORTEX (GMS)'
            & MB;EXTRA='MASS OF FRESH BEANS (GMS)'
            & NN;EXTRA='NUMBER OF NORMAL BEANS'
            & NW;EXTRA='NUMBER OF WILTED BEANS'
      READ MS,LG,WD,TX,TN,MC,MB,NN,NW
        348 15.19  7.67 1.08 0.81 242  99  38  2
        368 14.43  7.81 1.27 0.99 261  98  35  0
        327 15.15  7.38 1.23 0.92 216  95  36  2
        420 17.09  8.12 1.19 0.97 299 111  33  6
        242 13.65  6.75 0.97 0.70 166  68  26  2
        280 12.85  7.52 1.22 0.82 209  62  21  1
        334 14.55  7.71 1.27 1.02 242  82  26  4
        645 19.20  9.70 1.30 1.00 430 195  49  0
        387 15.81  7.94 1.32 1.10 281 100  35  3
        322 14.78  7.54 1.07 0.83 220  94  36  2
        392 14.88  8.01 1.26 0.93 280  99  32  3
        331 14.41  7.37 1.28 1.03 239  90  34  2
        400 15.93  7.89 1.41 1.19 308  87  31  1
        358 14.45  7.91 1.30 1.13 256  91  36  4
        292 14.04  6.89 1.08 0.84 209  77  29  5
        328 14.44  7.59 1.27 1.05 230  90  34  5
        266 13.39  7.09 1.14 0.90 188  71  25  6
        227 13.56  6.75 0.99 0.73 154  65  28  5
        358 15.93  7.50 1.18 0.98 263  85  31  3
        257 13.32  7.14 1.16 0.94 185  67  25  4
        409 16.39  7.79 1.31 1.06 321  81  30  7
        272 13.19  7.22 1.12 0.84 199  68  26  3
        364 15.59  7.69 1.26 1.05 258  95  33  2
```

```
  185 11.40  6.00 1.25 0.95 148  33  12  2
  352 14.65  7.78 1.34 1.03 240 104  33  0
  290 13.48  7.40 1.11 0.89 216  75  24  5
  209 12.57  6.53 1.02 0.81 149  56  26  8
  330 14.89  7.62 1.16 0.88 230  90  28  8
  440 15.20  7.70 1.30 1.00 325 110  45  1
  325 14.57  7.70 1.16 0.96 229  90  32  3
  334 14.90  7.45 1.17 0.96 232  92  32  3
  257 13.40  7.03 1.06 0.80 180  72  29  3
  424 17.35  7.98 1.18 0.90 288 126  37  9
:
VARIATE LW;EXTRA='VOLUME OF POD'
      & XN;EXTRA='RATIO OF MAX TO MIN THICKENSS'
      & PC;EXTRA='RATIO OF MASS OF BEANS TO MASS OF POD'
      & MN;EXTRA='MASS PER NORMAL BEAN'
CALCULATE LW=LG*WD*WD: & XN=TX/TN: & PC=MB/MS: & MN=MB/NN
POINTER [VALUES=MS,LG,WD,LW,TX,TN,XN,MC,MB,NW,NN,PC,MN] VTS
PCA [PRINT=R,L,S,G,V,I;METHOD=C;NROOTS=3] VTS
STOP
```

6. Appendix 2 – Example Results

