Revista Colombiana de Estadística
Current Topics in Statistical Graphics
Diciembre 2014, volumen 37, no. 2, pp. 261 a 283
DOI: http://dx.doi.org/10.15446/rce.v37n2spe.47934

Recent Advances in Visualizing Multivariate Linear Models

Avances recientes para la visualización de modelos lineales multivariados

Michael Friendly[1,a], Matthew Sigal[1,b]

[1]Department of Psychology, Faculty of Health, York University, Toronto, Canada

Abstract

This paper reviews our work in the development of visualization methods (implemented in R) for understanding and interpreting the effects of predictors in multivariate linear models (MLMs) of the form $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{B} + \boldsymbol{U}$, and some of their recent extensions.

We begin with a description of and examples from the Hypothesis-error (HE) plots framework (utilizing the `heplots` package), wherein multivariate tests can be visualized via ellipsoids in 2D, 3D or all pairwise views for the Hypothesis and Error Sum of Squares and Products (SSP) matrices used in hypothesis tests. Such HE plots provide visual tests of significance: a term is significant by Roy's test if and only if its H ellipsoid projects somewhere outside the E ellipsoid. These ideas extend naturally to repeated measures designs in the multivariate context.

When the rank of the hypothesis matrix for a term exceeds 2, these effects can also be visualized in a reduced-rank canonical space via the `candisc` package, which also provides new data plots for canonical correlation problems. Finally, we discuss some recent work-in-progress: the extension of these methods to robust MLMs, development of generalizations of influence measures and diagnostic plots for MLMs (in the `mvinfluence` package).

***Key words*:** Graphics, Multivariate Analysis, Software, Visualization.

[a]Professor. E-mail: friendly@yorku.ca
[b]Professor. E-mail: msigal@yorku.ca

Resumen

Este artículo hace una revisión de los desarrollos recientes en métodos de visualización (implementados en R) para la comprensión e interpretación de los efectos de los predictores en modelos lineales multivariados (MLMs) de la forma $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{B} + \boldsymbol{U}$ y sus extensiones recientes.

Comenzamos con una descripción y ejemplos de los gráficos de Hipótesis-Error (HE), (utilizando el paquete `heplots`) en los cuales los tests multivariados son visualizados vía elipsoides en 2D, 3D o todas las vistas pareadas de las matrices de sumas de cuadrados y productos (SSP por sus siglas en inglés) de Hipótesis y Error. Las gráficas HE permiten pruebas de significancia visuales: un término es significativo en el test de Roy si y solo si su elipsoide H es proyectado fuera del elipsoide E. Estas ideas se extienden a diseños de medidas repetidas en el contexto multivariado.

Cuando el rango de la matriz de hipótesis para un término es mayor a 2, estos efectos pueden ser visualizados en un espacio canónico de rango reducido vía el paquete `candisc`, que a su vez también permite nuevos gráficos para problemas de correlación canónica. Finalmente, se discuten algunas áreas de investigación en desarrollo: la extensión de estos métodos a MLMs robustos, generalizaciones de las medidas de influencia y gráficas de diagnóstico para MLMs (en el paquete `mvinfluence`).

***Palabras clave*:** análisis multivariado, gráficas, software, visualización.

1. Introduction

Multivariate response data are very common in applied research. A research outcome (e.g., depression, job satisfaction, academic achievement, childhood attention deficit hypheractivily disorders-ADHD) may have several observed measurement scales or related aspects. In this framework, the primary concern of the researcher is to ascertain the impact of potential predictors on two or more response variables. For example, if adolescent academic achievement is measured by reading, mathematics, science, and history scores, do predictors such as parent encouragement, socioeconomic status and school environmental variables affect all of these outcomes? And, do they affect these outcomes in similar or different ways?

Statistically, this is easy, because the classical univariate response model for ANOVA and regression, $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u}$, with $\boldsymbol{u} \sim \mathcal{N}(0, \sigma^2\boldsymbol{I})$ generalizes directly to an analogous multivariate linear model (MLM), $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{B} + \boldsymbol{U}$ for multiple responses, $\boldsymbol{Y} = (\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_p)$. Happily as well, hypothesis tests for the MLM are also straight-forward generalizations of the familiar F and t tests for univariate response models.

However, due to the possibly high-dimensional nature of the data, visualizations for multivariate response models are not as straightforward as they are for simpler, univariate models, either for understanding the effects of predictors, model parameters, or for standard model diagnostics, such as influence plots. Consequently, the results of such studies are often discussed solely in terms of coefficients and significance, and visualizations of relationships are presented for one response variable at a time.

This is unfortunate, because visualization affords us a window to truly see what is happening in our data, and can aid in interpretation, yet the univariate graphical methods cannot show the important *relations* among the multivariate responses. The aim of this paper is to present a few methods that are currently employed for the visualization of high-dimensional data, and then review a series of methods we have worked on to apply these methods to MLMs. These extensions involve the Hypothesis-Error (HE) plot framework and the use of rank reduction, and can be utilized under a range of circumstances.

