Data Ellipses, HE Plots and Reduced-Rank Displays for Multivariate Linear Models: SAS Software and Examples

Michael Friendly
York University, Toronto

Abstract

This paper describes graphical methods for multiple-response data within the framework of the multivariate linear model (MLM), aimed at understanding what is being tested in a multivariate test, and how factor/predictor effects are expressed across multiple response measures.

In particular, we describe and illustrate a collection of SAS macro programs for: (a) Data ellipses and low-rank biplots for multivariate data, (b) HE plots, showing the hypothesis and error covariance matrices for a given pair of responses, and a given effect, (c) HE plot matrices, showing all pairwise HE plots, and (d) low-rank analogs of HE plots, showing all observations, group means, and their relations to the response variables.

Keywords: biplot, canonical discriminant plot, data ellipse, HE plot, HE plot matrix, multivariate analysis of variance, MANOVA, multivariate multiple regression, MMRA, SAS, scatterplot matrix.

1. Introduction

The classical univariate general linear model (LM), \( y = X\beta + \epsilon \), with \( \epsilon \sim \mathcal{N}(0, \sigma^2 I) \), is among the most generally useful inventions in the history of statistics. As is well known, the LM includes as special cases all forms of regression, analysis of variance (ANOVA), analysis of covariance (ANCOVA), and response surface models. Extensions of this basic model include generalized linear models, \( g(y) = X\beta + \epsilon \), such as Poisson regression, logistic regression and loglinear models, all with non-Gaussian, heteroscedastic error structures, and versions that substitute robust estimation for standard least squares.

The applied use of these LM family methods is also well-supported by a wide range of graphical methods, both for assessing departures of the data from model assumptions, and for assisting
the viewer in understanding and communicating the nature of effects. Such graphical methods, including QQ plots of residuals, spread-level plots, influence-leverage plots and so forth are widely implemented in many (if not most) statistical software systems; see Fox (1991, 1997); Friendly (1991) for descriptions and examples of these.

The classical LM also extends quite naturally to the multivariate response setting, at least for a multivariate normal collection of $p$ responses, $(y_1, y_2, \ldots, y_p) \equiv Y$. The multivariate linear model (MLM) then becomes $Y = XB + U$. Thus, multivariate analysis of variance (MANOVA) extends the ideas and methods of univariate ANOVA in simple and straightforward ways, just as multivariate multiple regression (MMRA) extends univariate multiple regression (MRA) to the multiple response setting. It is therefore somewhat surprising that corresponding graphical methods for multivariate responses are not widely developed, or at least, are not widely known and used.

This paper describes a collection of graphical methods for multivariate data in the context of the multivariate LM (Friendly 2007) aimed at understanding how variation is reflected in multivariate tests and showing how factor/predictor effects are expressed across multiple response measures. The principal new graphical methods we introduce (in what we call HE plots) concern the use of data and covariance ellipses to visualize covariation against multivariate null hypotheses ($H$) relative to error covariation ($E$). These also combine with older ideas of low-rank projections for multivariate data to give other displays that can provide simpler summaries of complex multivariate data than are available by other means.

We focus here on the implementation of these methods in SAS macros (many of which originated in Friendly (1991, 2000)) and illustrations of their use. There are available with documentation and examples at http://www.math.yorku.ca/SCS/sasmac/. The principal programs used here are:

- **biplot** Generalized biplot display of variables and observations
- **canplot** Canonical discriminant structure plots
- **ellipses** Plot bivariate data ellipses
- **heplot** Plot H and E matrices for a bivariate MANOVA effect
- **hemat** HE plots for all pairs of response variables
- **hemreg** Extract H and E matrices for multivariate regression
- **panels** Display a set of plots in a rectangular layout
- **outlier** Robust multivariate outlier detection
- **robcov** Calculate robust covariance matrix via MCD or MVE
- **scatmat** Scatterplot matrices

The outline of this paper is as follows: Section 2 provides a graphic overview of these methods and illustrations of how the SAS macros are used, with a sampler of these displays for a single, well-known data set. Section 3 describes some of the statistical, graphic and computational details on which these methods are based. Section 4 gives some further examples highlighting some additional features of these plots and software. An appendix provides documentation for some of the macro programs.

## 2. Overview examples
It is easiest to illustrate the graphical ideas first with relatively simple and straight-forward
data. The focus here is on understanding what the various plots can reveal about multivariate
samples and their interpretation in the context of a well-known example. We also explain
how some of these plots are produced using our macro programs.

For this purpose, we use Anderson’s (1935) classic data on four measures of sepal and petal
size in three species of iris flowers found in the Gaspé Peninsula, later used by Fisher (1936) in
his development of discriminant analysis. Data sets of this general structure ($g > 1$ groups or
populations, $p > 1$ measures) can be used to address a variety of questions within the frame-
work of the MLM: Do the mean vectors differ across groups (MANOVA)? If so, which groups
differ, and on which variables (contrasts)? Are the regression relations between variables—
slopes and intercepts— the same across groups (homogeneity of regression, ANCOVA)?

2.1. Data ellipses

Figure 1 shows a scatterplot matrix of these data. Each pairwise plot also shows the regression
lines for predicting the row ($y$) variable from the column ($x$) variable, separately for each iris
species.

In addition, each plot shows a 68% data ellipse for each species, a bivariate analog of the
“standard univariate interval,” $\bar{y} \pm 1s$, centered at the bivariate mean. These have the prop-
ties (Monette 1990) that their projections on any axis are proportional to standard deviations,
the regression lines for $y \mid x$ pass through the loci of vertical tangents, and their eccentricities
reflect the correlations. The data ellipses show clearly that the means, variances, correlations,
and regression slopes differ systematically across the three iris species in all pairwise plots.
The most obvious features shown here are the consistent ordering of the species means from
setosa to versicolor to virginica, but the data ellipses show that these also differ consistently
in variances and within-sample correlation.

Figure 1 is produced using the scatmat macro, as shown below. The %include statements
cause SAS to read the named data and macro files from pre-assigned directories. In the
%scatmat call, interp=rl adds linear regression lines and anno=ellipse causes the program
to invoke the ellipses macro to add data ellipses for each group in each panel of the plot.

```sas
%include data(iris);
%include macros(scatmat); /*-- or place scatmat.sas in SASAUTOS;
%scatmat(data=iris,
 var=SepalLen SepalWid PetalLen PetalWid,
 group=Species,
 symbols=triangle plus square,
 colors= blue red green,
 hsym=4, htitle=9,
 interp=rl, /* draw regression lines */
 anno=ellipse); /* and add data ellipses */
```

2.2. Partial plots

In many MLM contexts we may wish to study the covariation among responses after some
one or more predictor effects have been taken into account or “adjusted for,” \( \text{VAR}(Y \mid X) = \)
Figure 1: Scatterplot matrix of Anderson’s iris data, showing separate 68% data ellipses and regression lines for each species. Key– Iris setosa: blue, △; Iris versicolor: red, +; Iris virginica: green, □.

VAR(\(U\)). For example, Figure 2 shows a scatterplot matrix of residuals from the one-way MANOVA model \(\text{SepalLen SepalWid PetalLen PetalWid} = \text{Species}\) fit using PROC GLM. Graphically, this plot is equivalent to translating each of the separate species in Figure 1 to a common origin at (0, 0) in each sub-plot. Statistically, it provides a visualization of the within-cell covariance matrix, \(\text{VAR}(Y | X)\), proportional to what we call the \(E\) matrix in HE plots. Scaled to a correlation matrix, \(\text{VAR}(Y | X)\) gives partial correlations among the responses. The variation among the separate data ellipses indicate the extent to which the assumption of homogeneity of covariance matrices, required in the MLM is met in the data. The sub-plots in Figure 2 show that the iris species appear to differ substantially in their variances and covariances on these variables, more directly than in Figure 1.

```sas
*-- remove group means to view within-cell relations;
proc glm data=iris noprint;
```
Figure 2: Scatterplot matrix of within-cell residuals for the iris data, with 68% data ellipses and regression lines for each species. Key– *Iris setosa*: blue, △s; *Iris versicolor*: red, +; *Iris virginica*: green, □.

class species;
model SepalLen SepalWid PetalLen PetalWid = Species /nouni;
output out=resids
r=seplen sepwid petlen petwid;
run;

%scatmat(data=resids,
  var=SepLen SepWid PetLen PetWid, group=Species,
  names=SepalLen SepalWid PetalLen PetalWid,
  symbols=triangle plus square,
  colors= blue  red  green,
  hsym=4, htitle=9,
  interp=rl,
  anno=ellipse);
2.3. Biplots: Reduced-rank displays

Each scatterplot in Figure 1 is a 2D (marginal) projection of the 4D space. Instead of showing all pairwise views, it is often more useful to project the multivariate sample into a low-dimensional space (typically 2D or 3D) accounting for the greatest variation in the (total sample) data.

The biplot (Gabriel 1971, 1981) is one such display that is extremely useful for multivariate data and can be enhanced for multivariate LMs, particularly in the MANOVA setting. The name “biplot” comes from the fact that this technique displays the observations (as points) and variables (as vectors) in the same plot, in a way that depicts their joint relationships. The (symmetric) scaling of the biplot described here is equivalent to a plot of principal component scores for the observations, together with principal component coefficients for the variables in the same 2D (or 3D) space (see Figure 3). When there are classification variables dividing the observations into groups, we may also overlay data ellipses for the scores to provide a low-dimensional visual summary of differences among groups in means and covariance matrices.