```
***** Principal components analysis *****

*** Latent Roots ***

 DLRV['Roots']
        1          2          3
    7.565      2.255      1.223

*** Percentage variation ***

 DLRV['Roots']
        1          2          3
   58.19      17.35       9.41

*** Trace ***

 DLRV['Trace']
    13.00

*** Principal Component Scores ***

                1          2          3
  1      -0.0290    -0.3593     0.0851
  2      -0.1906     0.0647     0.2124
  3      -0.0112    -0.1635     0.1120
  4      -0.4992    -0.0633    -0.2159
  5       0.6144    -0.4275     0.1691
  6       0.4004    -0.0234     0.5798
  7      -0.0585     0.2211    -0.0099
  8      -1.7123    -0.4214     0.2264
  9      -0.3931     0.2522    -0.0808
 10       0.0636    -0.3102     0.0077
 11      -0.2321    -0.0411     0.2036
 12      -0.0429     0.1467     0.0152
 13      -0.4193     0.5999     0.0790
 14      -0.2177     0.3568    -0.2166
 15       0.3527    -0.1284    -0.1053
 16      -0.0635     0.1523    -0.2051
 17       0.3760    -0.0143    -0.1534
 18       0.6603    -0.3882    -0.0642
 19      -0.1006     0.1741    -0.0875
 20       0.3725     0.1085    -0.0942
 21      -0.2713     0.4202    -0.1682
 22       0.4162    -0.0567     0.1260
 23      -0.2418     0.1879    -0.0414
 24       0.9193     0.5538     0.3078
 25      -0.2570     0.0134     0.2343
 26       0.2467    -0.0118    -0.0994
 27       0.7821    -0.1149    -0.3431
```

28	0.0288	-0.1857	-0.1757
29	-0.3918	0.0804	0.1724
30	-0.0330	-0.0010	-0.1121
31	-0.0331	-0.0008	-0.0963
32	0.4656	-0.2416	0.0375
33	-0.5007	-0.3791	-0.2994

*** Latent Vectors (Loadings) ***

DLRV['Vectors']

VTS	1	2	3
MS	-0.35854	-0.02352	0.03803
LG	-0.33590	-0.08956	-0.16082
WD	-0.34927	-0.05733	0.04774
LW	-0.35338	-0.08937	0.00427
TX	-0.22601	0.45659	0.20034
TN	-0.21503	0.51089	-0.10015
XN	0.11096	-0.37296	0.58241
MC	-0.34924	0.07230	0.04401
MB	-0.34207	-0.20713	0.00389
NW	0.08034	-0.07665	-0.72138
NN	-0.29534	-0.18933	-0.07383
PC	-0.08852	-0.52285	-0.20118
MN	-0.24824	-0.09133	0.11295

CONTRIBUTION OF VARIABLES
(Correlation between variables and principal axes and multiple squared correlation coefficient for each variable on chosen axes)

VTS	1.000	2.000	3.000	SUM
MS	-0.9861	-0.0353	0.0421	97.5483
LG	-0.9239	-0.1345	-0.1779	90.3270
WD	-0.9607	-0.0861	0.0528	93.3058
LW	-0.9719	-0.1342	0.0047	96.2701
TX	-0.6216	0.6856	0.2216	90.5625
TN	-0.5914	0.7672	-0.1108	95.0619
XN	0.3052	-0.5600	0.6441	82.1658
MC	-0.9606	0.1086	0.0487	93.6819
MB	-0.9408	-0.3110	0.0043	98.1941
NW	0.2210	-0.1151	-0.7978	69.8541
NN	-0.8123	-0.2843	-0.0816	74.7373
PC	-0.2435	-0.7851	-0.2225	72.5210
MN	-0.6828	-0.1371	0.1249	50.0598

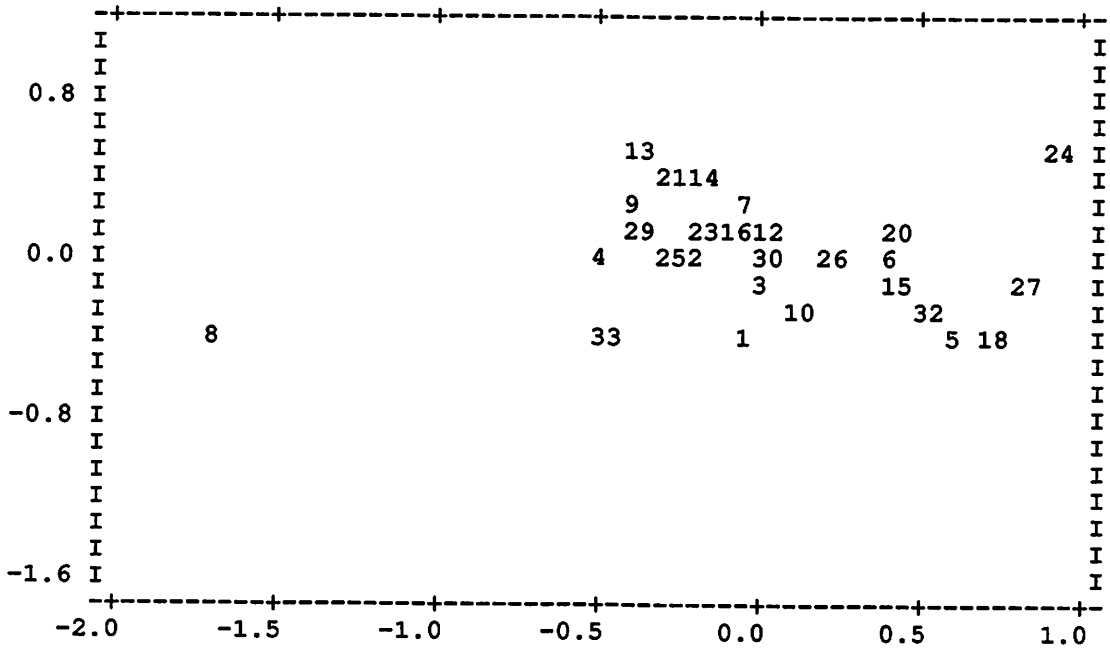
CONTRIBUTION OF INDIVIDUALS

(Correlation between individuals and principal axes and multiple squared correlation coefficient for each individual on chosen axes)

	1.000	2.000	3.000	SUM
1	-0.0672	-0.8318	0.1970	73.5237
2	-0.5474	0.1858	0.6100	70.6200
3	-0.0339	-0.4940	0.3382	35.9594
4	-0.8165	-0.1035	-0.3532	80.2152
5	0.7872	-0.5477	0.2166	96.6643
6	0.5106	-0.0298	0.7394	80.8271
7	-0.1830	0.6912	-0.0310	51.2232
8	-0.9567	-0.2355	0.1265	98.6699
9	-0.8103	0.5200	-0.1665	95.4727
10	0.1623	-0.7907	0.0196	65.1841

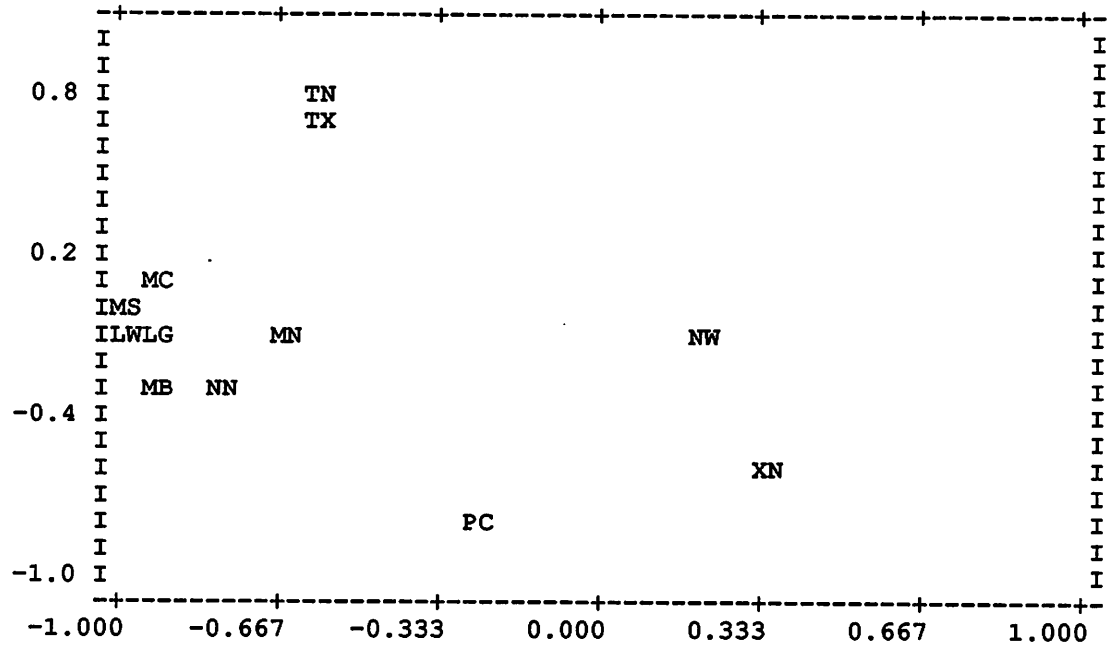
11	-0.6271	-0.1110	0.5500	70.8184
12	-0.1495	0.5115	0.0530	28.6806
13	-0.5622	0.8043	0.1060	97.4188
14	-0.4058	0.6650	-0.4037	76.9861
15	0.8703	-0.3167	-0.2598	92.5169
16	-0.1907	0.4575	-0.6160	62.5125
17	0.8611	-0.0327	-0.3513	86.5913
18	0.8459	-0.4973	-0.0822	96.9523
19	-0.3201	0.5543	-0.2785	48.7349
20	0.8857	0.2579	-0.2239	90.1061
21	-0.3947	0.6114	-0.2447	58.9547
22	0.9291	-0.1265	0.2813	95.8262
23	-0.6940	0.5393	-0.1189	78.6719
24	0.7954	0.4792	0.2663	93.3290
25	-0.5231	0.0274	0.4771	50.2027
26	0.6290	-0.0302	-0.2536	46.0876
27	0.8938	-0.1313	-0.3921	96.9887
28	0.0658	-0.4238	-0.4010	34.4807
29	-0.6235	0.1279	0.2744	48.0411
30	-0.1462	-0.0042	-0.4968	26.8161
31	-0.1716	-0.0041	-0.4990	27.8449
32	0.8612	-0.4469	0.0694	94.6243
33	-0.6371	-0.4823	-0.3809	78.3656

PROJECTION OF INDIVIDUALS ON PRINCIPAL AXES 1 AND 2



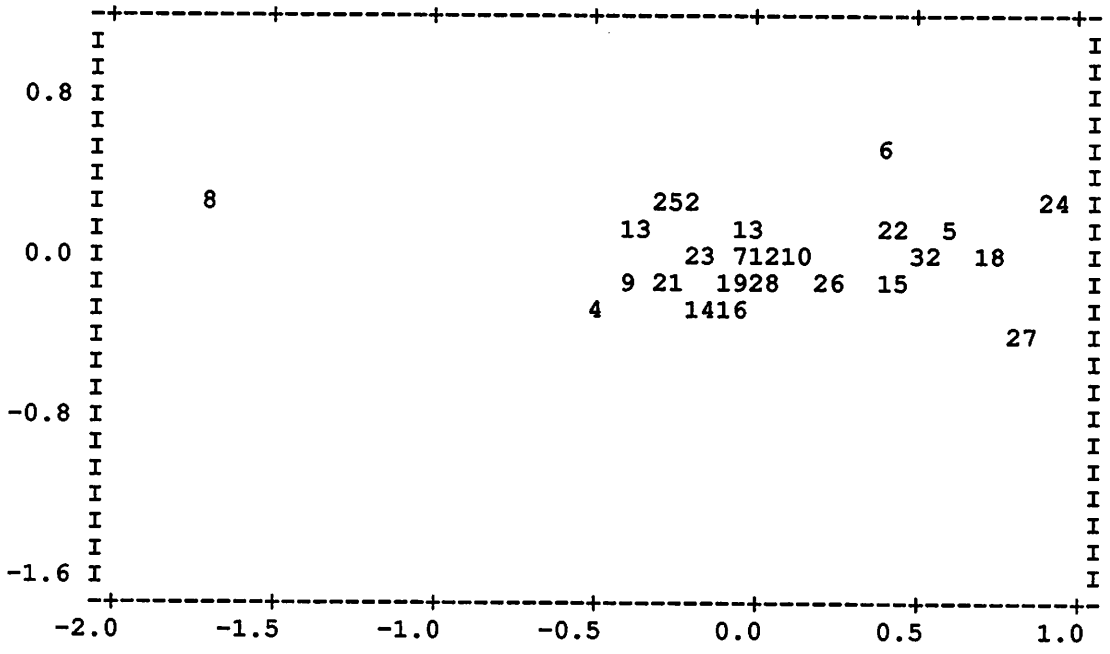
Points coinciding with 2 6 16 3 30
11 17 19 28 31
22

PROJECTION OF VARIABLES ON PRINCIPAL AXES 1 AND 2



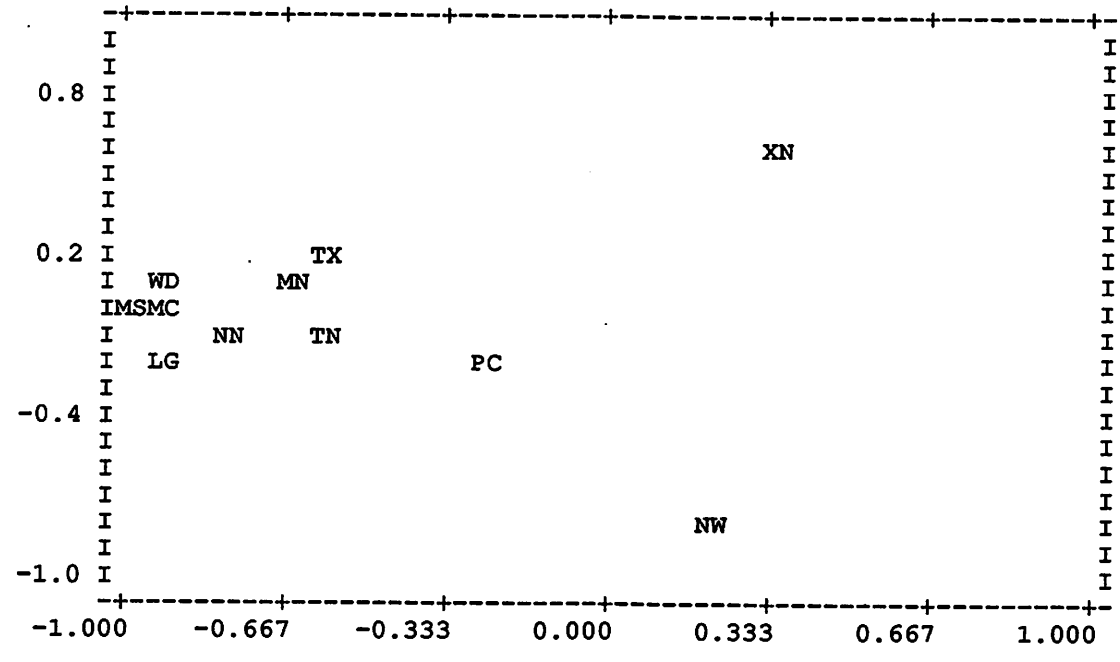
Points coinciding with LG
WD

PROJECTION OF INDIVIDUALS ON PRINCIPAL AXES 1 AND 3



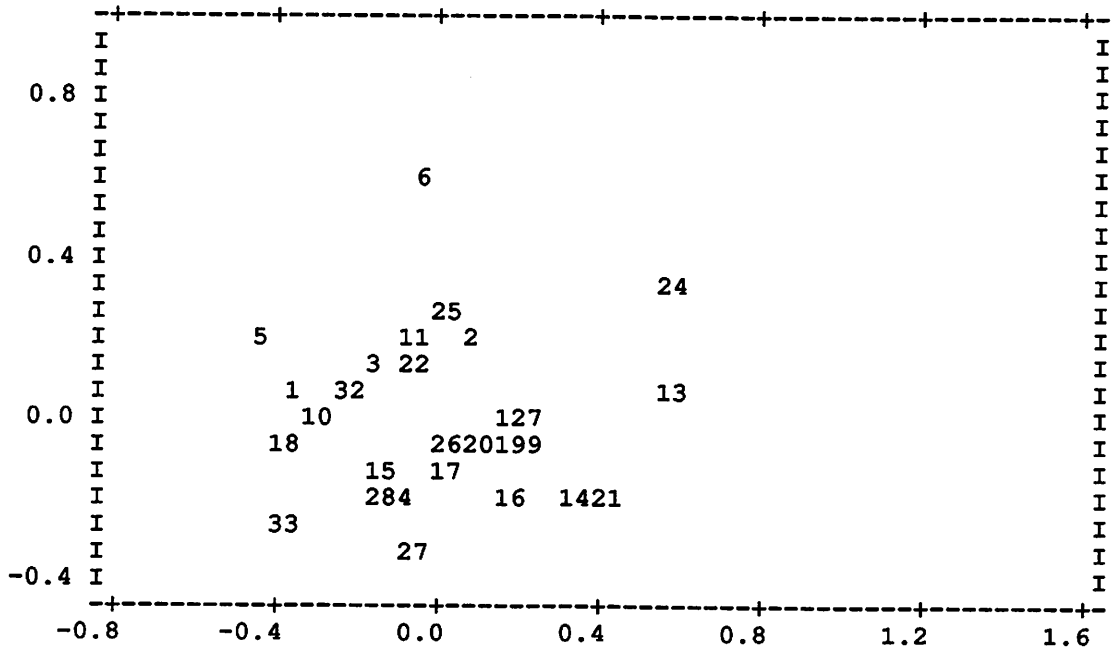
Points coinciding with 2 15 13 28 4
11 17 29 30 33
20 31

PROJECTION OF VARIABLES ON PRINCIPAL AXES 1 AND 3



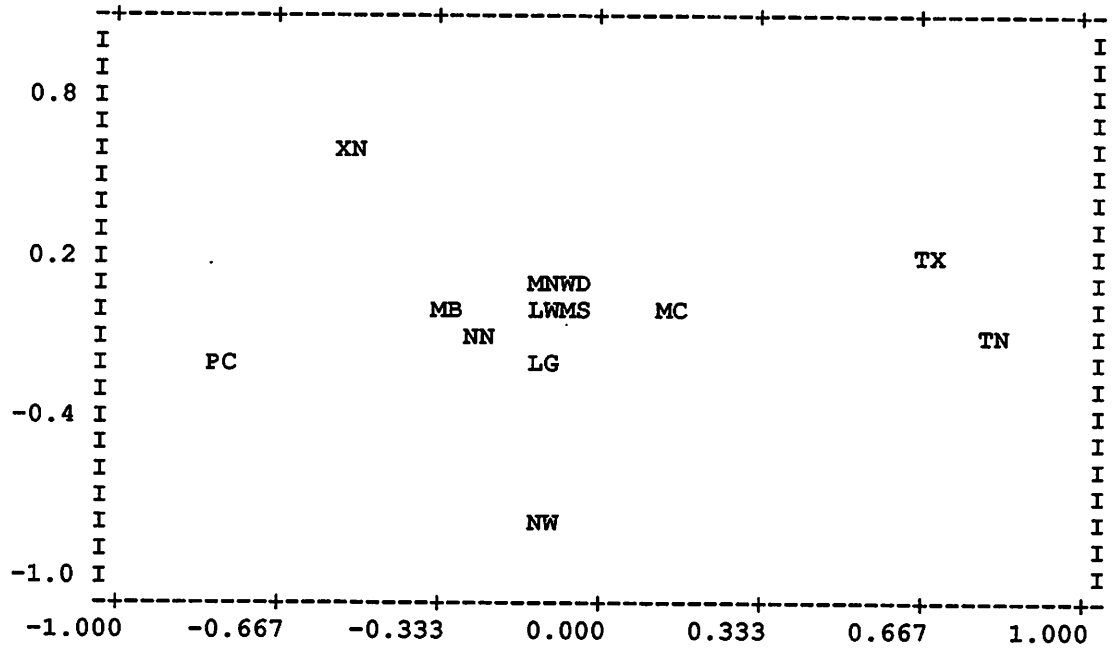
Points coinciding with MS MC
LW MB

PROJECTION OF INDIVIDUALS ON PRINCIPAL AXES 2 AND 3



Points coinciding with 5 19 2 17 26
8 23 29 30 31

PROJECTION OF VARIABLES ON PRINCIPAL AXES 2 AND 3



A Genstat 5 Procedure for Generating Discrete Distributions Belonging to an Exponential Family

L P Lefkovitch
Statistical Research
Research Program Service
Agriculture Canada
Bldg. 54
Central Experimental Farm
Ottawa
Ontario
Canada K1A OC6

1. Introduction

Suppose that for one of a pair of dice there is evidence that the true mean score is exactly 4.5 (or some other value), but that no reason is known for the departure from the value of 3.5, consistent with a fair die. In the biased die, a score of 4.5 may have arisen either from faces 1, 2, 3 and 6 each having a probability of zero, with each of faces 4 and 5 having a probability of 0.5, or from several other possibilities in which only two faces have equal or non-equal non-zero probabilities, or from very many others in which the probability of each of the six faces may be non-zero. If the only information about the die is the mean score, what probabilities should be assigned to each face?

Consider some similar questions arising from the following series of examples of contingency tables.

Example 1: Suppose there is a 2 by 2 contingency table for which only the marginal proportions are known, e.g.

a	b	$\frac{3}{4}$
c	d	$\frac{1}{4}$
$\frac{3}{4}$	$\frac{1}{4}$	1

which implies

$\frac{1}{2}+q$	$\frac{1}{4}-q$
$\frac{1}{4}-q$	q

What value should q take? Let $n = 4$; there are just two possible tables satisfying the marginal proportions:

2 1	3 .
1 .	. 1

In considering these two tables, which is to be preferred given ignorance of the actual values? One way to answer this is to consider how many different ways each table can be generated.

Letting a, b, c and d now denote the frequencies in the table, the number of ways each may be generated is given by

$$w = n!/(a!b!c!d!)$$

which are 12 and 4 respectively, i.e. the first table can be realised in the greater number of ways.

Suppose $n = 16$; the possible tables satisfying the same marginal proportions are:

8 4	9 3	10 2	11 1	12 .
4 .	3 1	2 2	1 3	. 4

for which the values of w are

900,900 1,601,600 720,720 87,360 1,820

i.e. the second table can be realised in the greatest number of ways.