The plan of this paper is as follows: In Section 2, we illustrate three basic techniques that can be utilized in presentation graphics to enhance the usefulness and applicability of traditional statistical graphic displays for multivariate data. Next, Section 3 describes the framework for generating HE plots, that are useful for visualizing the relationships found in MLMs. The idea of low-dimensional visualization introduced in Section 2.2 is extended to MLMs in Section 4. Finally, in Section 5, we describe some recent extensions of these methods, designed to make graphical methods for MLMs more closely approximate the range of techniques available for univariate response models.

2. Basic Approaches to Visualizing Multivariate Data

Attempts to visualize multivariate data using static graphs almost always have to proceed by reducing the complexity of the data to what can be shown on a 2D page or screen image. The most common methods involve a pairwise display of bivariate relationships in a scatterplot matrix and various dimension-reduction techniques, as we illustrate below.

An important adjunct to these basic ideas is the use of *plot annotations* to highlight important features of the data or the relationships among variables that might not be seen otherwise or as readily. Another consequential idea is the use of *plot summaries* to suppress some details of the data that obscure the features we want to see.

To frame our approach in a general context, we begin with some examples of visualizations of the well-known iris dataset, made famous by Fisher (1936). This contains data on measurements of four characteristics of iris flowers (sepal length and width, and petal length and width), for three species of the flower (*setosa, virginica*, and *versicolor*).

2.1. Bivariate Views: Scatterplot Matrices

One basic method often employed when investigating the relationship between multiple response variables is to generate a scatterplot matrix, which features every possible pairwise relationship between the variables.

These allow us to notice trends between the response categories, and can be annotated to include more information, such as via shading to differentiate between

levels of a categorical variable, data ellipses for visualizing confidence intervals, or through the addition of simple regression lines. In each case, the analyst gets to observe how each variable relates to the others. For example, see Figure 1 for a scatterplot matrix concerning the four variables in the iris dataset, annotated with 68% data ellipses and simple regression lines for each type of flower.

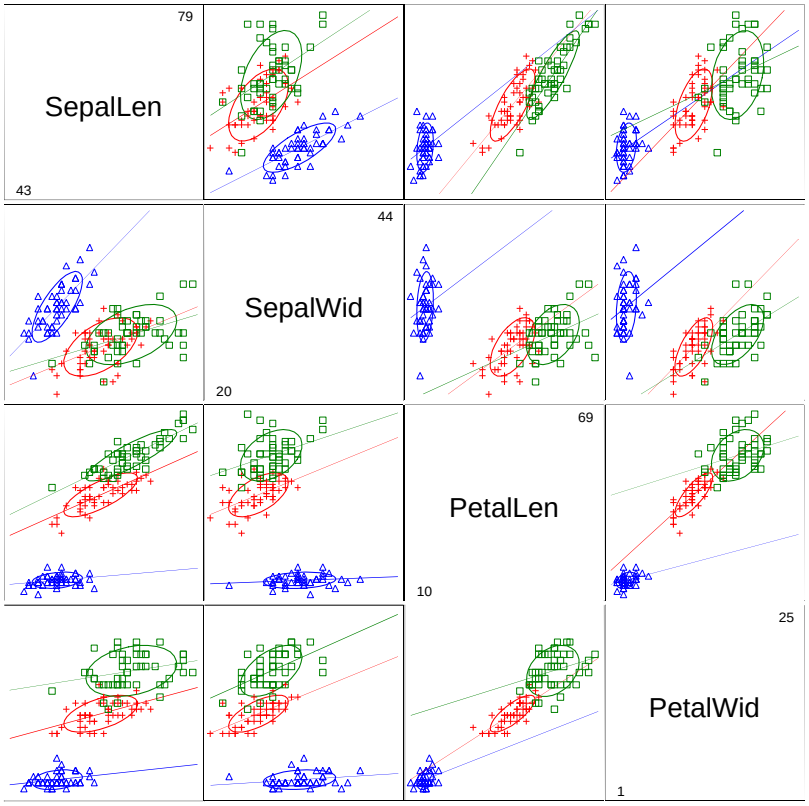

FIGURE 1: An enhanced scatterplot matrix, showing linear regression lines and 68% data ellipses by iris species, for all pairs of variables in the iris dataset.

However, this pairwise display does not yield any information on how the variables interact within a higher dimensional space. With three variables, we can no longer simultaneously plot and print the simple raw data on a static graphic, and must utilize software to view an interactive data cloud that can be rotated. Yet, with more variables, even this approach fails.

2.2. Low-Dimensional Views: Biplots

One way to understand how three or more dependent variables are related at once in a static graphic is to map the data into a low-dimensional space that preserves as much information in the data as possible. One such technique is the biplot (Gabriel 1971, Gabriel 1981, Gower, Lubbe & Roux 2011), which projects

the data into a 2D (or 3D) space accounting for the largest amount of variance in the data. The name "biplot" comes from the fact that this technique displays both the observations and the variables in a single, joint plot.