![Figure 3: Enhanced biplot for iris data, showing observations (points) and variables (vectors), together with the 68% data ellipses (calculated in the reduced space) for each species (setosa: blue (1); versicolor: green (2); virginica: red (3)) and for all species (in gray).](image)

Figure 3 shows the biplot for the iris data. The 2D projection of the 4D dataset accounts for...
95% of the total variation, of which most (73%) is accounted for by the first dimension. In such plots, it is crucial that the axes are “equated,” so that the units on the horizontal and vertical axes have equal lengths, in order to preserve the standard interpretation of lengths and angles. With this scaling, the observations and variables may be understood as follows:

- the variable vectors have their origin at the mean on each variable, and point in the direction of positive deviations from the mean on each variable.
- the angles between variable vectors portray the correlations between them, in the sense that the cosine of the angle between any two variable vectors approximates the correlation between those variables (in the reduced space). Thus, vectors at right angles reflect correlations of zero, while vectors in the same direction reflect perfect correlations;
- the relative length of each variable vector indicates the proportion of variance for that variable represented in the low-rank approximation;
- the orthogonal projections of the observation points on the variable vectors show approximately the value of each observation on each variable;
- by construction, the observations, shown as principal component scores are uncorrelated, as may be seen from the total sample ellipse (gray ellipse in Figure 3);
- within-sample correlations, means, and variances in the reduced space are shown by the separate data ellipses, in relation to the grand mean \(Y\) at the origin, and in relation to the variable vectors.

The interpretation of Figure 3 is as follows: In the total sample, petal width and petal length are nearly perfectly correlated, and these are both highly correlated with sepal length; the two sepal variables are nearly uncorrelated. As well, the three iris species differ primarily along the first dimension, and so are ordered by increasing means on both petal variables (cf. Figure 1, panel 3:4 in row 3, column 4), but the variances and covariances differ as well.

Figure 3 is produced using (a) the biplot macro to obtain the component scores for the observations and coordinates for the variable vectors, (b) the ellipses macro to obtain the outlines of the 68% data ellipses for the species and the total sample, and (c) SAS graphics programming involving “Annotate data sets” to produce a customized graphic display.

The complete code for this figure is contained in the file bipliris.sas, included in the accompanying archive. Here, we show just portions to illustrate the flexibility of SAS graphic displays using our macros.

The biplot macro constructs generalized biplot displays for multivariate data, and for two-way and multi-way tables of either quantitative or frequency data. By default, it produces a 2D labeled plot of observations (points) and variables (vectors from the origin) in the reduced-rank space of the first two dimensions. It also produces an output data set (out=biplot) containing coordinates for the observations and variables and a SAS/Graph Annotate data set (anno=bianno) for drawing the variable vectors and labels on a plot.

```sas
[bipliris.sas ...]
*-- Obtain biplot scores (_type_='OBS') and variable vectors (_type_='VAR');
%biplot(data=iris,
    var=SepalLen SepalWid PetalLen PetalWid,
    id=id,
    std=std, /* observation ID, here, species number */
    scale=0.36, /* standardize to mean=0, var=1 */
    ...)
```
Here, we suppress the default plot (gplot=no) and post-process the output data sets. For example, we assign different colors to the observation points (_type_='OBS') for different species, because the biplot macro doesn’t handle grouped data.

Finally, the ellipses macro is applied to the observation scores on the two biplot dimensions to give another Annotate data set that draws the ellipses in Figure 3. The actual figure is drawn with PROC GPLOT using the Annotate data sets bianno and ellipses (code not shown here).
2.4. HE plots

For the multivariate linear model, any hypothesis test may be calculated from an analog of the univariate $F$, where $p \times p$ matrices, $H$ and $E$ play the roles of univariate sums of squares, $SS_H$ and $SS_E$. But, in the multivariate case, the variation against the null hypothesis ($H$) may be large in one or more dimensions relative to the error variation ($E$).

The HE plot provides a two-dimensional visualization of the size and shape of the $H$ matrix relative to the size of the $E$ matrix for any multivariate test. Figure 4 shows data ellipses for Sepal and Petal length in the iris data, and the corresponding view of the $2 \times 2$ portions of the $H$ and $E$ matrices for the model SepalLen SepalWid PetalLen PetalWid = Species testing whether the species means are equal on all four variables.

![Figure 4: Data and HE plots for iris data](image)

The ellipse for the $H$ matrix in this plot is equivalent to a data ellipse of the fitted values $\hat{y}_{ij} = \bar{y}_j$ under this model. The ellipse for the $E$ matrix is equivalent to the plot of residuals from this model shown in panel (1:2) in Figure 2. The interpretation is that the variation due to group means is very large compared to within-group variance, but that the mean variation is essentially one-dimensional, for these two variables.

In SAS, we use PROC GLM to calculate an output data set (outstat=stats) containing the $H$ and $E$ matrices for any linear hypothesis. (In general, the outstat= data set contains one $H$ matrix for each effect listed on the right-hand side of the MODEL statement)
proc glm data=iris outstat=stats noprint;
class species;
model SepalLen SepalWid PetalLen PetalWid = species / nouniss3;
run;

Figure 4(b) is then produced using the heplot macro using the stats data set for the \( H \) and \( E \) matrices and the iris data to find and plot the species means:

\[
\text{axis1 label=(a=90) order=(40 to 80 by 10);}
\]
\[
%heplot(data=iris, 
stat=stats, /* Data set containing H & E matrices */
var=PetalLen SepalLen, /* Variables to plot */
effect=species, /* Effect to plot */
vaxis=axis1);
\]

This idea can be extended to show the pairwise relations for all response variables, using the framework of a scatterplot matrix, plotting all pairs of response variables, in an HE plot matrix, as shown in Figure 5.

Comparing this with the full scatterplot matrix (Figure 1) one can regard the HE plot matrix as a “visual thinning” of the data, here focused on the predictive variation due to group mean differences relative to within-group variation. As well, the negative relations of species means on sepal width again stand out, compared with the strong positive relations for all other variables (cf. Figure 3).

Figure 5 is drawn using the hemat macro, using the same outstat= data set:

\[
%hemat(data=iris, 
stat=stats, 
var=SepalLen SepalWid PetalLen PetalWid, 
effect=species);
\]

2.5. Reduced-rank HE plots

Just as with the biplot, we can visualize the variation in group means (or any MLM effect) on all response variables in a single plot by projecting the data and variable vectors into a 2-dimensional subspace that captures most of the variation due to hypothesis relative to error. This amounts to transforming the observed responses to canonical discriminant scores \( z_1 \) and \( z_2 \), defined as the linear combinations of the \( y \) variables that maximize between-group (hypothesis) variance relative to within-group (error) variance.

Figure 6 illustrates this canonical discriminant HE plot for the iris data. In this plot, the order and separation of the group means on each canonical variable indicates how that linear combination of the responses discriminates among the groups. As an additional aid to interpretation we also draw vectors on the plot indicating the correlation of each of the observed variables with the canonical dimensions. For a given response, a vector is drawn from the origin (representing the grand mean on the canonical variates) to a point proportional to the correlation (canonical structure coefficients) of that variable with each canonical variate,
Figure 5: HE plot matrix for iris data. Each panel displays the \( H \) (solid, black) and \( E \) (dashed, red) bivariate ellipsoids.

Figure 6: Canonical HE plot for the iris data.
(r_{y_i z_1}, r_{y_i z_2}). With axes equated, the relative length of each variable vector will be proportional to its contribution to discriminating among the groups. As well, the angles between the variable vectors approximately indicate the correlations among group mean differences, based on the standardized $H$ matrix projected into the space of the canonical dimensions.

In this plot,

- the origin corresponds to the grand mean for all species and on all variables, with positive values representing positive deviations from the mean;
- thus, nearly all (99.1%) of the variation in species means is accounted for by a single canonical dimension, which corresponds to larger values for Virginica, and smaller for Setosa, on all variables except for sepal width.

Similar to the biplot example (Figure 3 in Section 2.3), we first use the canplot macro for its side-effect to calculate scores, can1 and can2 on the first two canonical dimensions.

```sas
%canplot(
  data=iris,
  class=species,
  var=SepalLen SepalWid PetalLen PetalWid,
  plot=NO,
  scale=3.5, /* scale factor for variable vectors */
  out=canscores, /* output data set containing discrim scores */
  anno=cananno); /* output data set containing annotations */
```

The plot in Figure 6 is then a standard HE plot applied to can1 and can2.

```sas
*-- Get H and E matrices for canonical scores;
proc glm data=canscores outstat=stats;
  class species;
  model can1 can2 = species / nouni ss3;
  manova h=species;
run;

*-- Axis statements to equate axis units;
axis1 length=2.6 IN order=(-4 to 4 by 2) label=(a=90);
axis2 length=6.5 IN order=(-10 to 10 by 2);
%heplot(data=canscores, stat=stats,
  x=Can1, y=Can2,
  effect=species,
  haxis=axis2, vaxis=axis1,
  legend=none, hsym=1.6,
  anno=cananno);
```

3. Details

Here we provide a brief summary of the statistical and computational methods on which these graphic displays are based.
3.1. Data ellipse

As seen from the examples, the data ellipse (Dempster 1969; Monette 1990) provides a visual summary of variables in a scatterplot indicating the means, standard deviations, correlation, and slope of the regression line for two variables. For two variables, \( Y_1, Y_2 \), the sample data ellipse \( \mathcal{E}_c \) of size \( c \) is defined as the set of points \( y = (y_1,y_2)' \) whose squared Mahalanobis distance, \( D^2(y) = (y - \bar{y})' S^{-1} (y - \bar{y}) \), from the means, \( \bar{y} \), is less than or equal to \( c^2 \),

\[
\mathcal{E}_c(y; S, \bar{y}) \equiv \{ y : (y - \bar{y})' S^{-1} (y - \bar{y}) \leq c^2 \} ,
\]

where \( S \) is the sample variance-covariance matrix, \( S = (n-1)^{-1} \sum_{i=1}^n (y_i - \bar{y})(y_i - \bar{y})' \).