Example 2: Consider the following more elaborate example;

a	*b*	$\frac{1}{2}$
c	*d*	$\frac{3}{8}$
e	*f*	$\frac{1}{8}$
$\frac{3}{4}$	$\frac{1}{4}$	*n*

Assuming $n = 16$, there are 12 distinct tables, which together with the corresponding values of w are:

4 4	5 3	6 2	7 1
6 .	5 1	4 2	3 3
2 .	2 .	2 .	2 .
25,225,200	121,080,960	151,351,200	57,657,600
8 .	5 3	6 2	7 1
2 4	6 .	5 1	4 2
2 .	1 1	1 1	1 1
5,405,400	40,360,320	121,080,960	86,486,400
8 .	6 2	7 1	8 .
3 3	6 .	5 1	4 2
1 1	. 2	. 2	. 2
14,414,400	10,090,080	34,594,560	5,405,400

It is the third table in the first row which can be generated in the greatest number of ways.

2. A Generalization

Both the example of the biased die and the contingency table examples lead to the following generalization. Consider any problem of the following general form: there is a hypothesis space H_0, described by enumerating some perceived possibilities (e.g. various values for the probabilities of the faces of the die; different probabilities for the body of a contingency table) which are not regarded as equally likely because there is some additional evidence, E (e.g. the value of the true mean in the die; the marginal proportions in the contingency tables). Since E is not an event, it does not have a sampling distribution, and in consequence, cannot be used as the data, B, in Bayes' theorem

$$p(A|B) = p(A)p(B|A)/p(B);$$

nevertheless, the existence of E leads to some constraints on the probabilities assigned to the elements of H_0 which may force them to be non-uniform, but does not fully determine them (assuming that the number of such constraints is less than the number of elements in H_0).

Thus given some 'macroscopic' information (the mean score for the die; the marginal proportions and total in a contingency table), determine a reasonable set of 'microscopic' events (the probability of each face of a die; the elements of the contingency table) which imply them. One man's 'reasonable' is another's prejudice; but, intuitively, it is claimed that choosing that set which can be achieved in the greatest number of ways can be asserted as avoiding the introduction of more structure than necessary (structure here can be interpreted as association among the margins, such as in diagonal 2 by 2 tables).

With this assertion, the remaining steps are almost automatic. Define $p_i = i/n$ where $i \in [a,b,c,d]$, and let $n = a+b+c+d$; from the Stirling approximation to the factorials for large x, $x! = (2\pi x)^{\frac{1}{2}}x^x e^{-x}$, with $x = np_i$, it can be shown that

$$\log w \approx -n\Sigma p_i \log p_i$$

the right-hand side of which is (Shannon) entropy. Thus the table which has the maximum number of possible realizations is given by the set of p_i which maximizes entropy, $-\Sigma p_i \log p_i$, subject to constraints on the p_i which satisfy the known marginal values; the p_i are used to

obtain the nearest integer values to np_i. For the contingency table of Example 1, the p_i can be obtained by maximizing entropy subject to

$$p(a) + p(b) = 0.75$$
$$p(a) + p(c) = 0.75$$
$$p(a) + p(b) + p(c) + p(d) = 1;$$

with solution, not surprisingly, $p(a) = 9/16$, $p(b) = p(c) = 3/16$, $p(d) = q = 1/16$. The entropy, in natural logarithms, is 1.125.

For Example 2, the constraint set on the probabilities is:

$$p(a) + p(b) = 0.5$$
$$p(c) + p(d) = 0.375$$
$$p(e) + p(f) = 0.125$$
$$p(a) + p(c) + p(e) = 0.75$$
$$p(a) + p(b) + p(c) + p(d) + p(e) + p(f) = 1$$

and the solution probabilities and expected values, both of which can be obtained in other ways, are

Probabilities		**Predicted values**	
0.375	0.125	6	2
0.28125	0.09375	4.5	1.5
0.09375	0.03125	1.5	0.5

from which it can be seen that the expected values differ by only small amounts from the third table above.

A more usual situation arises in contingency table analysis when the elements of the table have been observed, and it is desired to find the probabilities associated with each. Modifying the second example, we have:

$$ap(a) + bp(b) = x_1$$
$$cp(c) + dp(d) = x_2$$
$$ep(e) + fp(f) = x_3$$
$$ap(a) + cp(c) + ep(e) = x_4$$
$$p(a) + p(b) + p(c) + p(d) + p(e) + p(f) = 1$$

where $[a,b,c,d,e,f]$ is known, and the x_i are specified expected values perhaps arising from some unsaturated model, e.g. omitting the 4th equation. The maximum entropy solution to such a set of equations provides a solution of interest.

3. A Further Generalization

According to Guiaşu [2], of all distributions agreeing with a set of constraints, the maximum entropy principle expresses the enumeration of the possibilities assuming nothing beyond the evidence. Counting arguments show that the vast majority of distributions satisfying the constraints have entropy close to the maximum, and the entropy concentration theorem (Jaynes, [3]) allows an accurate estimate of how sharply these are concentrated; in fact, even for small numbers of observations, those having entropy near the maximum predominate. Because of this, the maximum entropy estimate can be considered as representative of this class.

It is widely known that the distributions which maximize entropy, subject to constraints which are linear functions of the probabilities, are all members of an exponential family (Kagan, *et al.*, [6]), often coinciding with a named distribution depending on the measure. Further, most of the probability distributions commonly encountered in statistics maximize the entropy (van Campenhout and Cover, [11]), and since in addition there is a solid axiomatic base for the principle (Shore and Johnson, [9]), its direct use has some merit. It is also well known (e.g. Feller, [1]) that the complete set of moments uniquely define a probability distribution under a mild set of conditions.

None of this is remarkable; what is provocative, however, are some characterizations, especially some special cases. An important one is that: the only (discrete) distribution which maximizes entropy subject to the single constraint specifying the mean of unbounded non-negative values is the geometric (see Guiaşu, [2]). Using the expected value of the distribution, the geometric probabilities for each cell can be obtained by maximization of the entropy, that is,

$$-\Sigma p_i \log p_i \qquad (1)$$

subject to the first-order constraint

$$\Sigma i p_i = \mu, \quad i = 0,1,2,...$$

as well as the zero-order constraint

$$\Sigma p_i = 1.$$

Other distributions are obtained if a second-order condition is also included to constrain the probabilities, e.g.

$$\Sigma i^2 p_i = \nu^2.$$

By an appropriate choice of ν^2, there is a wide class of distributions in which the variance need not be that of the standard (positive) geometric distribution, which is $\mu(\mu-1)$, but takes some other value; it follows that the probabilities will not be those of a geometric distribution. There is no restriction to first- and second-order moments, so that third and higher order constraints can also be imposed if there is cause; there can also be more than one constraint of each order. There can also be a lower and upper limit on the value of i; for example, Lefkovitch [7] gives the maximum entropy probabilities of the faces for a number of biased dice based on up to third-order constraints.