The standard biplot is essentially based on a principal components (or singular value) decomposition, plotting the component scores for the observations on the first two (or three) components. On such a plot, vectors (component loadings) are drawn for each variable that indicate the relationship they hold with each of the components. For example, see Figure 2. In this plot, one can observe that petal length and width are strongly related to differentiating between the species on the first component (plotted along the X axis), while sepal width differentiates between species on the vertical axis. These two dimensions account for 95.8% of the variance of the four iris variables, ensuring the adequacy of this visual summary. In less structured data, when the two dimensions account for less variance, the biplot still provides a useful 2D summary between the data, the variables, and the strongest available components. See Gower et al. (2011) for a wider presentation on this topic.

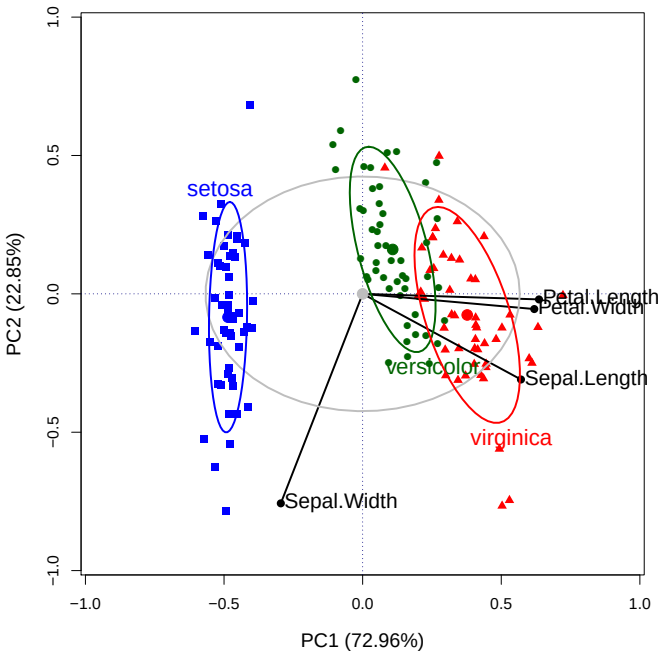

FIGURE 2: A biplot, visualizing the first two components from a PCA of the iris dataset. Data ellipses show the covariance structure for each species and for the total sample. Vectors show the projections of the original variables in this space.

2.3. Visual Summaries: Data Ellipses

The *data ellipse* (Monette 1990) (or *concentration ellipse*, Dempster, 1969, Ch. 7) provides a remarkably simple and effective display for viewing and understanding bivariate relationships in multivariate data. The data ellipse is typically used to add a visual summary to a scatterplot, indicating the means, standard deviations, correlation, and slope of the regression line for two variables. See Friendly, Monette & Fox (2013) for a complete discussion of the role of ellipsoids in statistical data visualization.

Formally, for two variables, Y_1, Y_2, the sample data ellipse $\mathcal{E}_c$ of size c is defined as the set of points $\boldsymbol{y} = (y_1, y_2)^\mathsf{T}$ whose squared Mahalanobis distance, $D^2(\boldsymbol{y}) = (\boldsymbol{y} - \bar{\boldsymbol{y}})^\mathsf{T} \boldsymbol{S}^{-1} (\boldsymbol{y} - \bar{\boldsymbol{y}})$, from the means, $\bar{\boldsymbol{y}}$, is less than or equal to c^2,

$$\mathcal{E}_c(\boldsymbol{y}; \boldsymbol{S}, \bar{\boldsymbol{y}}) \equiv \{\boldsymbol{y} : (\boldsymbol{y} - \bar{\boldsymbol{y}})^\mathsf{T} \boldsymbol{S}^{-1} (\boldsymbol{y} - \bar{\boldsymbol{y}}) \leq c^2\}, \tag{1}$$

where $\boldsymbol{S}$ is the sample variance-covariance matrix,

$$\boldsymbol{S} = (n-1)^{-1} \sum_{i=1}^{n} (\boldsymbol{y}_i - \bar{\boldsymbol{y}})^\mathsf{T} (\boldsymbol{y}_i - \bar{\boldsymbol{y}})$$

When $\boldsymbol{y}$ is (at least approximately) bivariate normal, $D^2(\boldsymbol{y})$ has a large-sample χ^2_2 distribution (χ^2 with 2 df), so taking $c^2 = \chi^2_2(0.68) = 2.28$ gives a "1 standard deviation bivariate ellipse", an analog of the standard interval $\bar{y} \pm 1s$, while $c^2 = \chi^2_2(0.95) = 5.99 \approx 6$ gives a data ellipse of 95% coverage. A bivariate ellipse of $\approx 40\%$ 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. These ideas generalize readily to p-dimensional ellipsoids, and we will use the term "ellipsoid" below to cover all cases.

Thus, in Figure 1, the data ellipses show clearly how the means, variances, correlations, and regression slopes differ systematically across the three iris species in all pairwise plots. The iris *setosa* flowers (in blue) are not only smaller on all variables than the other species, but its variances are also smaller, and the correlations differ from those of the other species.

Similarly, in the reduced-rank principal component analysis (PCA) view shown in Figure 2, you can see that the component scores are uncorrelated for *setosa* but slightly negatively correlated for the other two species. The data ellipse for the total sample (ignoring species), shows that the two components, PC1 and PC2, are of course uncorrelated, as guaranteed in PCA. The variable vectors in this plot show that PC1 is largely determined by three of the variables while PC2 is determined mainly by sepal width.