When \( y \) is (at least approximately) bivariate normal, \( D^2(y) \) has a large-sample \( \chi^2 \) distribution (\( \chi^2 \) with 2 df), so taking \( c^2 = \chi^2(0.68) = 2.28 \) gives a “1 standard deviation bivariate ellipse,” an analog of the standard interval \( \bar{y} \pm 1s \), while \( c^2 = \chi^2(0.95) = 5.99 \approx 6 \) gives a data ellipse of 95% coverage. A bivariate ellipse of approximately 40% coverage has the property that its shadow on the \( y_1 \) or \( y_2 \) axes (or any linear combination of \( y_1 \) and \( y_2 \)) corresponds to a univariate \( \bar{y} \pm 1s \) interval. In small samples, the distribution of \( D^2(y) \) can be approximated more closely by \( [2(n-1)/(n-2)]F_{2,n-2} \approx 2F_{2,n-2} \); except in tiny samples \((n < 10)\), the difference is usually too small to be noticed in a graphical display.

The boundary of the data ellipse, \( \mathcal{E}_c \) (where equality holds in Eqn. (1)) may easily be computed as a transformation of a unit circle, \( U = (\sin \theta, \cos \theta) \) for \( \theta = 0 \) to \( 2\pi \) in radians. Let \( A = S^{1/2} \) be the Choleski square root of \( S \) in the sense that \( S = AA' \), whose columns form an orthonormal basis for the inner product \( (u, v) = uS^{-1}v \). Then \( \mathcal{E}_c = \bar{y} + cAU \) is an ellipse centered at the means, \( \bar{y} = (\bar{y}_1, \bar{y}_2) \), whose size reflects the standard deviations of \( y_1 \) and \( y_2 \) and whose shape (eccentricity) reflects their correlation. For bivariate normal data, the data ellipse is a level curve through the bivariate density.

All of the above results extend immediately to \( p \) variables, \( y_1, y_2, \ldots, y_p \), giving a \( p \)-dimensional \((1-\alpha) \) data ellipsoid \( \mathcal{E}_c \) with \( c^2 = \chi^2_p (1-\alpha) \) or \( c^2 = [p(n-1)/(n-p)]F_{p,n-p} (1-\alpha) \) in small samples.

3.2. Robust data ellipses

We recognize that a normal-theory summary (first and second moments), shown visually or numerically, can be distorted by multivariate outliers, particularly in smaller samples. Such effects can be countered by using robust covariance estimates such as multivariate trimming (Gnanadesikan and Kettenring 1972) or the high-breakdown bound Minimum Volume Ellipsoid (MVE) and Minimum Covariance Determinant (MCD) methods developed by Rousseeuw and others (Rousseeuw and Leroy 1987; Rousseeuw and Van Driessen 1999). In what follows, it should be noted that robust covariance estimates could, in principle, be substituted for the classical, normal-theory estimates in all cases.

To save space, we don’t illustrate these possibilities here. However, our outlier macro implements multivariate trimming, and the robcov macro implements MVE and MCD covariance estimation. Both return a modified dataset containing a \_WEIGHT_ variable, set to 0 for observations identified as potential outliers. Using this variable as the value of the WEIGHT= parameter in the ellipses macro will then give robust data ellipses.

3.3. Brief review of the multivariate LM
To establish notation and context for HE plots, we provide a capsule summary of the multivariate LM and the general linear test for any hypothesis. For details, see, e.g., Timm (1975) or Muller, LaVange, Ramey, and Ramey (1992).

When there are \( p \) responses, \((y_1, y_2, \ldots, y_p) = Y\), the multivariate LM

\[
Y_{(n \times p)} = X_{(n \times q)}B_{(q \times p)} + U_{(n \times p)}, \tag{2}
\]

with \( \text{vec}(U) \sim \mathcal{N}_p(0, I_n \otimes \Sigma) \), where \( \otimes \) is the Kronecker product, is a natural extension of the univariate version. Except for the fact that hypotheses are tested using multivariate tests, model Eqn. (2) is equivalent to the set of \( p \) models for each separate response, \( y_i = X_i \beta_i + \epsilon_i \) for \( i = 1, 2, \ldots, p \), where the columns of \( B = (\beta_1, \beta_2, \ldots, \beta_p) \) in Eqn. (2) are the model coefficients for the separate responses. As in the univariate case, the columns of the predictor matrix \( X \) may include any combination of: (a) quantitative regressors (age, income, education); (b) transformed regressors (\( \sqrt{\text{age}}, \log(\text{income}) \)); (c) polynomial regressors (\( \text{age}^2, \text{age}^3, \ldots \)); (d) categorical predictors and factors (treatment, sex—coded as “dummy” (0/1) variables or contrasts); (e) interaction regressors (treatment \( \times \) age, or sex \( \times \) age); (f) more general regressors (e.g., basis vectors for smoothing splines). In all cases, the least squares estimates of the coefficients, \( \hat{B} \) can be calculated as

\[
\hat{B} = (X'X)^{-1}X'Y, \tag{3}
\]

Then, just as all linear models can be cast in the form of the LM, all tests of hypotheses (for null effects) can be represented in the form of a general linear test,

\[
H_0 : C_{(h \times q)}B_{(q \times p)} = 0_{(h \times p)}, \tag{4}
\]

where \( C \) is a matrix of constants whose rows specify \( h \) linear combinations or contrasts of the parameters to be tested simultaneously by a multivariate test. (For repeated measures designs, an extended form of the general linear test, \( CBA = 0 \), where \( A \) is a \( p \times k \) matrix of constants, provides analogous contrasts or linear combinations of the responses to be tested. We don’t pursue this straightforward extension here.)

For any such hypothesis of the form Eqn. (3), the analogs of the univariate sums of squares for hypothesis (\( SS_H \)) and error (\( SS_E \)) are the \( p \times p \) sum of squares and crossproducts (SSCP) matrices (Timm 1975, Ch. 3, 5)

\[
H = (C\hat{B})' [C(X'X)^{-1}C']^{-1} (C\hat{B}), \tag{5}
\]

and

\[
E = Y'Y - \hat{B}'(X'X)\hat{B} = \hat{U}'\hat{U}. \tag{6}
\]

For example, in a one-way MANOVA, using the cell-means model for the vector of responses of subject \( j \) in group \( i \), \( y_{ij} = \mu_i + e_{ij} \), the \( H \) and \( E \) SSCP matrices for testing \( H_0 : \mu_1 = \mu_2 = \cdots \mu_g \) are

\[
H = \sum_{i=1}^{g} n_i (\bar{y}_i - \bar{y}_.) (\bar{y}_i - \bar{y}_.)', \tag{7}
\]

\[1 \text{ For simplicity, we don’t distinguish here among various parameterizations for factor variables (e.g., sum-to-zero constraints, first/last parameter =0, contrasts, etc.) that provide different unique solutions for parameter estimates, but which all yield identical overall tests for model effects.} \]
and

$$E = \sum_{i=1}^{g} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i) (y_{ij} - \bar{y}_i)' .$$

(7)

Then, the multivariate analog of the univariate $F$ statistic,

$$F = \frac{SS_H/df_h}{SS_E/df_e} = \frac{MS_H}{MS_E},$$

which implies $(MS_H - F MS_E) = 0$

is

$$\det(H - \lambda E) = 0 ,$$

(8)

where $\lambda$ represents the $s = \min(p, df_h)$ non-zero latent roots of the $H$ matrix relative to ("in the metric of") the $E$ matrix, or equivalently, the ordinary latent roots of the matrix $HE^{-1}$

$$\det(HE^{-1} - \lambda I) = 0 .$$

(9)

The ordered latent roots, $\lambda_1 \geq \lambda_2 \geq \cdots \lambda_s$, measure the "size" of $H$ relative to the "size" of $E$ in up to $s$ orthogonal directions, and are minimal sufficient statistics for all multivariate tests. These tests can also be expressed in terms of the eigenvalues $\theta_i = \lambda_i/(1+\lambda_i)$ of $HT^{-1}$, where $T = H + E$, and $\theta_i = \rho_i^2$ are the generalized squared canonical correlations. The various multivariate test statistics (Wilks’ $\Lambda$, Pillai’s trace criterion, Hotelling-Lawley trace criterion, Roy’s maximum root criterion) reflect different ways of combining this information across the dimensions, ranging from functions of their product (Wilks’ $\Lambda$), to functions of their sum (Pillai, Hotelling-Lawley), to their maximum (Roy).

Thus, in univariate problems ($p = 1$), or for 1 degree-of-freedom hypotheses ($df_h = 1$), there is $s = 1$ non-zero latent root, that has an exact relation to a univariate $F$. When $s > 1$, the number of “large” dimensions indicate how many different aspects of the responses are being explained by the hypothesized effect. These relations provide the motivation for HE plots.

From the description above, it is relatively easy to provide a visual explanation of the essential ideas behind all multivariate tests, particularly in the MANOVA context, as shown in Figure 7.

Figure 7(a) shows the individual-group data ellipses for two hypothetical variables, $Y_1, Y_2$. The variation due to differences in the group means is captured by the $H$ matrix, while the pooled within-sample variation is captured by the $E$ matrix, as illustrated in panel (b). The answer to the question, “How big is $H$ relative to $E$?” is shown geometrically in the last two panels.

The transformation from Eqn. (8) to Eqn. (9) can be represented (panel (c)) as a rotation of the variable space to the principal components of $E$, giving the matrix $E^*$. The same transformation applied to the $H$ matrix gives $H^*$. The axes in panels (c) and (d) turn out to be the canonical discriminant dimensions discussed in Section 3.7. In this space, the errors (residuals) are all uncorrelated, i.e., $E^*$ is diagonal, but with possibly different variances. Standardizing then transforms $E^* \mapsto I$ and $H^* \mapsto HE^{-1}$.