Example 3: Returning to the example in the first paragraph, in a die with expected value, $\mu = 4.5$, the probabilities p_i which maximize entropy

$$-\Sigma p_i \log p_i$$

subject to

$$\Sigma i p_i = \mu = 4.5 \quad \text{and} \quad \Sigma p_i = 1, \quad i = 1...6$$

are approximately

[0.0054 0.0788 0.1142 0.1655 0.2398 0.3475]

(Lefkovitch, [7]).

Suppose entropy is maximized subject to

$$\Sigma i p_i = \mu, \quad i = 0,1,2,...$$
$$\Sigma p_i = 1,$$
$$\Sigma i^2 p_i = \nu^2 = \mu^2 + \mu,$$

then it follows that $\sigma^2 = \mu$; but the third moment is not equal to μ, nor is the fourth moment equal to $\mu + 3\mu^2$, i.e. this maximum entropy distribution, while a member of an exponential family, and having its variance equal to the mean, is not the Poisson. The explanation is that this distribution is the closest to the uniform distribution while satisfying the constraints. Another way of expressing this is that the non-negative (counting) measure, π_i, implicitly assigned to each of the cells is equal to a constant. Including this measure into the entropy function introduces a more general expression based on the objective function,

$$\Sigma p_i \log (p_i / c\pi_i). \qquad (2)$$

If $\Sigma p_i = 1$, then in extremal problems c is irrelevant. It is convenient, **but not necessary**, to choose c so that $\Sigma c\pi_i = 1$. If $\pi_i = 1$, $\forall i$, it is easy to see that minimizing this new function will yield the same estimates as maximizing entropy, that is the geometric family (Lefkovitch, [8]) of distributions having maximum entropy will be obtained.

It follows immediately that if the measure is not uniform, another family will be obtained. One important example is given by $\pi_i = 1/i!$, which can be recognised as being the counting measure of the Poisson distribution. In fact, one of the characterizations of this distribution (Guiaşu, [2]) is that it minimizes the **cross-entropy** with the normalized measure, i.e. where π_i is replaced by $\pi_i / \Sigma\pi_i$ and c by 1 in expression (2). Expression (2) with the standardised measure is known as cross-entropy; its relationship with the Kullback-Leibler measure is well known, and $2n\times$(cross-entropy) has an asymptotic χ^2 distribution. Notice that with this measure there is no need to specify any further constraints in order to obtain the probabilities associated with the Poisson distribution. However, if second, third, ... order terms are also included in the constraints, maximum entropy distributions belonging to the Poisson family, as

defined by Lefkovitch [8], are obtained, including some with smaller variance, while others with a larger variance may well provide a useful set of probabilities in situations where the negative binomial or Taylor's power law, Taylor [10], has often been employed. In like manner, the measure can be such that for some subset of $\{i\}$, $p_i = 0$, and so a very wide class of distributions exist within the family. This is further enlarged when it is recognised that although the number of cells in the geometric and Poisson families are unbounded, there is no reason for excluding finite sets, such as in dice, and so obtain multinomial probabilities with different measures.

4. The Genstat Procedure

The examples above give rise to a set of linear constraints in the probabilities, and to a non-linear objective function, which is either entropy or cross-entropy. In most circumstances, a numerical solution is needed to obtain the probabilities. Procedures to do this have been published several times (see the citations in Lefkovitch, [8], Section 4); that used in the Genstat 5 procedure in the Appendix is based on the method given by Johnson [4], but modified by replacing the computation of $(\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T$ by $\mathbf{A}^-$, where $\mathbf{X}^-$ denotes a generalized inverse (via a singular decomposition) of $\mathbf{X}$. This modification results in an increase in numerical accuracy, because of the reduction in the number of computational steps in each cycle, and since there is protection against singularity arising from redundant constraints, there is no need for any preprocessing. The Genstat procedure assumes that all constraints are equal, and so standardises the latter so that each of the expected values is unity. If the constraints are consistent with an exponential family, this standardisation has no impact on the solution.

While the procedure gives protection against redundancy in the constraints, there is none against the possibility of their inconsistency, which can occur if the expected values are wrongly specified, or if the probabilities do not belong to an exponential family. In these circumstances, the computed solution is the centre of attraction (Jupp and Mardia, [5]) of the algorithm, whch can be considered to be the closest exponential family approximation to the true distribution.

A listing of the procedure is given in the Appendix, together with a numerical example (Example 3). The procedure may also be used to provide discrete approximations to continuous densities.

5. References

[1] Feller, W.
An Introduction to Probability Theory and its Applications. Vol. 2.
Wiley, New York, 1971.

[2] Guiaşu, S.
Information theory with applications.
McGraw-Hill, New York, 1977.

[3] Jaynes, E.T.
On the rationale of maximum entropy methods.
Proc. IEEE, 70, pp. 939-952, 1982.

[4] Johnson, R.W.
Determining probability distributions by maximum entropy and minimum cross-entropy.
APL79 conference Proceedings, pp. 24-29, 1979.

[5] Jupp, P.E. and Mardia, K.V.
A note on the maximum-entropy principle.
Scand. J. Statist. 10, pp. 45-47, 1983.

[6] Kagan, A.M., Linnik, Yu.V. and Rao, C.R.
Characterization problems in mathematical statistics.
Wiley, New York, 1973.

[7] Lefkovitch, L.P.
Entropy and set covering.
Information Sciences, 36, pp. 283-294, 1985.

[8] Lefkovitch, L.P.
Estimation by maximum entropy subject to second-order conditions.
Biom. J., 31, pp. 75-91, 1989.

[9] Shore, J.E. and Johnson.
Axiomatic derivation of the principle of maximum entropy and the principle of maximum cross entropy.
IEEE Trans. Inform. Theory, IT-26, pp. 26-37, 1980.

[10] Taylor, L.R.
Assessing and interpreting the spatial distributions of insect populations.
Ann. Rev. Entomol., 29, pp. 321-357, 1984.

[11] Van Campenhout, J.M. and Cover, T.M.
Maximum entropy and conditional probability.
IEEE Trans. Inf. Theory IT29, pp. 483-489, 1981.

6. The Procedure

Editors' Note: this procedure uses the SET parameter of the PARAMETER directive introduced in Release 2.