3. Hypothesis-Error (HE) Plots

Hypothesis-Error (HE) plots build upon the idea of the data ellipse, but attempt to also display indicators of significance and effect size for predictors (Friendly 2007),

by visualizing hypothesis and error covariation as ellipsoids. These plots can be generated using the `heplots` package (Fox, Friendly & Monette 2013) in R (R Core Team 2013), as shown in the Appendix, and, by default, will display significance in terms of the Roy's largest root test.

The MLM we want to visualize here is given by:

$$\boldsymbol{Y}_{n\times p} = \boldsymbol{X}_{n\times q}\, \boldsymbol{B}_{q\times p} + \boldsymbol{U}, \tag{2}$$

for p responses, $\boldsymbol{Y} = (\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_p)$, and q regressors in $\boldsymbol{X}$. The regressors can comprise quantitative predictors, factors (represented as dummy variables or contrasts), interaction terms, or any other term (e.g., a polynomial or spline) that can be represented within the framework of the general linear model.

The essential ideas here are that:

- Every multivariate hypothesis test is carried out by a multivariate analog of the general linear test, $H_0 : \boldsymbol{C}_{h\times q}\, \boldsymbol{B}_{q\times p} = \boldsymbol{0}_{h\times p}$, where the matrix of constants $\boldsymbol{C}$ selects subsets or linear combinations (contrasts) of the coefficients in $\boldsymbol{B}$ to be tested.

- Every such hypothesis gives rise to $(p \times p)$ matrices, $\boldsymbol{H}$ and $\boldsymbol{E}$ that are the multivariate analogs of the familiar sums of squares, SS_H and SS_E, used in univariate F tests.

$$\boldsymbol{H} = (\boldsymbol{C}\widehat{\boldsymbol{B}})^{\mathsf{T}}\, [\boldsymbol{C}(\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X})^{-}\boldsymbol{C}^{\mathsf{T}}]^{-1}, (\boldsymbol{C}\widehat{\boldsymbol{B}}), \tag{3}$$

$$\boldsymbol{E} = \boldsymbol{U}^{\mathsf{T}}\boldsymbol{U} = \boldsymbol{Y}^{\mathsf{T}}[\boldsymbol{I} - \boldsymbol{H}]\boldsymbol{Y}. \tag{4}$$

- The univariate F test statistic, $F = \frac{SS_H/df_h}{SS_E/df_e}$ has a direct multivariate analog in terms of the latent $s = \min(p, df_h)$ non-zero latent roots, λ_i, of $\boldsymbol{H}$ relative to $\boldsymbol{E}$ that solve

  $$\det(\boldsymbol{H} - \lambda\boldsymbol{E}) = 0$$

  The various multivariate test statistics such as Wilks' Λ, the Pillai and Hotelling trace criteria, and Roy's maximum root test are all functions of the λ_i that reflect different geometric properties of the size of the H ellipsoid relative to the size of the E ellipsoid. The statistical and geometric details are described in Friendly (2007) and Friendly et al. (2013).

An animated display of these ideas and the relations between data ellipses and HE plots can be seen at `http://www.datavis.ca/gallery/animation/manova/`.

3.1. Visualizing H and E (Co)variation

A standard ellipsoid representing residual (error) variation reflected in the $\boldsymbol{E}$ matrix is simply the data ellipse of the residuals in $\boldsymbol{U}$, scaled as $\boldsymbol{E}/df_e$. In an HE plot, we typically show this as an ellipse of 68% coverage, but centered at the overall means of the variables plotted, so that we can also show the means for

factors on the same scale. This also allows you to "read" the residual standard deviation as the half-length of the shadow of the E ellipsoid on any axis.

An ellipsoid representing variation in the means of a factor (or any other term reflected in (3) in the $\boldsymbol{H}$ matrix is simply the data ellipse of the fitted values for that term. Scaling the $\boldsymbol{H}$ matrix as $\boldsymbol{H}/df_e$ puts this on the same scale as the E ellipse, as shown in the left panel of Figure 3. We refer to this as *effect size scaling*, because it is similar to an effect size index used in univariate models, e.g., $ES = (x_2 - \bar{x}_2)/s$ in a two-group, univariate design.

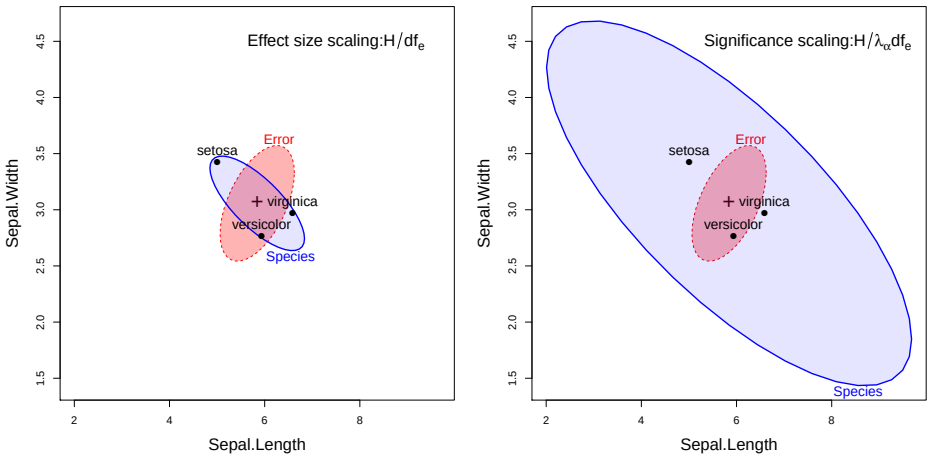

FIGURE 3: HE plots for two variables in the iris data set. Left: effect scaling of the $\boldsymbol{H}$ matrix. Right: significance scaling of the $\boldsymbol{H}$ matrix.