Because the transformed errors are now uncorrelated and standardized to unit variance, we can focus only on the ellipse for $HE^{-1}$ as shown in panel (d), where the latent roots, $\lambda_1, \lambda_2$ are the half-lengths of the major and minor axes.
Data Ellipses, HE Plots and Reduced-Rank Displays

Figure 7: Conceptual plots showing the essential ideas behind multivariate tests, in terms of the hypothesis ($H$) and error ($E$) matrices, for a 1-way MANOVA design with two response variables ($Y_1, Y_2$): (a) Bivariate means (points) and within-group variance-covariance matrices (ellipses); (b) The hypothesis ($H$) matrix reflects the variation of bivariate group means around the grand mean. The error ($E$) reflects the pooled within-group dispersion and covariation. (c) Standardizing: The $E$ matrix can be standardized, first to its principal components ($E^*$) and then normalized. The same transformations are applied to the $H$ matrix, giving $HE^{-1}$. (d) The ellipsoid of $HE^{-1}$ then shows the size and dimensionality of the variation in the group means in relation to a spherical error ellipsoid.

3.4. Varieties of HE plots

From Figure 7 and the preceding discussion it may be seen that there are several different ways to display the $H$ and $E$ matrices for a given effect in a multivariate test, as illustrated
in Figure 8.

Panel (a) shows the observations and the data ellipses for sepal length and petal length, as in panel (1:3) in Figure 1. The $H$ and $E$ matrices are shown in panel (b). The shape of the $H$ ellipse shows that the variation in the group means is essentially 1D, a dimension of overall (petal + sepal) length.

Alternatively, it is sometimes useful to plot the ellipses for $H + E$ and $E$ as shown in panel (c). This form is particularly useful for multi-way designs, so that each effect (e.g., $H_A$, $H_B$, $H_{AB}$) can be seen in relation to error ($E$) variation (see Figure 11). When the variation due to a given hypothesis is small relative to error—leading to acceptance of $H_0$—the $H$ and $E$ ellipses will nearly coincide. The lengths of the major/min or axes of $H + E$ are $1 + \lambda_i$, and Wilks’ $\Lambda = \prod_{i=1}^{s}(1 + \lambda_i)^{-1}$ is inversely proportional to the area (volume when $s > 2$) of the $H + E$ ellipse.

In these plots (and all those shown so far), $E$ in Eqn. (5) is scaled to a covariance matrix (giving $S_{\text{pooled}} = E/df_e$ for a MANOVA), so that it is on the same scale as the data
Data Ellipses, HE Plots and Reduced-Rank Displays

...and the same scaling is applied to \(H\) (or \(H + E\)), so we plot \(E_c(y; H/\text{df}_e, \bar{y})\) and \(E_c(y; E/\text{df}_e, \bar{y})\). This scaling also allows us to show the group means on the plot as an aid to interpretation, and the \(H\) matrix then reflects the effect size (similar to the square of Cohen’s (1977) \(d = (\bar{x}_1 - \bar{x}_2)/s_{\text{pooled}}\)) as well as its orientation and shape. We use the \(1/\text{df}_e\) scaling factors for \(H\) and \(E\) implicitly in the \texttt{heplot} macro, corresponding to the default \(\text{scale}=1\) for the \texttt{scale} parameter.

Finally, one may plot the ellipse for \(HE^{-1}\) (or the equivalent, symmetric matrix, \(H^* = E^{-1/2}HE^{-1/2}\)) in relation to the identity matrix, \(I\), representing uncorrelated errors of unit variance, as shown in panel (d). The Hotelling-Lawley trace statistic, \(\text{HLT} = \text{tr}(HE^{-1}) = \sum \lambda_i\), reflects the sum of lengths of the major and minor axes; the length of the major axis reflects Roy’s criterion, \(\theta_1 = \lambda_1/(1 + \lambda_1)\). The group means could be shown on such a plot (as in Figure 6) by calculating means on the transformed (canonical) variables.

There is also justification for considering \(H/\text{df}_h\) and \(E/\text{df}_e\) as an alternative and natural scaling, analogous to \(MS_H/MS_E\) in the univariate case, that would provide an indication of the strength of evidence against a null hypothesis \(CB = 0\) (Eqn. (3)). Figure 9 shows the same data as in Figure 8 and three scalings of \(H\) and \(E\), specified with the \texttt{scale=} parameter in the \texttt{heplot} macro. Panel (b) is the default (\texttt{scale}=1 1) seen so far. Panel (c) uses \texttt{scale=dfe/dfh 1} to give \(H/\text{df}_h\) and \(E/\text{df}_e\), keeping \(E/\text{df}_e\) while expanding \(H\), while panel (d) equivalently uses \texttt{scale=1 df/dfh}, keeping \(H/\text{df}_h\) and the means as in panel (b), while shrinking \(E\). However, because various multivariate tests focus on different aspects of the “size” of the matrices displayed, we cannot provide an unambiguous metric to indicate when \(H\) is sufficiently large to reject the null. Thus, all other plots shown here use the default (\texttt{scale}=1 1) scaling and standard 1 s.d. (68%) coverage, unless otherwise noted.

3.5. Contrasts

Just as in univariate ANOVA designs, important overall effects (\(df_h > 1\)) in MANOVA may be usefully explored and interpreted by the use of contrasts among the levels of the factors involved. In the general linear test Eqn. (3), contrasts are easily specified as one or more \((h_i \times q)\) \(C\) matrices, \(C_1, C_2, \ldots\), each of whose rows sum to zero.

As an important special case, for an overall effect with \(df_h\) degrees of freedom (and balanced sample sizes), a set of \(df_h\) pairwise orthogonal \((1 \times q)\) \(C\) matrices \((C_i'C_j = 0)\) gives rise to a set of \(df_h\) rank 1 \(H_i\) matrices that additively decompose the overall hypothesis SSCP matrix,

\[
H = H_1 + H_2 + \cdots + H_{df_h},
\]

exactly as the univariate \(SS_H\) may be decomposed in an ANOVA. Each of these rank 1 \(H_i\) matrices will plot as a vector in an HE plot, and their collection provides a visual summary of the overall test, as partitioned by these orthogonal contrasts.

To illustrate, we show in Figure 10 an HE plot for the sepal width and sepal length variables in the iris data, corresponding to panel (1:2) in Figure 1. Overlayed on this plot are the 1 df \(H\) matrices obtained from testing two orthogonal contrasts among the iris species: \textit{setosa} vs. the average of \textit{versicolor} and \textit{virginica} (labeled “S:VV”), and \textit{versicolor} vs. \textit{virginica} (“V:V”), for which the contrast matrices are

\[
C_1 = \begin{pmatrix} -2 & 1 & 1 \end{pmatrix} \\
C_2 = \begin{pmatrix} 0 & 1 & -1 \end{pmatrix}
\]
Figure 9: Data and HE plots for iris data, showing different scalings for the $H$ and $E$ matrices. (a) data ellipses; (b) $H/dfe$ and $E/dfe$; (c) $H/dfh$ and $E/dfe$; (d) $H/dfe$ and $(df_h/dfe)E/dfe$.

where the species (columns) are taken in alphabetical order.

These contrasts are tested with PROC GLM as shown below, using CONTRAST statements to specify the the contrast weights:

```sas
proc glm data=iris outstat=stats;  
class species;  
model SepalLen sepalwid PetalLen petalwid = species / nouni ss3;  
contrast 'S:VV' species -2 1 1;  
contrast 'V:V' species 0 -1 1;  
manova H=species /short summary;  
run;
```

This HE plot shows that, for the two sepal variables, the greatest between-species variation is accounted for by the contrast between *setosa* and the others, for which the effect is very large in relation to error (co-)variation. The second contrast, between the *versicolor* and *virginica* species is relatively smaller, but still explains some variation of the sepal variables among the
Figure 10: $H$ and $E$ matrices for sepal width and sepal length in the iris data, together with $H$ matrices for testing two orthogonal contrasts in the species effect.

The general method described above applies more widely than we have illustrated. Multiple-df tests may be expressed in terms of $C$ matrices with $h_i > 1$ rows. In a bivariate HE plot, their $H$ matrices will appear as ellipses for these contrasts contained within the H ellipse for the overall test.

The `heplot` macro was initially designed to plot $H$ and $E$ matrices for just a single effect. To show them all together, we produce three plots (with display suppressed), then overlay them in a single plot with the `panels` macro. To make this work, the axes must be scaled identically in all plots, which is done with the `AXIS` statements and the `VAXIS=` and `HAXIS=` parameters.

```sas
axis1 label=(a=90) order=(45 to 70 by 5);
axis2 order=(24 to 38 by 2);
legend1 position=(bottom center inside) offset=(0,1) mode=share frame;
goptions nodisplay;
%heplot(data=iris,stat=stats, var= SepalWid SepalLen, effect=species, vaxis=axis1, haxis=axis2, legend=legend1);
*-- Contrasts;
%heplot(data=iris,stat=stats, var= SepalWid SepalLen, effect=Species, vaxis=axis1, haxis=axis2, legend=legend1);
```
3.6. MMRA

Multivariate multiple regression is just another special case of the MLM, where all columns in the $X$ matrix are quantitative. For MMRA, the overall test, $B = 0$, of no linear relation between the $X$ and $Y$ variables collectively corresponds to $C = I$ in Eqn. (4) and the $(p \times p)$ $H$ matrix becomes

$$H = \hat{\beta}'(X'X)\hat{\beta} = \hat{Y}'\hat{Y},$$

where $H$ is of rank $s = \min(p, q)$ and therefore has $s$ non-zero latent roots. (For simplicity, we assume that all response variables are expressed in terms of deviations from their means, so all intercepts are zero and can be ignored here.)