```
            PROCEDURE 'CRSENT'
" To obtain the minimum cross-entropy probabilities for a
  discrete distribution having M cells, subject to N constraints.
"
OPTION    'MAXCYCLE', "SCALAR : THE MAXIMUM NUMBER OF ITERATIONS " \
          'TOLERANCE' "SCALAR : A NUMERICAL VALUE FOR CONVERGENCE " \
                      ; MODE=P ; DEFAULT=10, 0.0001

PARAMETER 'N',          "I:  THE NUMBER OF CONSTRAINTS (SCALAR)" \
     'M',         "I:  THE NUMBER OF CELLS (SCALAR)" \
     'CONS',      "I:  THE CONSTRAINTS (N BY M MATRIX)" \
     'RHS',       "I:  THE RIGHT-HAND-SIDE(S) (N BY 1 MATRIX)" \
     'PRIOR',     "I:  PRIOR PROBABILITIES (OR MEASURES) (M BY 1 MATRIX)" \
         "(A UNIFORM PRIOR MUST BE PROVIDED IN THE ABSENCE OF ANY OTHER)" \
     'CYCLE',     "O:  THE ACTUAL NUMBER OF ITERATIONS (SCALAR)" \
     'PROB',      "O:  THE MINIMUM CROSS-ENTROPY ESTIMATE OF THE
                       PROBABILITIES (M BY 1 MATRIX)" \
     'ENTROPY',   "O:  THE ENTROPY IN NATS (SCALAR)" \
     'XENTROPY'   "O:  THE CROSS-ENTROPY WITH THE PRIOR IN NATS (SCALAR)" \
       ; MODE=P; SET = 5(yes), 4(no)

" Declare necessary structures "

MATR  [ROWS=N;COLU=M] C & [ROWS=N;COLU=N] W & [ROWS=M;COLU=N] A \
   &  [ROWS=N;COLU=1] U & [ROWS=M;COLU=1] R,Q,PROB
DIAG  [ROWS=N] S & [ROWS=M] T
SCAL  CYCLE,ENTROPY,XENTROPY,Z

" Computation begins "

EQUATE  OLDS=RHS ; NEWS=S
CALC    C = ((1/S)*+CONS)-1 & Z = SUM(PRIOR) & R = SQRT(PRIOR/Z) & PROB = R
   &    U = 0 & CYCLE=0
FOR     [NTIMES=MAXCYCLE]
   EQUATE  OLDS = PROB ; NEWS = T & OLDS = PROB ; NEWS = Q
   CALC    A = TRANS(C*+T)
   SVD     A; SING = S ; LEFT = A ; RIGHT = W
   CALC    U = U-W*((S.NE.0)/(S+(S.EQ.0)))*+(TRAN(A))*+PROB
      &    PROB = R*EXP(0.5*(TRAN(C)*+U)) & CYCLE = CYCLE+1
   EXIT    SUM(ABS(PROB-Q))/M < TOLERANCE
ENDF
CALC    PROB = PROB*PROB & PROB = PROB/SUM(PROB)
   &    ENTROPY = -SUM(PROB*LOG(PROB)) & XENTROPY = SUM(PROB*LOG(PROB*Z/PRIOR))
DELETE  C,W,A,U,R,Q,S,T,Z
ENDP
```

7. Appendix

The Main Program

```
SCALAR [VALU=2] NN & [VALU=6] MM & [VALU=20] MAXT & [VALU=0.0001] TOLL
MATRIX [ROWS=2;COLU=6;VALU=1,2,3,4,5,6,1,1,1,1,1,1] CNSTRS
   &  [ROWS=2;COLU=1;VALU=3.5,13.0] EXVAL
   &  [ROWS=6;COLU=1;VALU=6(0.166667)] PRR

"Print the size of the problem, maximum iterations, tolerance, constraints,
  the rhs (expected values) and the prior"

PRINT NN ; DECI=0 & MM ; DECI=0 & MAXT ; DECI=0
   &  TOLL  & CNSTRS & EXVAL & PRR

CRSENT [MAXCYCLE=MAXT; TOLERANCE=TOLL] N=NN; M=MM; CONS=CNSTRS;\
  RHS=EXVAL; PRIOR=PRR; PROB=PRBS; CYCLE=ITS ; ENTROPY=ENTNATS; XENTROPY=XENT

"Print the probabilities, the number of iterations actually used,
the entropy and cross entropy in nats"

PRINT PRBS ; DECI=4 & ITS ; DECI=0 & ENTNATS ; DECI=4 & XENT ; DECI=4
STOP
```

Edited Output

```
Input:  NN = 2;  MM = 6;  MAXT = 20; TOLL = 0.0001

        CNSTRS
            1          2          3          4          5          6
1       1.000      2.000      3.000      4.000      5.000      6.000
2       1.000      1.000      1.000      1.000      1.000      1.000

        EXVAL
1        4.500
2        1.000

        PRR
1       0.1667
2       0.1667
3       0.1667
4       0.1667
5       0.1667
6       0.1667

Output:  ITS = 4;  ENTNATS = 1.6136;  XENT = 0.1782
        PRBS
1       0.0544
2       0.0788
3       0.1142
4       0.1654
5       0.2398
6       0.3475
```

Minimization of a Function

P W Lane
AFRC Institute of Arable Crops Research
Rothamsted Experimental Station
Harpenden
Hertfordshire
United Kingdom *AL5 2JQ*

Genstat provides the ability to search for parameter values that minimize a function of those parameters. The method of doing this is described in the Genstat Manual, Chapter 8, Section 8.6.4. However, because that section is buried at the end of a chapter, and a chapter otherwise devoted to regression analysis at that, it may escape the notice of someone looking for a minimization technique. The index does contain entries for 'function minimization' and for 'optimization', but it is notoriously difficult to provide index entries to satisfy all the ways that may be tried to look up a subject.

This article is intended to summarize briefly the available facilities, leaving out the regression framework that surrounds them in the Manual. I also give details of some changes made in Genstat 5 Release 2, and of some errors in the Genstat Manual and in output printed by Genstat following a function minimization.

1. How to Minimize a Function

There are four directives that are usually needed to carry out a function minimization, though the first of these, the EXPRESSION directive, is optional in simple cases. Firstly, then, the form of the function should be given in the form of one or more Genstat expressions; the expressions should be in the form expected by the CALCULATE directive to calculate the function from current values of the parameters. All the parameters, and the function value itself, will be assumed by Genstat to be scalar structures.

For a simple example, consider the problem of minimizing a function of just one parameter, p say:

$$f(p) = 2 + p/4 - \log(p).$$

This can be solved algebraically to give the result $f = 1.614$ when $p = 4$, with second derivative 0.0625 at the solution, so it is easy to evaluate the results that Genstat gives. The required expression in Genstat is specified by the statement:

```
EXPRESSION [VALUE=(f=2+p/4-LOG(p))] ex
```

This uses the standard function LOG (natural logarithm) and the identifiers f and p. The outer pair of round brackets are optional here.

The function is minimized by the following three statements:

```
MODEL [FUNCTION=f]
RCYCLE p
FITNONLINEAR [CALCULATION=ex]
```

2. Output from the FITNONLINEAR Directive

The FITNONLINEAR statement above gives the following output, using Release 2 of Genstat.

```
***** Results of optimization *****

*** Minimum function value: ***

      1.61371


*** Estimates of parameters ***
                            sq. root of
                  estimate   2nd derivs
p                     4.00         5.64
```

The solution $p = 4$ has thus been correctly found, but the associated number 5.64 does not seem right. In fact, it is the heading that is wrong, because the quantity is, as intended, the square root of the inverse of the second derivative of half the function. The reason for giving this particular quantity is discussed in Sections 4 and 5.

The output can be modified by the PRINT option of FITNONLINEAR. The available settings are 'summary' and 'estimates', which give the two default sections of output already shown, 'monitoring' to give information about the search process, and 'correlations' to indicate the interdependence of the parameter estimates, as in a regression analysis. Here are the two extra sections of output for the simple function above.

```
*** Convergence monitoring ***

Cycle Eval Move     Function value     Current parameters
    0    1    0          2.2500000       1.00000
                              Steps      0.050000
    1    4    0          1.8779203       1.74989
    2    8    0          1.6144749        4.1590
    3   12    0          1.6137969        3.9462
    4   15    0          1.6137056        3.9994
                              Steps     0.0125000
    5   18    1          1.6137056        4.0002
    6   22    6          1.6137056        4.0001

                              Steps       0.39739
    1   26    0          1.6137056        4.0001

*** Scaled 2nd derivatives ***

estimate       ref     scaled 2nd derivatives

p                1     1.000
                                 1
```

These 'correlations' are of little use with a single parameter, but can show up potential problems of parameterization when there are more parameters. Their heading, 'Scaled 2nd derivatives', is also wrong and is discussed further in Sections 4 and 5.

After the minimization, the scalar *f* will store the minimum function value, and *p* will store the best parameter value.

3. Extensions for More Complicated Functions

If a function has more than one parameter, then a list of parameter names can be given in the RCYCLE statement. There is no formal limit to the number of parameters in a function to be minimized in Genstat, but the speed of solution and the likelihood of success decrease as the number of parameters increases.

When the function is more complicated than in this simple example, it may be difficult to specify in a single expression; in that case, a series of expressions may be used.