The geometry of ellipsoids and multivariate tests allow us to go further with a re-scaling of the H ellipsoid that gives a *visual test of significance* for any term in a MLM, simply by dividing H/df_e further by the α-critical value of the corresponding test statistic. Among the various multivariate test statistics, Roy's maximum root test gives $H/(\lambda_\alpha df_e)$ which has the visual property that the scaled H ellipsoid will protrude somewhere outside the standard E ellipsoid if and only if Roy's test is significant at significance level α. For these data, the HE plot using significance scaling is shown in the right panel of Figure 3.

You can interpret the HE plot by recalling that they reflect data ellipsoids of the fitted values (group means, here) and the residuals. The direction of the H ellipsoid relative to that of E reflects the linear combinations of the response variables shown that depart from the null hypothesis.

In this example, the iris data has $p = 4$ dimensions, but with three groups, $df_h = 2$, so the H and E ellipsoids all have $s = \min(p, df_h) = 2$ non-zero dimensions. To see these relations for all variables together, it is easy to use a scatterplot matrix format, as shown in Figure 4. The plot shown in Figure 3 is just the panel in row 1, column 2, and the entire plot in Figure 4 is just an alternative visualization of the data-based scatterplot matrix shown in Figure 1, focussed on the evidence for differences among the species means.

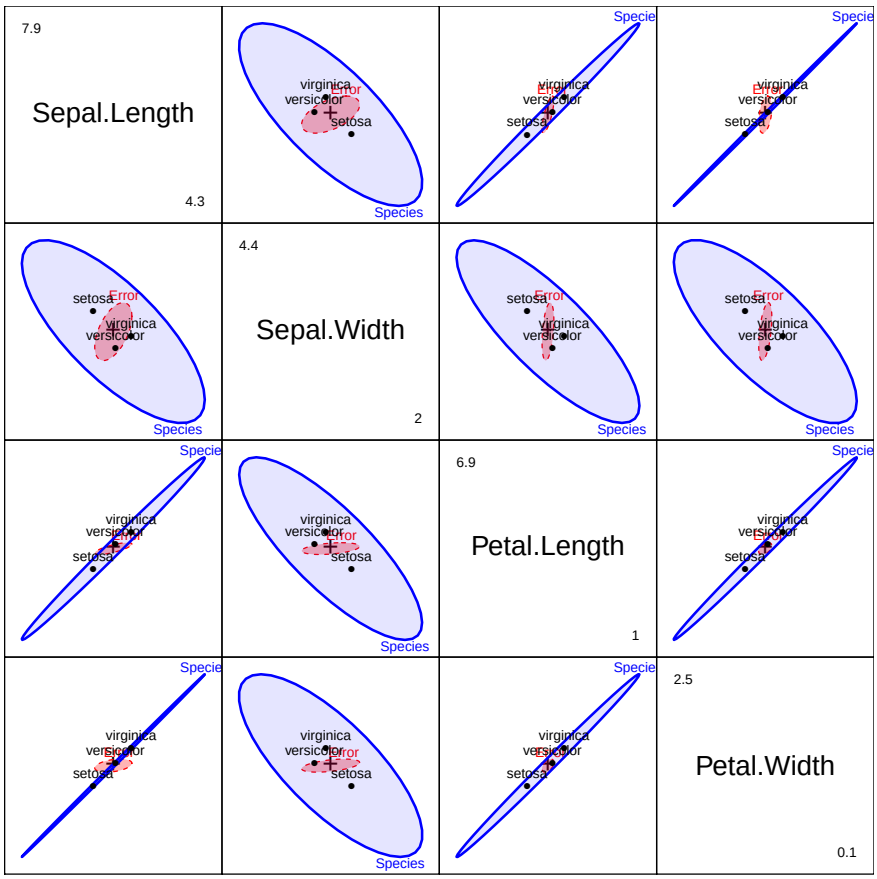

FIGURE 4: HE plot matrix for all variables in the iris data set. Each panel shows the HE plot as projected onto the plane of two response variables.

It is now easy to see the relationships among all four variables that are reflected in the multivariate tests of H_0: no differences among species means. For all variables except sepal width, the variation of the species means is essentially one-dimensional; however, sepal width has an opposite pattern to the others.