For any two responses, the overall test can be shown as an HE plot as we have shown before. The $H$ ellipse is simply the data ellipse of the fitted values $\hat{Y}$; the $E$ ellipse is the data ellipse of the residuals, $U = Y - \hat{Y}$ (shifted to the centroid). For an individual regressor, the test of $H_0 : \beta_i = 0$ for the $i$th row of $B$ also gives rise to a $(p \times p)$ $H$ matrix, obtained using the $1 \times q$ matrix $C = (0, 0, \ldots, 1, 0, \ldots, 0)$, with a 1 in the $i$th position. In this case $H_i = \beta_i'(X'X)\hat{\beta}_i$, is a matrix of rank 1, with one non-zero latent root, so the ellipse for the $H$ matrix degenerates to a line.

Unfortunately, given the model

```r
proc glm outstat=stats;
  model y1 y2 y3 = x1-x5;
```

PROC GLM will produce the five $3 \times 3, 1$ df $H$ matrices for the separate predictors, but does not produce a $H$ matrix for the overall test, $B = 0$. The overall $H$ matrix is produced by PROC REG, though in a slightly different format. The `hemreg` macro uses PROC REG to extract the overall $H$ and $E$ matrices, and massages them into the form expected by the `heplot` macro. Examples are described in Section 4.3.

3.7. Canonical discriminant plots

Canonical discriminant analysis (CDA), used in our reduced-rank HE plots, can be regarded as an extension of MANOVA, where the emphasis is on dimension-reduction.

Formally, for a one-way design with $g$ groups and $p$-variate observations $y_{ij}$, CDA finds a set of $s = \min(p, g - 1)$ linear combinations, $z_1 = c_1'y$, $z_2 = c_2'y$, $\ldots$, $z_s = c_s'y$, so that: (a) all $z_k$ are mutually uncorrelated; (b) the vector of weights $c_1$ maximizes the univariate $F$-statistic for the linear combination $z_1$; (c) each successive vector of weights $c_k, k = 2, \ldots, s$
maximizes the univariate $F$-statistic for $z_k$, subject to being uncorrelated with all other linear combinations.

The canonical weights, $c_i$ are simply the eigenvectors of $HE^{-1}$ associated with the ordered eigenvalues, $\lambda_i, i = 1, \ldots, s$, and a MANOVA of all $s$ linear combinations is statistically equivalent to that of the raw data. The $\lambda_i$ are proportional to the fractions of between-group variation expressed by these linear combinations. Hence, to the extent that the first one or two eigenvalues are relatively large, a two-dimensional display will capture the bulk of between group differences. The canonical discriminant HE plot is then simply an HE plot of the scores $z_1$ and $z_2$ on the first two canonical dimensions.

Because the $z$ scores are all uncorrelated, the $H$ and $E$ matrices will always have their axes aligned with the canonical dimensions; when, as here, the $z$ scores are standardized, the $E$ ellipse will be circular, assuming that the axes are equated so that a unit data length has the same physical length on both axes, as in Figure 6. The example in Section 4.2 illustrates how these methods can be extended to two-way designs.

4. Further examples

4.1. MANOVA Examples

Sex, Drugs and Weight Loss

For two-way and higher-order MANOVA designs, HE plots provide a compact, visual summary of the multivariate tests for various main effects and interactions. To illustrate, Figure 11 uses a text-book example (Morrison 1990, p. 217, Table 5.5) dealing with possible toxic effects of three drugs (A, B, C) on rats, also classified by sex (M, F), where the responses are weight losses on two consecutive weeks (Week1, Week2), treated here as a two-way MANOVA design.

From the data ellipses (Figure 11 (a)) it is apparent that groups given drug C differ substantially from the remaining groups, which don’t appear to differ among themselves, with the possible exception of group M:A. These are very small samples ($n = 4$ per cell); with larger samples, the assumption of equal within-group covariance matrices might be questioned. The HE plots (Figure 11 (b)–(d)) show that differences among drugs are quite large; the main effect of sex is inconsequential, and any hint of a sex*drug interaction is confined to the larger and opposite sex difference with drug C than the other two. Note that for a one degree-of-freedom test ($s = 1$), such as sex in this example, the $H$ ellipse degenerates to a line, a result we exploit below to show separate effects in a single plot.

These plots are produced in a way similar to previous examples (e.g., Figure 8). The code is contained in the file heplot2.sas in the accompanying archive.

Captive and Maltreated Bees

A graduate student (Noli Pabalan) in biology studied the effects of captivity and maltreatment on reproductive capabilities of queen and worker bees in a complex factorial design (Pabalan, Davey, and Packe 2000). Bees were placed in a small tube and either held captive or shaken periodically for one of 5, 7.5, 10, 12.5 or 15 minutes, after which they were sacrificed and two measures, ovarian development and ovarian reabsorption, were taken. A single control
group was measured with no such treatment, i.e., at time 0, n = 10 per group. The design is thus nearly a three-way factorial, with factors Caste (Queen, Worker), Treatment (Cap, Mal) and Time, except that there are only 11 combinations of Treatment and Time; we call these TrtTime below.

To analyze this data, we treat the design as a two-way factorial, Caste (2) × TrtTime (11), and use contrasts to resolve the 10 df for TrtTime into questions of interest. For example, tests of the control condition vs. the average of the Captive groups and vs. the average of the Captive groups are shown below, along with a test of the Treatment effect (Cap vs. Mal). Because Time is quantitative (and expected to have mainly linear effects), we also use orthogonal polynomial contrasts to test for linear and non-linear effects of time, within each of Cap and Mal treatment groups. Other contrasts (not shown here) are used to resolve the interaction of Caste with Treatment and Time into constituent components.

```sas
title 'Two-way factorial with appended control group';
title2 'Ovarian development in captive and maltreated bees';
data bees;
```
input caste $ treat $ time Iz Iy;
length trtime $8;
   label Iy='Index of ovarian development'
   Iz='Index of ovarian reabsorption';
   *-- Since treat is missing @ time zero, construct
   a new variable with 11 levels;
   if time = 0 then trtime = '0 3';
   else trtime = trim(treat) || put(int(time),z2.);

datalines;
Queen . 0 1.33 1.50
Queen . 0 1.50 0.00
Queen . 0 1.83 0.00
Queen . 0 1.67 0.00
Queen . 0 0.67 0.83
Queen . 0 1.83 0.00
... Worker MAL 15 0.17 0.83
Worker MAL 15 0.00 1.33
Worker MAL 15 0.17 0.83
run;
proc glm data=bees outstat=HEstats noprint;
class caste trtime;
model Iz Iy = caste|trtime /ss3;
   * Treatment: Captive Maltreated ;
   * Time: 0 5 7 10 12 15 5 7 10 12 15;
   contrast '0 vs CAP' trtime 5 -1 -1 -1 -1 -1 0 0 0 0 0;
   contrast '0 vs MAL' trtime 5 0 0 0 0 0 -1 -1 -1 -1 -1;
   contrast 'treat' trtime 0 1 1 1 1 1 -1 -1 -1 -1 -1;
   *-- Contrasts for time, within each treatment;
   contrast 'CAP t:lin' trtime 0 -2 -1 0 1 2 ;
   *-- Test all non-linear terms for captive;
   contrast 'CAP t:2-4 ' trtime 0 2 -1 2 -1 2 ,
   trtime 0 -1 2 0 -2 1 ,
   trtime 0 1 -4 6 -4 1 ;
   ...
   manova h = caste|trtime;
run;

As a result of the above PROC GLM step, the outstat=HEstats data set contains \( H \) matrices for the factorial of caste|trtime as well as for all contrasts, plus, of course, the \( E \) matrix. A portion of this is shown below. (\( F \) and PROB values are those for the univariate tests.)

<table>
<thead>
<tr>
<th></th>
<th><em>SOURCE</em></th>
<th><em>TYPE</em></th>
<th><em>NAME</em></th>
<th>Iz</th>
<th>DF</th>
<th>SS</th>
<th>F</th>
<th>PROB</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ERROR</td>
<td>ERROR</td>
<td>Iz</td>
<td>22.8031 -14.7162 224</td>
<td>22.8031</td>
<td>.</td>
<td>.</td>
<td></td>
</tr>
<tr>
<td>caste</td>
<td>SS3</td>
<td>Iy</td>
<td>0.2540</td>
<td>2.2696</td>
<td>1</td>
<td>0.2540</td>
<td>2.495</td>
<td>0.11558</td>
</tr>
<tr>
<td>caste</td>
<td>SS3</td>
<td>Iz</td>
<td>54.3386</td>
<td>-50.5192</td>
<td>10</td>
<td>54.3386</td>
<td>53.378</td>
<td>0.00000</td>
</tr>
<tr>
<td>trtime</td>
<td>SS3</td>
<td>Iy</td>
<td>-50.5192</td>
<td>47.6545</td>
<td>10</td>
<td>47.6545</td>
<td>51.841</td>
<td>0.00000</td>
</tr>
</tbody>
</table>
The superimposed HE plots for some of these effects (Treatment, Time and Caste) is shown in Figure 12. This shows that the overall effect of Time for all treated groups is such that ovarian development increases, while ovarian reabsorption decreases over time, and the effect appears largely linear. Maltreatment as opposed to mere captivity increases these effects in an approximately additive fashion.