```
EXPRESSION [VALUE=(f1=LOG(p))] e[1]
& [VALUE=(f=2+p/4-f1)] e[2]
...
FITNONLINEAR [CALCULATION=e]
```

Here, the identifier e is of a pointer that points (automatically, by the syntax of the Genstat language) to the two expression structures e[1] and e[2].

The minimization process is done by an algorithm known as 'modified Newton-Raphson'. The modification is evidenced by the absence of the need to specify the derivatives of the function with respect to the parameters: these are estimated internally by a differencing method. The process is not guaranteed to succeed, and indeed is likely to fail with complicated functions unless careful thought is given to the parameterization of the function, to starting values for the parameters, to initial steplengths for the search process, and to upper and lower bounds to exclude values of the parameters that would cause the calculations to break down. Initial values, steplengths and bounds can all be specified in the RCYCLE directive:

```
RCYCLE p; STEP=0.1; LOWER=0; UPPER=1000; INITIAL=1
```

To find the solution, or perhaps one of several solutions, of a function, it may be necessary to try several starting values. It may be helpful to inspect a grid of values of the function, which can be formed by CALCULATE statements, or by use of the NGRID option of FITNONLINEAR in conjunction with the bounds parameters of RCYCLE:

```
RCYCLE p; LOWER=1; UPPER=11
FITNONLINEAR [PRINT=grid; NGRID=6; CALCULATION=e]
```

This gives the following output.

```
*** Grid of function values ***

          p      1.00      3.00      5.00      7.00      9.00     11.00

              2.250     1.651     1.641     1.804     2.053     2.352
```

4. Errors

There are some errors in the Genstat Manual and in the output from the FITNONLINEAR directive.

(1) Page 385, Section 8.6.4, Line 8
Replace 'log likelihood' by 'log-likelihood ratio'.

(2) Page 388, Line 4
Replace 'estimated matrix' by 'inverse of the estimated matrix'.

(3) Page 388, Line 7
Replace 'second-derivative matrix' by 'inverse of the second-derivative matrix'.

(4) Page 388, Line 9
Replace 'a likelihood' by 'of the form –log(likelihood ratio)'; but see (8) below for a further change.

(5) Output from PRINT=estimates
The title should read 'sq. root of inverse of 2nd derivs' rather than 'sq. root of 2nd derivs'; but see (9) below for a further change.

(6) Output from PRINT=correlations
The title should read 'Scaled inverse of 2nd derivatives' rather than 'Scaled 2nd derivatives'.

In Genstat 5 Release 2, the results printed after function minimization for the option setting PRINT=estimates were changed. The intention was to print standard errors of the parameters if the function minimized was actually a deviance function. The deviance is –2*log(likelihood ratio); for example, the deviance for the Normal distribution is the residual sum of squares. The standard errors are, however, based on the second derivatives of log(likelihood ratio), and this factor of 2 introduces differences to the output. Thus for Release 2, the following further changes should be made to the Manual:

(7) Page 386, Output from PRINT=estimates
Replace '0.579' by '0.820' and '1.16' by '1.64'.

(8) Page 388, Line 9
Instead of (4) above, replace 'a likelihood' by 'a deviance, of the form –2*log(likelihood ratio)'.

(9) Output from PRINT=estimates
Instead of (5) above, the title should read 'sq. root of twice the inverse of 2nd derivs' rather than 'sq. root of 2nd derivs'.

5. Interpretation of the Output

As a result of the statistically motivated change in Release 2, the estimates of variability for arbitrary functions of parameters, like the simple example in this article, are now based on the inverse second-derivative matrix of half the target function. This is not actually as perverse as it may seem. One use of the second-derivative of the function at the solution, p_0 say, is to provide an approximate form of the function in the neighbourhood of the solution, using the Taylor expansion:

$$f(p) = f(p_0) + (p-p_0)^2 * f''(p_0)/2 + \ldots$$

This expansion also has a factor of 2! So using the quantity displayed by Genstat, called *se* say, the formula reduces to:

$$f(p) = f(p_0) + \{(p-p_0)/se\}^2 + \ldots$$

So we can approximate $f(4.5)$ by $1.61371 + \{0.5/5.64\}^2 = 1.622$; the actual value is of course easy to evaluate here as 1.62092. The *se* can be seen to be the change in p that gives a one unit increase in this approximation of the target function near the solution.

Regression Analyses for Multicollinear Data Using Genstat

A J Rook and M S Dhanoa
AFRC Institute of Grassland and Environmental Research
Hurley
Maidenhead
Berks
United Kingdom SL6 5LR

1. Introduction

A major problem arises with ordinary least-squares (OLS) multiple regression analyses when there is collinearity among the explanatory variables. This leads to unstable estimates of the regression coefficients which are difficult to interpret in terms of the underlying causal processes and results in poor prediction in independent data sets. Selection of variables by stepwise procedures is also less efficient in collinear data. A number of methods have been proposed to overcome this problem including principal component regression and ridge regression (Hoerl and Kennard, [2]). This paper summarises these methods and describes their implementation in a Genstat procedure.

2. Statistical Methods

Apart from examination of correlation coefficients between the explanatory variables, there are further tests for the degree of collinearity. The variance inflation factors (VIF), which are the diagonal elements of the inverse of the correlation matrix, may be examined. VIF > 10 indicate severe collinearity (Chatterjee and Price, [1]). The ratio of the squared error in the ordinary least-squares regression coefficients to the expected squared error if the data were orthogonal

$$R_L = \sigma^2 \sum_{i=1}^{p} \frac{\mathrm{VIF}_i}{p\sigma^2} = \sum_{i=1}^{p} \frac{\mathrm{VIF}_i}{p}$$

gives an overall measure of the degree of collinearity in the data. R_i values in excess of 5 indicate severe collinearity (Chatterjee and Price, [1]).

Let the regression model be represented by

$$Y = X\beta + u$$

where Y is an $n\times1$ vector of observations on a response variable, X is an $n\times p$ matrix of observations on p explanatory variables, β is a $p\times1$ vector of regression coefficients and u is $n\times1$ vector of residuals. Let X and Y be scaled such that $X'X$ and $X'Y$ are matrices of correlation coefficients. The least-squares estimator of β is

$$\hat{\beta} = (X'X)^{-1}X'Y.$$

It can be shown that the total mean square error

$$E[(\hat{\beta}-\beta)'(\hat{\beta}-\beta)] = \sigma^2 \sum_{j=1}^{p} \lambda_j^{-1}$$

where $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p$ are the ordered latent roots of $X'X$. It follows that when one or more latent roots are small the total mean square error is large indicating imprecision in the least-squares estimates. The ith latent root, λ_i, may be viewed as the sample variance of the ith principal component. If $\lambda_i = 0$ then all observations on the corresponding component are also 0. Since the principal component is a linear function of the original variables a latent root ≈ 0 indicates an approximate linear dependence among the original explanatory variables. By examining the latent vectors associated with small latent roots it is possible to identify which of the original variables give rise to serious collinearity.

The original regression model can be restated in terms of the principal components as

$$Y = W\alpha + u$$

where W is an $n\times p$ matrix of principal component scores. It is thus possible to calculate regression coefficients (α) in terms of the principal components and transform them to the coefficients on the original scale (β). The collinearity associated with principal components

with small latent roots may be removed by excluding these components from the regression. The reduced model can then be transformed to the original scale. Estimated regression coefficients produced in this way are biased, since some information is excluded, but they should be more stable and more in line with theoretical expectations.

Ridge regression (Hoerl and Kennard, [2]) allows a unified approach to the detection of collinearity and the estimation of new coefficients to overcome the problem. The estimates produced are biased but have a smaller mean square error than OLS estimates and are thus more stable. Prediction of values outwith the estimation data is thus more precise. Ridge regression can also be applied to the maximal model and used to assist in variable selection.

The ridge estimates are indexed by a parameter $k > 0$ such that

$$\hat{\beta}(k) = (X'X+kI)^{-1}X'Y = (X'X+kI)^{-1}X'X\hat{\beta}$$

total mean square error is

$$E[(\hat{\beta}(k)-\beta)'(\hat{\beta}(k)-\beta)] = \sigma^2\text{trace}[(X'X+kI)^{-1}X'X(X'X+kI)^{-1}] + k^2\beta'(X'X+kI)^{-2}\beta$$
$$= \sigma^2\sum_{i=1}^{p}\lambda_i(\lambda_i+k)^{-2} + k^2\beta'(X'X+kI)^{-2}\beta.$$

The first term in this equation is the total variance of the $\hat{\beta}(k)$ while the second term is the square of the bias. The aim in ridge regression is to select a value of k for which the reduction in total variance is greater than the increase in bias. This may be achieved by examination of the ridge trace, a graph of the p regression coefficients plotted against k. Large fluctuations in estimated coefficients in response to small increments in k indicate instability. Guidelines for selection of k (Hoerl and Kennard, [2]) are: (1) at a certain value of k the system will stabilize; (2) coefficients will not have theoretically unreasonable values or improper signs; (3) the residual sum of squares will not have been excessively inflated. Vinod [3] introduced a quantitative measure of the stability of the ridge trace called the 'index of stability of relative magnitudes'

$$\text{ISRM} = \sum_i[(p(\lambda_i/(\lambda_i+k))^2/\bar{s}\lambda_i)-1]^2$$

where $\bar{s} = \sum_i \lambda_i/(\lambda_i+k)^2$. For orthogonal data ISRM = 0. Vinod [3] also pointed out that if $|\hat{\beta}_i| > |\hat{\beta}_j|$ in standardised orthogonal data then $|t_i| = |\hat{\beta}_i|/\text{s.e.}(\hat{\beta}_i) > |t_j|$ should also hold. This may be tested by calculating the correlation between the $|\hat{\beta}_i|$ and the $|t_i|$ which Vinod [3] termed the 'numerical largeness of more significant regression coefficients' (NLMS). For orthogonal data NLMS = 1. k is thus chosen so as to minimise residual sums of squares and ISRM and maximise R^2 and NLMS, taking account of the theoretical expectations for the values of the parameters.

Ridge regression can also be applied to the maximal model to assist in variable selection. The ridge trace of the maximal model is examined and variables are eliminated: (1) if they are unstable and tend to 0 (i.e. lose their predicting power); (2) if they are stable but very small; (3) if, after 1 and 2, remaining variables are unstable. The final equation is then estimated from the full model using an appropriate value of k rather than by least-squares using a reduced model.

Both principal component regression and ridge regression de-emphasize the minor principal axes in order to overcome collinearity. However in some circumstances the response variate may be highly correlated with these axes and the methods will then perform poorly. Vinod [3] proposed an overall quantitative measure of the suitability of data for these methods termed the 'positive correlation spread association'. This is the simple correlation between the absolute values of the correlations of the standardised response variate with each principal component and the square root of the latent roots associated with each component. This should be close to 1 for these methods to be justified.

Vinod [3] suggested replacing the scale k in the ridge trace with a new scale termed the multicollinearity allowance

$$m = p - \sum_i \lambda_i/(\lambda_i+k)$$

m may be interpreted as the assigned deficiency in the rank of $(X'X)$, that is it is an index of the degree of de-emphasis of the minor principal components.

3. Description of Procedure

Procedure RIDGE has two parameters. The paremeter Y is used to pass the response variate to the procedure. The parameter X is set to a pointer which contains the explanatory variates. Because the procedure makes use of Genstat's matrix algebra functions, the parameters must neither be restricted nor contain missing values.

The procedure also has two options. Option PRINT controls the analysis and printing of results while option DGRAPH controls the production of high-resolution graphics. Setting PRINT=corr produces the correlation matrix among the explanatory variates using the CORRELATE directive, the variance inflation factors and R_L. Setting PRINT=pcp produces a principal component analysis using the statement