As a supplementary example, let's look at the Romano-British pottery dataset (Tubb, Parker & Nickless 1980). This data pertains to 26 ancient pottery samples found at four kiln sites in Britain (Ashley Rails, Caldicot, Isle of Thorns, and Llanedryn), and their chemical makeup. Each sample was quantified in terms of it's aluminum (Al), iron (Fe), magnesium (Mg), calcium (Ca), and sodium (Na) content, which naturally yields a one-way multivariate analysis of variance (MANOVA) design with 4 groups (site) and 5 responses (chemical composition). The primary question posed by this data is: Can the chemical composition of the samples differentiate the sites? And, how can we understand the contributions of the chemical elements to such a discrimination?

Using the HE plot framework, we can visualize each pairwise combination of the chemical compounds, and assess the sites in terms of effect or significance scaling, see the left panel of Figure 5. In this view, we can see that there does appear to be significant separation between the locations, with the Caldicot and Llanedryn being relatively high in iron but low in aluminum, while the Ashley Rails and Isle of Thorns samples being the opposite.

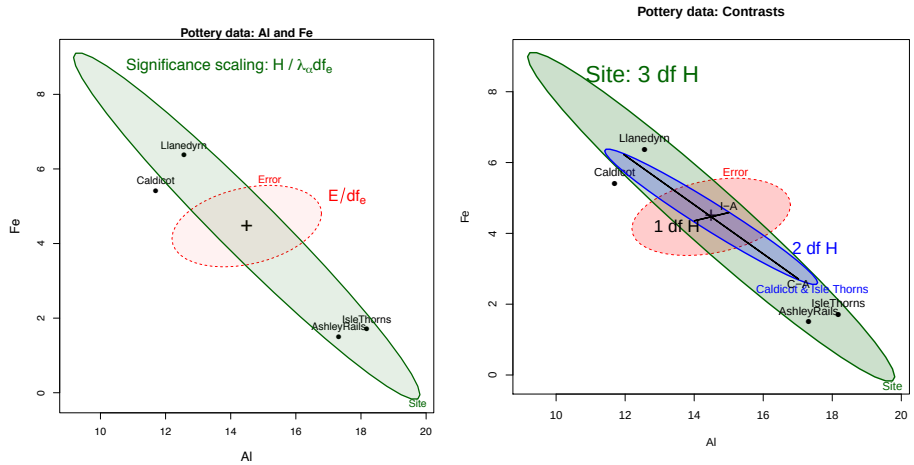

FIGURE 5: HE plots for a one-way MANOVA design using the Romano-British pottery dataset. Left: Significance scaling of the $\boldsymbol{H}$ matrix. Right: Visual decomposition of a multi-degree of freedom hypothesis test.

Further, we can utilize this approach to decompose a multi-degree of freedom hypothesis test into a set of orthogonal contrasts, exactly as the univariate SS_H may be decomposed in an ANOVA. Each subhypothesis matrix with rank greater than 1 will have hypothesis ellipse, while those of rank 1 will plot as a vector in the HE graphic. In this example, the overall 3 *df* hypothesis tests the equality of the four Site mean vectors. Also overlaid on this plot is the 2 *df* sub-hypothesis test, which is the contrast between the average of Caldicot and Isle Thorns against Ashley Rails. This in turn can be decomposed into two 1 df tests for each of Caldicot and Isle Thorns versus Ashley Rails, in which only the former is significant. The hypothesis ellipsoids in this plot have the interesting visual property that they form conjugate directions with respect to the ellipsoid for the joint test, provided the sub-hypotheses are statistically independent. More details about this approach can be found in Friendly et al. (2013).

3.2. More General Models

The HE plot framework extends quite naturally to all cases of the general multivariate linear model. This includes multiple MANOVA models with two or more factors (and their interactions), multivariate analysis regression (MMRA), models with mixtures of factors and quantitative regressors (MANCOVA) (Fox,

Friendly & Weisberg 2013) and repeated measure designs (Friendly 2010). More importantly, HE plots provide a visualizations of such complex models from which many features can be "read" far more directly than from the collection of many numerical tables that they summarize.

All hypothesis tests for these models can be formulated as special cases of the general linear test, $H_0 : \boldsymbol{C}_{h\times q}\boldsymbol{B}_{q\times p} = \boldsymbol{0}_{h\times p}$, giving rise to $\boldsymbol{H}$ and $\boldsymbol{E}$ matrices as indicated in (3) and (4). For example, in a MMRA model, the test of the multivariate hypothesis $H_0 : \boldsymbol{B} = \boldsymbol{0}$ that all predictors have no effect on any responses is given by specifying $\boldsymbol{C} = \boldsymbol{I}_{q\times q}$, while the multivariate test for the i^{th} predictor is given by using $\boldsymbol{C} = (0, 0, \ldots, 1, 0, \ldots 0)$, with a value of 1 in position i.

We illustrate the flexibility of these models and HE plots using data from an experiment by William Rohwer (Timm 1975, Table 4.7.1) on kindergarten children ($n = 37$ of low socioeconomic status (SES), $n = 32$ of high SES) designed to examine how well performance on a set of paired-associate (PA) tasks can predict performance on measures of aptitude and achievement: the Peabody Picture Vocabulary Test (PPVT), a Student Achievement Test (SAT), and the Raven Progressive matrices test. The PA tasks varied in how the stimuli were presented, and are called *named* (n), *still* (s), *named still* (ns), *named action* (na), and *sentence still* (ss).