This plot is produced as shown below. Again, note the use of AXIS statements to ensure that the plots are scaled and labeled identically.

```sas
%heplot(data=bees,stat=HEstats, var= Iz Iy, effect=time, ss=contrast, 
        efflab=Time, vaxis=axis1, haxis=axis2, legend=legend1);

%heplot(data=bees,stat=HEstats, var= Iz Iy, effect=treat, ss=contrast, 
        efflab=Treat, vaxis=axis1, haxis=axis2, legend=none);

%heplot(data=bees,stat=HEstats, var= Iz Iy, effect=caste, ss=ss3, 
        efflab=Caste, vaxis=axis1, haxis=axis2, legend=none);

%displa(ON);
%panels(rows=1, cols=1, replay=1:1 1:2 1:3);
```
4.2. CDA Examples

Soil composition

This example illustrates the use of canonical discriminant HE plots for a two-way design, in a situation where there are more response variables than can easily be viewed in bivariate HE plot or in HE plot matrices.

Horton, Russell and Moore 1968 considered the problem of discriminating among populations of gilgaied soil types based on physical and chemical characteristics, for the purpose of being able to identify the important variables that could be used to classify new samples. These data are discussed in Khattree and Naik (2000).

Microtopographic areas, categorized as Top, Slope and Depression were sampled at four different depth layers (0–10 cm, 10–30, 30–60, 69–90). The area was divided into four Blocks and four soils samples were taken for each of these 12 populations, and nine variables were measured for each for each. These are pH value (pH), total nitrogen in % (N), bulk density in gm/cm$^3$ (Dens), total phosphorous in ppm (P), calcium (Ca), magnesium (Mg), potassium

\begin{footnote}{The surface topography of gilgaied soils resembles a battlefield covered by bomb craters.}\end{footnote}
(K), sodium(Na, the last four in me/100 gms) and conductivity (Conduc).

With nine response variables and 12 groups in a $3 \times 4$ randomized block design there is too much data to be understood comprehensively in a few displays using plots of means or even HE plot matrices. For example, Figure 13 shows the HE plot matrix for just six of the nine response variables (as many as we could fit legibly), and only for the effect of depth.

![HE plot matrix for the Depth effect in the soils data](image)

Figure 13: HE plot matrix for the Depth effect in the soils data, showing six of the nine responses

Canonical discriminant analysis is almost always applied to one-way designs, and most software allows only a single classification factor. Yet, from the MLM it is not hard to see how the method can be extended to two-way and higher designs. This can be done in several ways. First, one may simply code the combinations of all factors interactively, so that $H$ expresses all group differences, e.g., $H = H_A + H_B + H_{AB}$ and $E = E_w$ (within-cell error) in a two-way design. The result, using the `canplot` macro, is shown in Figure 14. The two canonical dimensions account for 92.6% of between group variation.

This figure shows that the first dimension is largely reflecting soil depth, with smaller depth associated with higher values for phosphorous, pH, calcium and potassium, and lower values...
on the remaining variables. The second dimension distinguishes the three ellipses types.

It is also possible to study effects individually, ignoring other factors, whose effects get pooled with error. For a two-way design, this would correspond to $H = H_A$ and $E = H_B + H_{AB} + E_w$ in an HE plot of the canonical scores. Figure 15 shows the canonical discriminant plot for Depth, ignoring Contour. The result indicates that the effect of Depth is largely linear in the depth value. The second dimension, accounting for only 5% of between-group variation is associated with deviation from linearity.

These plots (Figure 14 and Figure 15) are produced using the `ellipses` macro as shown below. They differ mainly in the variable specified for the `class=` parameter. (The `caption` macro, not shown, is a simple utility to add a plot title inside the plot frame.)

```sas
legend1 position=(inside bottom left) mode=protect;
%caption(All groups, x=50, htext=2, out=_title_);
%canplot(data=soils,
   class=Gp,
   var=pH--Conduc,
   inc=2 2, yextra=1 0, xextra=0 1, scale=5,
   annoadd=mean gplabel _title_,
   colors=red blue green brown,
   legend=legend1
);

%caption(Soil depth, x=50, htext=2, out=_title_);
```
Second, the method may be applied to adjusted response variate vectors, which are essentially the residuals (deviations from the means) of the model effects adjusted for. In the two-way setting, for example, the reduced-rank HE plot for the AB effect, $H_{AB}$, is equivalent to the analysis of $y | (A, B)$, i.e., $y_{ijk} - \bar{y}_A(i) - \bar{y}_B(j)$.

For example, the PROC GLM step below fits the main-effect model and produces residuals to an output data set named Resids via the output statement. Because variables of the same names exist on the input data set, SAS appends 2 to each variable name.

```sas
%caption(Depth*Contour Effect, htext=2, out=_title_);
%canplot(data=Resids, class=Gp, var=pH2--Conduc2, inc=2 2, scale=5, conf=.50, colors=red blue green brown, annoadd=mean _title_, legend=legend1, yextra=1 0);
```

---

Figure 15: Canonical plot for soils data, showing the main effect of soil depth, ignoring ellipses
out=canscores,
anno=cananno);

Figure 16: Canonical HE plot for the interaction of Contour × Depth

This plot (not shown) is too cluttered to be of much use. Instead, we use the canonical scores (out=canscores) to produce a canonical HE plot version in a similar way to that shown earlier for the iris data (Section 2.5). The result, shown in Figure 16, summarizes all interaction variation of the means in the $H$ ellipse and simply plots the means as points. In the MANOVA analysis, the Depth × Contour interaction is significant ($p = 0.01$) by Roy’s greatest root test ($\lambda_1$) but not by any of the other tests. It is not clear if there is any interpretation of the pattern of interaction residuals in relation to the response variables.

The following statements are used to produce Figure 16 from the canonical scores and annotations generated above:

```sas
*** Remove conf. circles;
data cananno;
set cananno;
where comment not in ('CIRCLE');
```
4.3. MMRA Examples

Cognitive ability and paired-associate learning

We illustrate the use of these methods for MMRA data with a study by William Rower used as a textbook example (Timm 1975, Table 4.7.1). In this study, \( n = 37 \) kindergarten children of low socio-economic status (SES) were first assessed on standard measures of “cognitive skills and ability” using (a) the Peabody Picture Vocabulary Test (PPVT), a non-verbal measure of receptive vocabulary and verbal ability; (b) a student achievement test (SAT), unspecified; (c) the Raven Progressive matrices test (RAVEN), a non-verbal, “culture-fair” intelligence test thought by some to tap a latent dimension of general intelligence.

In the study, Rower also administered a set of five paired-associate (PA) learning tasks, where stimulus:response pairs (Toronto:YYZ, Los Angeles:LAX; or YYZ:\( \Box \), LAX:\( \triangle \), etc.) are first presented for study and then the stimuli are presented alone for testing (Toronto:?, LAX:?), with the subject required to indicate the appropriate response. The PA tasks varied in how the stimuli were presented (not described) and are called named (N), still (S), named still (NS), named action (NA), and sentence still (SS).

An interesting feature of this data is that separate, univariate multiple regressions carried out for each response variable, testing \( H_0 : \beta_i = 0 \) for the \( i \)th row of \( B \), show that the SAT and RAVEN fail significance on an overall test for the \( q = 5 \) predictors. For the PPVT, the overall univariate test is significant \( (F(5, 31) = 6.47, R^2 = 0.510) \), but among the partial tests for individual predictors, only one (NA) attains significance. From these results, one might conclude that PA tasks are at best marginally related to the intellectual and achievement tests. However, the overall multivariate test, \( B = 0 \), is highly significant.

The HE plot for SAT and PPVT in Figure 17 helps to understand these results. It may be seen that although the error covariance for these variables is nearly circular, the H matrix structure is more highly concentrated, with generally positive correlations among the predictors, for two subsets, (NA, S) and (NS, SS) that appear to have different relations to the responses. This
allows the multivariate tests to “pool strength” across predictors, resulting in greater power for overall test.

In this example, we first use the `hemreg` macro to obtain the $H$ and $E$ matrices for the overall test, and `PROC GLM`, as usual, to obtain the $H$ matrices for the separate predictors. We specify `SS3` (via a macro variable) to request partial (type III) sums of squares and crossproducts, corresponding to the usual tests in MRA.

```sas
data lo_ses (label='Timm Example 4.7');
  input SAT PPVT RAVEN N S NS NA SS;
  label sat='Student achievement test'
    ppvt='Peabody picture vocabulary test'
    raven='Raven progressive matrices test'
    n='Named P.A. learning test'
    s='Still P.A. learning test'
    ns='Named Still P.A. learning test'
    na='Named Action P.A. learning'
    ss='Sentence Still P.A. learning';
cards;
  49 48 8 1 2 6 12 16
  47 76 13 5 14 14 30 27
  11 40 13 0 10 21 16 16
  9 52 9 0 2 5 17 8
```

Figure 17: HE plot for MMRA, showing the H ellipse for an overall test, $B = 0$, and the H ellipses for individual predictors, using type III (partial) sums of squares.
The plot in Figure 17 is then produced by overlaying the result of six calls to `heplot` with the `panels` macro:

```sas
*-- Plot separate predictors;
%heplot(stat=HE_Xs, data=lo_ses, var=sat ppvt raven, effect=N, ss=&ss, efflab=N, class=%str( ));
%heplot(stat=HE_Xs, data=lo_ses, var=sat ppvt raven, effect=S, ss=&ss, legend=none, efflab=S, class=%str( ));
%heplot(stat=HE_Xs, data=lo_ses, var=sat ppvt raven, effect=NS, ss=&ss, legend=none, efflab=NS, class=%str( ));
%heplot(stat=HE_Xs, data=lo_ses, var=sat ppvt raven, effect=NA, ss=&ss, legend=none, efflab=NA, class=%str( ));
%heplot(stat=HE_Xs, data=lo_ses, var=sat ppvt raven, effect=SS, ss=&ss, legend=none, efflab=SS, class=%str( ));
goptions display;
```

Re-running this program with the macro statement `%let ss=ss1;` generates a HE plot using
Type I (sequential) sum of squares $H$ matrices for the individual predictors. These have the nice property that they add to the $H$ matrix for the overall test; however, they depend on the order that the variables are listed in the `model` statement and usually require some justification for testing effects in a hierarchical fashion.