```
PCP [PRINT=loading,roots; method=corr]
```

The standardised response variate is then regressed on the principal component scores using the usual MODEL and FIT directives. The correlations of the standardised response variate with each principal component are also calculated and printed explicitly to increase the clarity of the output. The coefficients of the three smallest principal components are then set to 0 consecutively and each time the regression coefficients are transformed back to those of the original variate both on standardised and unstandardised scales and printed. The positive correlation spread association is also printed.

Setting PRINT=ridge calculates ridge regression coefficients on the standardised scale for values of k between 0 and 1. These are printed in parallel with k and m. The standard errors of the coefficients and the coefficients on the unstandardised scale are printed in subsequent blocks. Finally the residual sum of squares, R^2, total variance of the ridge coefficients, ISRM and NLMS are printed for each k.

Setting DGRAPH=yes produces high-resolution ridge traces of the ridge coefficients against both k (Hoerl-Kennard trace) and m (Vinod trace). These are output to the graphics device set up prior to the call by the user. The trace for each explanatory variate is labelled by its ordinal position in the pointer X.

4. Example

The data for this example are taken from Chatterjee and Price [1] and are shown in Table 1. The procedure was run using the program shown below

```
JOB 'test program for procedure ridge'
UNITS [NVALUES=11]
READ import,doprod,stock,consum
POINTER [VALUES=doprod,stock,consum] indep
OPEN 'rtrace.dat'; CHANNEL=9; FILE=graphics
DEVICE 9
RIDGE [PRINT=corr,pcp,ridge; DGRAPH=yes] Y=import; X=indep
ENDJOB
STOP
```

Year	Imports	Domestic production	Stock formation	Domestic consumption
49	15.9	149.3	4.2	108.1
50	16.4	161.2	4.1	114.8
51	19.0	171.5	3.1	123.2
52	19.1	175.5	3.1	126.9
53	18.8	180.8	1.1	132.1
54	20.4	190.7	2.2	137.7
55	22.7	202.1	2.1	146.0
56	26.5	212.4	5.6	154.1
57	28.1	226.1	5.0	162.3
58	27.6	231.9	5.1	164.3
59	26.3	239.0	0.7	167.6

Table 1

Milliard of French francs (Chatterjee and Price, [1])

The output from the example is shown in the Appendix with the output from DGRAPH=yes as shown in Figure 1.

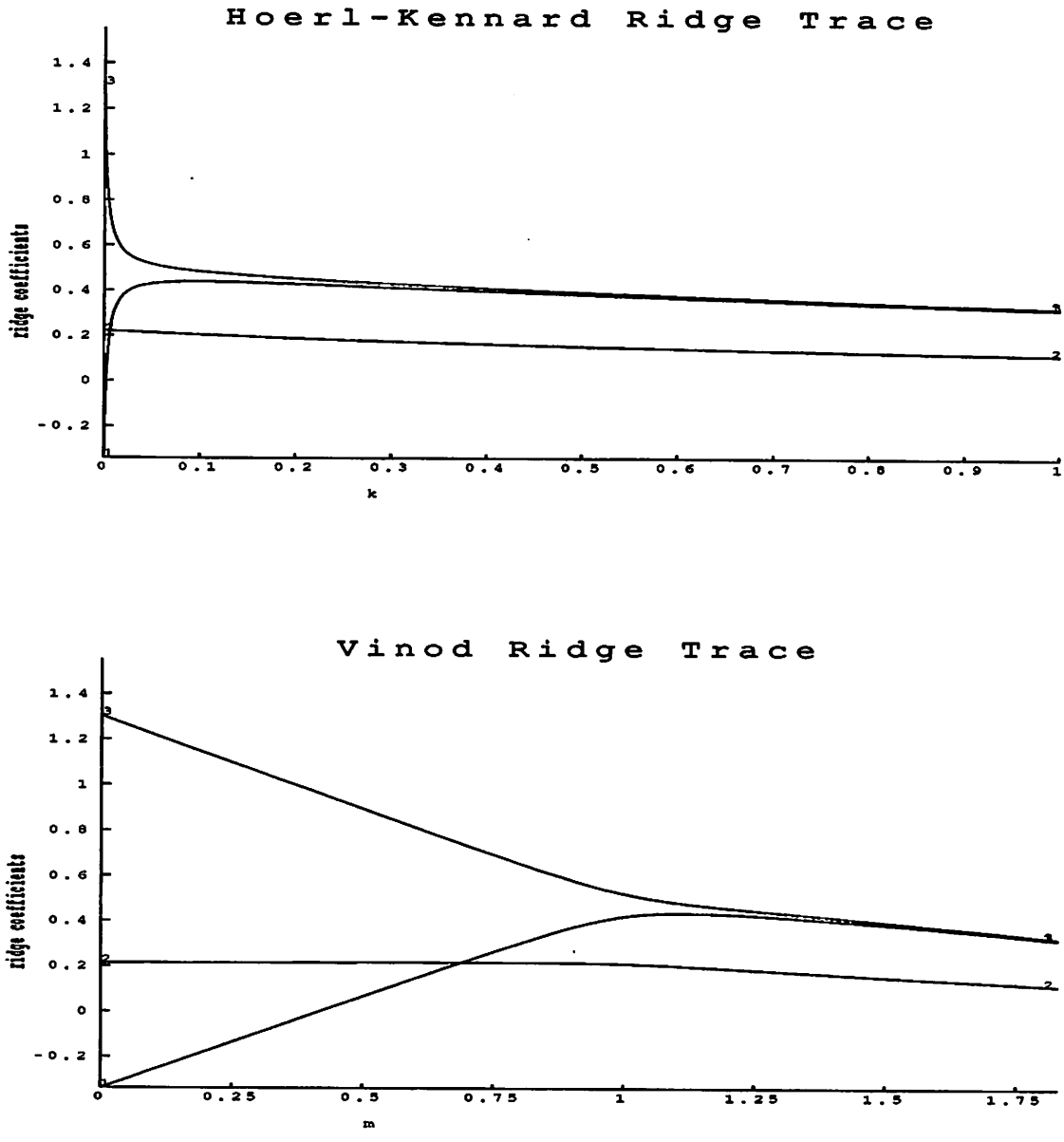


Figure 1

5. References

[1] Chatterjee, S. and Price, B.
Regression Analysis by Example.
Wiley, New York, 1977.

[2] Hoerl, A.E. and Kennard, R.W.
Ridge regression: biased estimation for nonorthogonal problems.
Technometrics, 12, pp. 55-67, 1970.

[3] Vinod, H.D.
Application of new ridge regression methods to a study of Bell system scale economics.
J. Amer. Statist. Assoc., 71, pp. 835-841, 1976.

6. Appendix

Edited output from example.

The output from PRINT=corr is shown below.

```
 ******* REGRESSION ANALYSES FOR MULTICOLLINEAR DATA *******


*** Correlation matrix ***

       doprod  1.000
        stock  0.026  1.000
       consum  0.997  0.036  1.000

              doprod  stock consum



 *** Variance Inflation Factors ***

      doprod       stock      consum
       186.0       1.019       186.1

 *** Ratio of squared error in OLS estimates of regression
     coefficients to error if data were orthogonal ***
       124.4
```

The output from PRINT=pcp follows.