A simple MANCOVA model, allowing different intercepts (means) for the SES groups, but assuming that the regression slopes for the PA tasks are the same for both groups can be fit in R as follows:

```
rohwer.mod <- lm(cbind(SAT, PPVT, Raven) ~ SES + n + s + ns + na + ss,
                 data=Rohwer)
```

Multivariate hypothesis tests show that only SES, ns and na have significant effects. But, how can we understand these results and the nature of the relationships in these data?

The left panel of Figure 6 shows the HE plot for this model, for the first two variables, SAT and PPVT. The ellipsoid labeled "Regr" gives the result of an overall multivariate test for all of the PA tests jointly. As can be seen based upon its extension past the error ellipse, this test is highly significant. Note that the predictors are all 1 df_h terms, so the $\boldsymbol{H}$ matrices are all of rank 1, and their H ellipsoids degenerate to lines.

We can interpret this display as follows: (a) The *length* of each predictor line indicates the strength of its relationship to the two responses jointly. (b) Only the predictor lines for na and ns lie outside the E ellipsoid, and the latter is just barely significant. (c) The *orientation* of each predictor line shows its relationship to SAT and PPVT. (d) The means for the SES groups show that the high group performs better on both the SAT and the PPVT, but more so on the PPVT. (e) The Regr ellipsoid for the overall test of the regression variables can be seen as a sum of the contributions of the individual predictors, some of which make it larger in the direction of SAT, others in the direction of PPVT. (f) The orientation of the E ellipsoid indicates a small positive correlation between the residuals for SAT and

PPVT; the shadows of this ellipsoid on the horizontal and vertical axes reflect the residual standard deviations, apparently larger for SAT than PPVT.

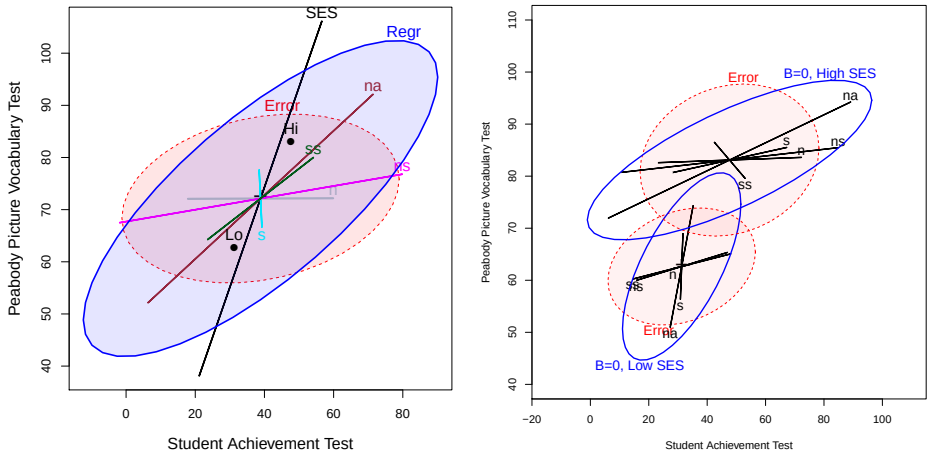

FIGURE 6: HE plots for SAT and PPVT in the Rohwer data. Left: Fitting a MANCOVA model, assuming equal slopes for the SES groups. Right: Fitting a model of heterogeneous regressions.

To test the assumption of equal slopes in the simple MANCOVA model, we can fit extended models that allow heterogeneous slopes and intercepts in the two groups. One way to do this is to fit separate multivariate regression models for the two groups, and overlay the HE plots on common scales. This result is shown in the right panel of Figure 6.

Some additional aspects that can be seen here are: (a) The centers of the ellipsoids show the group means that were reflected in the SES factor in the MANOVA. (b) The overall regression ellipsoid testing $H_0 : \boldsymbol{B} = \boldsymbol{0}$ is more aligned with the SAT axis for the high SES group than for the low group, reflecting the better prediction of SAT than PPVT for the high group. (c) Among the individual predictors, na and ns are more important predictors in the high SES group, and are also more important in predicting SAT and PPVT.

In closing this section, we hope we have convinced the reader that HE plots, once you learn how to read them, provide much more direct information about the relationships between predictors and responses in complex MLMs than is easy to understand from tables or univariate displays.

In Figure 6 we only showed the HE plots for two of the response variables. As shown in Figure 7, we can extend this to any number of responses using an HE plot matrix. However, the idea of a low-dimensional view for multivariate data *per se* discussed in Section 2.2 has a direct connection with similar low-D views for MLMs, that we will discuss in the following section.