*Cognitive ability and paired-associate learning by SES*

Rohwer’s study, described above, was also carried out with 32 high SES children from an upper-class, white residential school. By joining the data sets, one can test a variety of hypotheses about the multivariate regression relations for the two groups: Are the regression relations coincident for the two groups (equal slopes and intercepts)? If not, are they parallel? If they are parallel, it also makes sense to test whether intercepts are equal, which corresponds to a test of equal response means.

Figure 18: HE plots the High and Low SES groups showing the H ellipse for an overall test, $B = 0$, and the H ellipses for individual predictors, using type III (partial) sums of squares, separately for each group.

We don’t fully follow this testing approach here. For the present purposes it is sufficient to
illustrate what may be seen from a comparison of HE plots for the two groups, shown in Figure 18. In regression terms, this corresponds to fitting a full model that allows different slopes and intercepts for all measures in the two groups. It may be seen that there is a large difference in the means of the Low and High SES groups on these two response measures. As well, the predictive relations of the paired associate tests to the responses appear to differ somewhat for the two groups, with the PA measures more strongly related to the SAT in the High SES group than in the Low group.

5. Discussion

In this paper we have described and illustrated a variety of graphical displays for multivariate LMs, designed to focus on the relationships between two sets of variables: predictors (regressors) and responses. Some of these methods are new (HE plots), some are old (biplots), and some have been extended here to a wider context (data ellipse). There are several general themes, statistical ideas, and graphical notions that connect the cases we have described here. First, the data ellipse, as used here, provides a visual summary of bivariate relations, depicting means, variances, and covariances (or correlations), for either the classical, normal-theory estimators, or any robust estimator. These provide useful exploratory and confirmatory displays in a variety of multivariate contexts, can be used to show multiple-group MANOVA data, and can be embedded in a scatterplot matrix form to show all pairwise, bivariate relations.

Second, the idea of HE plots provides ways to visualize and understand the results of multivariate tests in both the MANOVA and MMRA contexts. Group means (for MANOVA) or 1-df H matrix vectors (for MMRA) can be overlayed on these plots to aid interpretation, and the pairwise relations for all responses can be seen in the IIE plot matrix.

Third, we have used several dimension-reduction techniques (biplot, canonical discriminant analysis) to display two-dimensional summaries of the salient characteristics of multivariate data related to various aspects of the MLM. Overlaying variable vectors, data ellipses, and reduced-rank scores for observations, helps to make these plots more interpretable in relation to both the original data and the low-dimensional summary.

The collection of SAS macros we have developed makes these methods accessible and easily used for a wide range of data problems.

A. Software

As mentioned above, the SAS macros used here are available with documentation and examples at http://www.math.yorku.ca/SCS/sasmac/. The current versions as of this writing are contained in the accompanying archive, together with SAS code for the named examples in this paper. A README file gives installation and usage instructions.

For those who are not primarily SAS users, or who might wish to translate these methods to other software, the following programs are documented in this appendix:

- canplot: Canonical discriminant structure plots
- heplot: Plot H and E matrices for a bivariate MANOVA effect
- hemat: HE plots for all pairs of response variables
hemreg  Extract H and E matrices for multivariate regression

For R users, the online documentation for the heplot macro contains a link to heplot.R, which contains a rudimentary function ellipse.manova() for drawing HE plots from an mlm object. A more mature R implementation is in progress.

A.1. The CANPLOT macro: Canonical discriminant structure plot

The CANPLOT macro constructs a canonical discriminant structure plot. The plot shows class means on the two largest canonical variables, confidence circles for those means, and variable vectors showing the correlations of variables with the canonical variates.

Method

Discriminant scores and coefficients are extracted from PROC CANDISC and plotted. Other designs may be handled either by (a) coding factor combinations 'interactively', so, e.g., the combinations of A*B are represented by a GROUP variable, or (b) by applying the method to adjusted response vectors (residuals) with some other predictor (class or continuous) partialled out. The latter method is equivalent to analysis of the residuals from an initial PROC GLM step, with the effects to be controlled or adjusted for as predictors.

E.g., to examine Treatment, controlling for Block and Sex,

```plaintext
proc glm data=..;
  model Y1-Y5 = block sex;
  output out=resids r=E1-E5;
  %canplot(data=resids, var=E1-E5, class=Treat, ...);
```

Usage

The CANPLOT macro is defined with keyword parameters. Values must be supplied for the CLASS= and VAR= parameters. The arguments may be listed within parentheses in any order, separated by commas. For example:

```
%canplot(data=inputdataset, var=predictors, class=groupvariable...,);
```

The interpretation of the angles between variable vectors relies on the units for the horizontal and vertical axes being made equal (so that 1 data unit measures the same length on both axes. The axes should be equated either by using the GOPTIONS HSIZE= VSIZE= options, or using the macro HAXIS= and VAXIS= parameters and AXIS statements which specify the LENGTH= value for both axes. The current version now uses the EQUATE macro if the HAXIS= and VAXIS= arguments are not supplied.

Parameters

DATA=  Name of data set to analyze. [Default: DATA=_LAST_]
CLASS= Name of one class variable, defining the groups to be discriminated.

VAR= List of classification variables

ID= Observation ID variable, used to label observations in the plot.

VARLAB= How to label variables? _NAME_ or _LABEL_. [Default: VARLAB=_NAME_]

DIM= Number of canonical dimensions to be extracted. [Default: DIM=2]

SCALE= Scale factor for variable vectors in plot. The variable vectors are multiplied by the SCALE= value, which should be specified (perhaps by trial and error) to make the vectors and observations fill the same plot region. [Default: SCALE=4]

CONF= Confidence probability for canonical means, determining the ra. [Default: CONF=.99]

OUT= Output data set containing discrim scores. [Default: OUT=_DSCORE_]

OUTVAR= Output data set containing coefficients. [Default: OUTVAR=_COEF_]

ANNO= Output data set containing annotations. [Default: ANNO=_DANNO_]

ANNOADD= Additional annotations to add to the plot. Can include 'MEAN' and/or 'GPLABEL' and/or the name(s) of additional input annotate data sets. [Default: ANNOADD=MEAN]

PLOT= YES (or NO to suppress plot) [Default: PLOT=YES]

HAXIS= The name of an optional AXIS statement for the horizontal axis. The HAXIS= and VAXIS= arguments may be used to equate the axes in the plot so that the units are the same on the horizontal and vertical axes. If neither HAXIS= nor VAXIS= are supplied, the EQUATE macro is called to generate axis statements.

VAXIS= The name of an optional AXIS statement for the vertical axis.

INC= X, Y axis tick increments, in data units. [Default: INC=1 1]

XEXTRA= # of extra X axis tick marks on the left and right. Use this to extend the axis range. [Default: XEXTRA=0 0]

YEXTRA= # of extra Y axis tick marks on the bottom and top. [Default: YEXTRA=0 0]

LEGEND= Name of a LEGEND statement to specify legend for groups. Use LEGEND=NONE to suppress the legend (perhaps with ANNOADD=GPLABEL to plot group labels near the means).

HSYM= Height of plot symbols. [Default: HSYM=1.2]

HID= Height of ID labels. [Default: HID=1.4]

IDCOLOR= Color of ID labels

HTEXT= Height of variable and group labels. [Default: HTEXT=1.5]

CANX= Horizontal axis of plot. [Default: CANX=CAN1]

CANY= Vertical axis of plot. [Default: CANY=CAN2]

DIMLAB= Dimension label prefix. [Default: DIMLAB=Canonical Dimension]

COLORS= List of colors to be used for groups (levels of the CLASS= variable). The values listed are recycled as needed for the number of groups. [Default: COLORS=RED GREEN BLUE BLACK PURPLE BROWN ORANGE YELLOW]

SYMBOLS= List of symbols to be used for the observations within the groups, recycled as needed. [Default: SYMBOLS=dot circle triangle square star - : \$ =]

LINES= List of line style numbers used for the confidence circles. [Default: LINES=20 20 20 20 20 20 20]

NAME= Name for graphic catalog entry. [Default: NAME=CANPLOT]

GOUT= The name of the graphics catalog. [Default: GOUT=GSEG]
A.2. The **HEPLOT** macro: Plot hypothesis and error matrices for a bivariate MANOVA effect

The **HEPLOT** macro plots the covariance ellipses for a hypothesized (H) effect and for error (E) for two variables from a MANOVA. The plot helps to show how the means of the groups differ on the two variables jointly, in relation to the within-group variation. The test statistics for any MANOVA are essentially saying how 'large' the variation in H is, relative to the variation in E, and in how many dimensions. The **HEPLOT** shows a two-dimensional visualization of the answer to this question. An alternative two-dimensional view is provided by the **CANPLOT** macro, which shows the data, variables, and within-group ellipses projected into the space of the largest two canonical variables—linear combinations of the responses for which the group differences are largest.

Typically, you perform a MANOVA analysis with **PROC GLM**, and save the output statistics, including the H and E matrices, using the **OUTSTAT=** option. This must be supplied to the macro as the value of the **STAT=** parameter. If you also supply the raw data for the analysis via the **DATA=** parameter, the means for the levels of the **EFFECT=** parameter are also shown on the plot.

Various kinds of plots are possible, determined by the **M1=** and **M2=** parameters. The default is **M1=H** and **M2=E**. If you specify **M2=I** (identity matrix), then the H and E matrices are transformed to $H^* = eHe$ (where $e = E^{-1/2}$), and $E^* = eEe = I$, so the errors become uncorrelated, and the size of $H^*$ can be judged more simply in relation to a circular $E^* = I$.