```
*****  Principal components analysis  *****

***  Latent Roots  ***

          roots
              1           2           3
          1.999       0.998       0.003

***  Percentage variation  ***

          roots
              1           2           3
          66.64       33.27        0.09

***  Trace  ***

        trace
        3.000

***  Latent Vectors (Loadings)  ***

                   vectors
                         1           2           3
         indep
        doprod     0.70633     0.03569    -0.70698
         stock     0.04350    -0.99903    -0.00697
        consum     0.70654     0.02583     0.70720


***** Regression Analysis *****

 Response variate: stany
     Fitted terms: pcp[1], pcp[2], pcp[3]


*** Summary of analysis ***

              d.f.        s.s.        m.s.      v.r.  F pr.
Regression       3     9.91896     3.30632    326.41  <.001
Residual         8     0.08103     0.01013
Total           11    10.00000     0.90909

Percentage variance accounted for 99.0

* MESSAGE: The following units have high leverage:
                  11        0.70
```

*** Estimates of regression coefficients ***

	estimate	s.e.	t	t pr.
pcp[1]	2.1819	0.0712	30.65	<.001
pcp[2]	-0.605	0.101	-6.01	<.001
pcp[3]	3.67	1.94	1.89	0.095

*** Coefficients of original variables excluding effect of 1, 2 or 3 smallest p

One Principal component(s) excluded

Standardised scale

doprod	stock	consum
1.520	0.6993	1.526

Original scale

int	doprod	stock	consum
-76.21	0.2301	1.927	0.3360

Two Principal component(s) excluded

Standardised scale

doprod	stock	consum
1.541	0.09491	1.542

Original scale

int	doprod	stock	consum
-71.83	0.2334	0.2615	0.3395

Three Principal component(s) excluded

Standardised scale

doprod	stock	consum
0	0	0

Original scale

int	doprod	stock	consum
21.89	0	0	0

*** Correlation of Standardised Response Variable with Principal Component Scor

pcp[1]	pcp[2]	pcp[3]
0.9756	-0.1911	0.0602

*** Positive correlation spread association ***

0.8180

Output from PRINT=ridge follows.

*** Ridge Coefficients ***

k	m	doprod	stock	consum
0.0000	0.000	-0.3393	0.2130	1.3027
0.0010	0.272	-0.1174	0.2150	1.0802
0.0020	0.429	0.0097	0.2161	0.9525
0.0030	0.532	0.0922	0.2167	0.8696
0.0040	0.604	0.1499	0.2171	0.8114
0.0050	0.658	0.1925	0.2173	0.7683
0.0060	0.699	0.2253	0.2174	0.7351
0.0070	0.733	0.2512	0.2175	0.7086
0.0080	0.760	0.2723	0.2175	0.6871
0.0090	0.783	0.2897	0.2174	0.6692
0.0100	0.803	0.3043	0.2174	0.6541
0.0200	0.911	0.3786	0.2161	0.5751
0.0300	0.962	0.4060	0.2144	0.5430
0.0400	0.995	0.4196	0.2127	0.5249
0.0500	1.021	0.4271	0.2109	0.5128
0.0600	1.043	0.4315	0.2091	0.5038
0.0700	1.062	0.4341	0.2073	0.4968
0.0800	1.080	0.4356	0.2055	0.4909
0.0900	1.097	0.4363	0.2038	0.4858
0.1000	1.112	0.4364	0.2021	0.4813
0.2000	1.245	0.4265	0.1864	0.4499

0.3000	1.353	0.4112	0.1730	0.4274
0.4000	1.446	0.3957	0.1614	0.4082
0.5000	1.528	0.3809	0.1513	0.3911
0.6000	1.602	0.3669	0.1424	0.3755
0.7000	1.668	0.3538	0.1345	0.3613
0.8000	1.727	0.3415	0.1275	0.3482
0.9000	1.782	0.3300	0.1212	0.3360
1.0000	1.831	0.3192	0.1155	0.3247

*** Standard Errors of Ridge Coefficients ***

k	m	doprod	stock	consum
0.0000	0.000	0.43405	0.03213	0.43418
0.0010	0.272	0.31664	0.03196	0.31673
0.0020	0.429	0.24933	0.03187	0.24941
0.0030	0.532	0.20572	0.03181	0.20578
0.0040	0.604	0.17517	0.03176	0.17522
0.0050	0.658	0.15260	0.03172	0.15264
0.0060	0.699	0.13524	0.03168	0.13528
0.0070	0.733	0.12149	0.03164	0.12152
0.0080	0.760	0.11033	0.03161	0.11036
0.0090	0.783	0.10109	0.03157	0.10112
0.0100	0.803	0.09333	0.03154	0.09336
0.0200	0.911	0.05381	0.03122	0.05382
0.0300	0.962	0.03901	0.03091	0.03901
0.0400	0.995	0.03149	0.03062	0.03149
0.0500	1.021	0.02706	0.03032	0.02706
0.0600	1.043	0.02421	0.03004	0.02420
0.0700	1.062	0.02225	0.02975	0.02224
0.0800	1.080	0.02083	0.02948	0.02083
0.0900	1.097	0.01978	0.02921	0.01977
0.1000	1.112	0.01896	0.02894	0.01895
0.2000	1.245	0.01559	0.02653	0.01558
0.3000	1.353	0.01438	0.02449	0.01437
0.4000	1.446	0.01359	0.02273	0.01358
0.5000	1.528	0.01295	0.02122	0.01294
0.6000	1.602	0.01240	0.01989	0.01240
0.7000	1.668	0.01191	0.01872	0.01191
0.8000	1.727	0.01147	0.01768	0.01146
0.9000	1.782	0.01106	0.01675	0.01105
1.0000	1.831	0.01068	0.01591	0.01067

*** Ridge Coefficients on Original Scale ***

k	m	int	doprod	stock	consum
0.0000	0.000	-10.128	-0.05139	0.5869	0.2868
0.0010	0.272	-9.841	-0.01778	0.5924	0.2379
0.0020	0.429	-9.670	0.00148	0.5953	0.2097
0.0030	0.532	-9.553	0.01396	0.5970	0.1915
0.0040	0.604	-9.467	0.02270	0.5980	0.1787
0.0050	0.658	-9.398	0.02915	0.5986	0.1692
0.0060	0.699	-9.342	0.03412	0.5989	0.1619
0.0070	0.733	-9.294	0.03805	0.5991	0.1560
0.0080	0.760	-9.252	0.04124	0.5991	0.1513
0.0090	0.783	-9.215	0.04388	0.5990	0.1474
0.0100	0.803	-9.181	0.04609	0.5989	0.1440
0.0200	0.911	-8.928	0.05734	0.5954	0.1266
0.0300	0.962	-8.734	0.06150	0.5908	0.1196
0.0400	0.995	-8.558	0.06354	0.5859	0.1156
0.0500	1.021	-8.392	0.06469	0.5809	0.1129
0.0600	1.043	-8.231	0.06536	0.5760	0.1109
0.0700	1.062	-8.074	0.06575	0.5711	0.1094
0.0800	1.080	-7.920	0.06598	0.5662	0.1081
0.0900	1.097	-7.768	0.06608	0.5615	0.1070
0.1000	1.112	-7.618	0.06610	0.5568	0.1060
0.2000	1.245	-6.217	0.06459	0.5135	0.0991
0.3000	1.353	-4.952	0.06228	0.4766	0.0941
0.4000	1.446	-3.800	0.05994	0.4446	0.0899
0.5000	1.528	-2.744	0.05769	0.4168	0.0861
0.6000	1.602	-1.773	0.05557	0.3924	0.0827
0.7000	1.668	-0.877	0.05359	0.3706	0.0796
0.8000	1.727	-0.047	0.05173	0.3513	0.0767
0.9000	1.782	0.724	0.04999	0.3338	0.0740
1.0000	1.831	1.442	0.04835	0.3181	0.0715

```
*** RIDGE Regression Fit and Stability Parameters ***
```

k	m	RSS	Rsq	TVARB	ISRM	NLMS
0.0000	0.000	0.0810	0.9919	0.37794	5.928	0.1971
0.0010	0.272	0.0837	0.9916	0.20160	5.865	0.2445
0.0020	0.429	0.0876	0.9912	0.12539	5.784	0.2761
0.0030	0.532	0.0911	0.9909	0.08568	5.686	0.2545
0.0040	0.604	0.0940	0.9906	0.06240	5.572	0.2407
0.0050	0.658	0.0964	0.9904	0.04759	5.443	0.2351
0.0060	0.699	0.0984	0.9902	0.03759	5.301	0.2386
0.0070	0.733	0.1000	0.9900	0.03053	5.147	0.2518
0.0080	0.760	0.1015	0.9899	0.02535	4.983	0.2752
0.0090	0.783	0.1027	0.9897	0.02144	4.810	0.3091
0.0100	0.803	0.1038	0.9896	0.01842	4.630	0.3533
0.0200	0.911	0.1102	0.9890	0.00677	2.745	0.9037
0.0300	0.962	0.1139	0.9886	0.00400	1.312	0.9951
0.0400	0.995	0.1170	0.9883	0.00292	0.539	1.0000
0.0500	1.021	0.1201	0.9880	0.00238	0.235	0.9998
0.0600	1.043	0.1234	0.9877	0.00207	0.189	0.9995
0.0700	1.062	0.1271	0.9873	0.00187	0.264	0.9994
0.0800	1.080	0.1310	0.9869	0.00174	0.386	0.9994
0.0900	1.097	0.1353	0.9865	0.00164	0.518	0.9995
0.1000	1.112	0.1400	0.9860	0.00156	0.644	0.9995
0.2000	1.245	0.2052	0.9795	0.00119	1.299	0.9998
0.3000	1.353	0.2981	0.9702	0.00101	1.454	0.9999
0.4000	1.446	0.4112	0.9589	0.00089	1.491	1.0000
0.5000	1.528	0.5385	0.9462	0.00079	1.495	1.0000
0.6000	1.602	0.6756	0.9324	0.00070	1.490	1.0000
0.7000	1.668	0.8191	0.9181	0.00063	1.482	1.0000
0.8000	1.727	0.9667	0.9033	0.00058	1.475	1.0000
0.9000	1.782	1.1163	0.8884	0.00052	1.469	1.0000
1.0000	1.831	1.2666	0.8733	0.00048	1.465	1.0000

Editors' Note: the final form of the procedure RIDGE may be different from that illustrated above