For multi-factor designs, it is sometimes useful to specify **M1=H+E**, so that each factor can be examined in relation to the within-cell variation.

**Usage**

The **HEPLOT** macro is defined with keyword parameters. The **STATS=** parameter and either the **VAR=** or the **X=** and **Y=** parameters are required. You must also specify the **EFFECT=** parameter, indicating the H matrix to be extracted from the **STATS=** data set. The arguments may be listed within parentheses in any order, separated by commas. For example:

```plaintext
proc glm data=dataset outstat=HEstats;
   model y1 y2 = A B A*B / ss3;
   manova;
   %heplot(data=dataset, stat=HEstats, var=y1 y2, effect=A );
   %heplot(data=dataset, stat=HEstats, var=y1 y2, effect=A*B );
```

**Parameters**

- **STAT=**: Name of the **OUTSTAT=** dataset from **PROC GLM** containing the SSCP matrices for model effects and ERROR, as indicated by the _SOURCE_ variable.
- **DATA=**: Name of the input, raw data dataset (for means)
- **X=**: Name of horizontal variable for the plot
- **Y=**: Name of vertical variable for the plot
- **VAR=**: 2 response variable names: x y. Instead of specifying **X=** and **Y=** separately, you can specify the names of two response variables with the **VAR=** parameter.
EFFECT=  Name of the MODEL effect to be displayed for the H matrix. This must be one of the terms on the right hand side of the MODEL statement used in the PROC GLM or PROC REG step, in the same format that this effect is labeled in the STAT= dataset. This must be one of the values of the _SOURCE_ variable contained in the STAT= dataset.

CLASS=  Names of class variables(s), used to find the means for groups to be displayed in the plot. The default value is the value specified for EFFECT=, except that '*' characters are changed to spaces. Set CLASS=(null) for a quantitative regressor or to suppress plotting the means.

EFFLAB=  Optional label (up to 16 characters) for the H effect, annotated near the upper corner of the H ellipse

MPLOT=  Matrices to plot. MPLOT=1 plots only the H ellipse. [Default: MPLOT=1 2]

GPFMT=  The name of a SAS format for levels of the group/effect variable used in labeling group means.

ALPHA=  Non-coverage proportion for the ellipses. [Default: ALPHA=0.32]

PVALUE=  Coverage proportion, 1-alpha. [Default: PVALUE=0.68]

SS=  Type of SS to extract from the STAT= dataset. The possibilities are SS1-SS4, or CONTRAST (but the SSn option on the MODEL statement in PROC GLM will limit the types of SSCP matrices produced). This is the value of the _TYPE_ variable in the STAT= dataset. [Default: SS=SS3]

WHERE=  To subset both the STAT= and DATA= datasets

ANNO=  Name of an input annotate data set, used to add additional information to the plot of y * x.

ADD=  Specify ADD=CANVEC to add canonical vectors to the plot. The PROC GLM step must have included the option CANONICAL on the MANOVA statement.

M1=  First matrix: either H or H+E. [Default: M1=H]

M2=  Second matrix either E or I. [Default: M2=E]

SCALE=  Scale factors for M1 and M2. This can be a pair of numeric values or expressions using any of the scalar values calculated in the PROC IML step. The default scaling [SCALE=1 1] results in a plot of E/dfe and H/dfe, where the size and orientation of E shows error variation on the data scale, and H is scaled conformally, allowing the group means to be shown on the same scale. The natural scaling of H and E as generalized mean squares would be H/dfh and E/dfe, which is obtained using SCALE=dfe/dfh 1. Equivalently, the E matrix can be shrunk by the same factor by specifying SCALE=1 dfh/dfe.

VAXIS=  Name of an AXIS statement for the y variable

HAXIS=  Name of an AXIS statement for the x variable

LEGEND=  Name of a LEGEND statement. If not specified, a legend for the M1 and M2 matrices is drawn beneath the plot. Specify LEGEND=NONE to suppress the legend.

COLORS=  Colors for the H and E ellipses. [Default: COLORS=BLACK RED]

LINES=  Line styles for the H and E ellipses. [Default: LINES=1 21]

WIDTH=  Line widths for the H and E ellipses. [Default: WIDTH=3 2]

HTEXT=  Height of text in the plot. If not specified, the global graphics option HTEXT controls this.
A.3. The \texttt{HEMAT} macro: HE plots for all pairs of response variables

The \texttt{HEMAT} macro plots the covariance ellipses for a hypothesized (H) effect and for error (E) for all pairs of variables from a MANOVA or multivariate multiple regression.

\textit{Method}

The macro calls \texttt{HEPLOT} within nested \texttt{%do} ... \texttt{%end} loops to plot all pairs of responses. This is wrapped with calls to the \texttt{GDISPLAY} macro to suppress display of the individual plots. The final display is produced by \texttt{PROC GREPLAY}. In order to make this macro reusable with a single SAS session, the separate plots are saved in a temporary graphics catalog (\texttt{GTEMP=work.gtemp}), which is normally deleted at the end (\texttt{GKILL=Y}).

\textit{Usage}

The \texttt{HEMAT} macro is defined with keyword parameters. The arguments may be listed within parentheses in any order, separated by commas. For example:

\begin{verbatim}
%hemat(data=iris, stat=stats, 
  var=SepalLen SepalWid PetalLen PetalWid, 
  effect=species);
\end{verbatim}

\textit{Parameters}

\begin{itemize}
  \item \texttt{DATA=} Name of the raw data set to be plotted. [Default: \texttt{DATA=\_LAST\_}]
  \item \texttt{STAT=} Name of \texttt{OUTSTAT=} dataset from \texttt{PROC GLM}
  \item \texttt{EFFECT=} Name of MODEL effect to be displayed for the H matrix. This must be one of the terms on the right hand side of the MODEL statement used in the \texttt{PROC GLM} or \texttt{PROC REG} step, in the same format that this effect is labeled in the \texttt{STAT=} dataset. This must be one of the values of the \_SOURCE\_ variable contained in the \texttt{STAT=} dataset.
  \item \texttt{VAR=} Names of response variables to be plotted - can be a list or \texttt{X1-X4} or \texttt{VARA-VARB}. [Default: \texttt{VAR=\_NUMERIC\_}]
  \item \texttt{NAMES=} Alternative variable names (used to label the diagonal cells.)
  \item \texttt{M1=} First matrix: either H or H+E. [Default: \texttt{M1=H}]
  \item \texttt{M2=} Second matrix either E or I. [Default: \texttt{M2=E}]
  \item \texttt{SCALE=} Scale factors for M1 and M2. See description in \texttt{HEPLOT}
  \item \texttt{HTITLE=} Height of variable name in diagonal cells
  \item \texttt{SYMBOLS=} Not used
  \item \texttt{COLORS=} Colors for the H and E ellipses. [Default: \texttt{COLORS=BLACK RED}]
\end{itemize}
A.4. The HEMREG macro: Extract H and E matrices for multivariate regression

The HEMREG macro extracts hypothesis (H) and error (E) matrices for an overall test in a multivariate regression analysis, in a form similar to that provided by the OUTSTAT= option with PROC GLM. This is typically used with the HEPLOT macro, or the MPOWER macro for MMRA.

Method
For a multivariate regression analysis, using

```
proc glm outstat=stats;
   model y1 y2 y3 = x1-x5;
```

PROC GLM will produce 5 separate 3x3, 1 df SSCP matrices for the separate predictors X1-X5, in the OUTSTAT= data set, but no SSCP matrix for the overall multivariate test. The HEMREG macro uses PROC REG instead, obtains the HypothesisSSCP and ErrorSSCP tables using ODS, and massages these into the same format used by PROC GLM.

Usage
The HEMREG macro is defined with keyword parameters. The Y= and X= parameters are required. One or more overall hypotheses involving subsets of the X= variables may be specified with the MTEST= parameter. The arguments may be listed within parentheses in any order, separated by commas. For example:

```
%hemreg(y=SAT PPVT RAVEN, x=N S NS NA SS);
%hemreg(y=SAT PPVT RAVEN, x=N S NS NA SS, mtest=%str(N,S,NS), hyp=N:S:NS);
```

Parameters

- **DATA=** Name of input dataset. [Default: DATA=_LAST_]
- **Y=** List of response variables. Must be an explicit, blank-separated list of variable names, and all variables must be numeric.
- **X=** List of predictor variables. Must be an explicit, blank-separated list of variable names, and all variables must be numeric.
- **HYP=** Name for each overall hypothesis tested, corresponding to the test(s) specified in the MTEST= parameter (to be used as the EFFECT= parameter in the HEPLOT macro). [Default: HYP=H1]
MTEST= If MTEST= is not specified (the default), a multivariate test of all X= predictors is carried out, giving an overall H matrix. Otherwise, MTEST= can specify one or more multivariate tests of subsets of the predictors, separated by '/', where the variables within each subset are separated by ','. In this case, the embedded ','s must be protected by surrounding the parameter value in %str(). For example,

MTEST = %str(group / x1, x2, x3 / x4, x5)

In this case you might specify HYP=Group X1:X3 X4:X5 to name the H matrices.

SS= Type of SSCP matrices to compute: Either SS1 or SS2, corresponding to sequential and partial SS computed by PROC REG. If SS=SS2, the _TYPE_ variable in the output data set is changed to _TYPE_='SS3' to conform with PROC GLM. [Default: SS=SS2]

OUT= The name of output HE dataset. [Default: OUT=HE]
References


**Affiliation:**
Michael Friendly  
Psychology Department  
York University  
Toronto, ON, M3J 1P3  
Canada  
E-mail: friendly@yorku.ca  