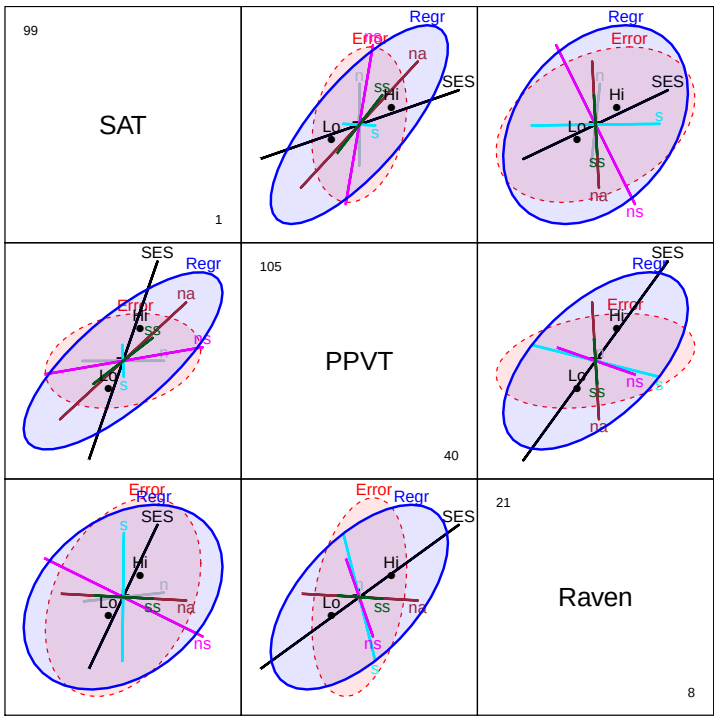

FIGURE 7: HE plot matrix for all three responses in the Rohwer data.

4. Generalized Canonical Discriminant HE Plots

For even a one-way MANOVA design with three or more response variables, it is difficult to visualize how the groups vary on all responses together, and how the different variables contribute to discrimination among groups. In this situation, canonical discriminant analysis (CDA) is often used, to provide a low-D visualization of between-group variation, analogous to the biplot technique for purely quantitative variables.

CDA amounts to a transformation of the p responses, $\boldsymbol{Y}_{n\times p}$ into the canonical space, $\boldsymbol{Z}_{n\times s} = \boldsymbol{Y}\boldsymbol{E}^{-1/2}\boldsymbol{V}$, where $\boldsymbol{V}$ contains the eigenvectors of $\boldsymbol{H}\boldsymbol{E}^{-1}$ and $s = \min(p, df_h)$. It is well-known (e.g., Gittins, 1985) that *canonical discriminant plots* of the first two (or three, in 3D) columns of $\boldsymbol{Z}$ corresponding to the largest canonical correlations provide an optimal low-D display of the variation between groups relative to variation within groups.

For a one-way design, the *canonical HE plot* is simply the HE plot of the canonical scores in the analogous MLM model that substitutes $\boldsymbol{Z}$ for $\boldsymbol{Y}$. This is shown in Figure 8 for the iris data. The interpretation of this plot is the same as before: if the hypothesis ellipse extends beyond the error ellipse, then that dimension is significant. Vectors for each predictor are then superimposed and demonstrate the relation between each and the two canonical dimensions.

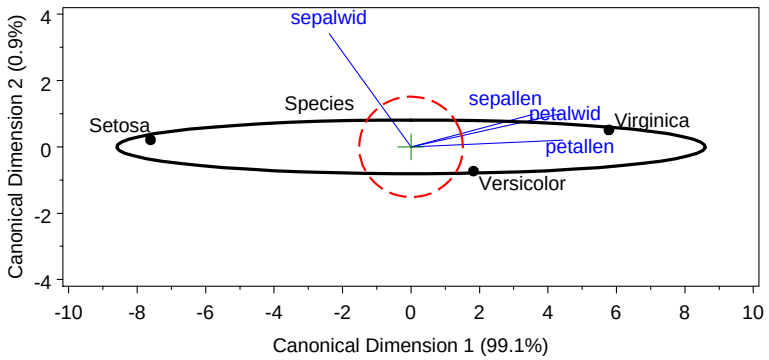

FIGURE 8: A canonical HE plot, visualizing the two canonical dimensions of the iris dataset.

The interpretation of this plot is quite simple: In canonical space, variation of the means for the iris species is essentially one-dimensional (99.1% of the effect of species), and this dimension corresponds to overall size of the iris flowers. All variables except for sepal width are aligned with this axis.

We have extended this to *generalized canonical discriminant plots* for the general MLM as follows: (a) Let t index the various hypothesized terms in an arbitrary MLM. Canonical discriminant analysis can be extended by performing the canonical analysis of the $\boldsymbol{H}_t$ and $\boldsymbol{E}$ matrices for each term in the model, based on the eigenvalues and eigenvectors of $\boldsymbol{H}_t\boldsymbol{E}^{-1}$. Then, for term t: (b) The canonical discriminant HE plot is the HE plot of canonical scores in the model $\boldsymbol{Z}_t \sim \bullet$, where $\bullet$ symbolizes all terms in the original MLM for $\boldsymbol{Y}$. These and other extensions of canonical analysis are implemented in the `candisc` package in R (Friendly & Fox 2013).

An application of this idea is shown in Figure 9 for the Rohwer data. In this case, the two dimensions account for 93.7% of the variability in achievement scores, and each predictor is plotted in relation to the two canonical dimensions. It is directly observable how both the predictors in the model and the outcome variables relate to the canonical dimensions. For instance, the SAT and PPVT extend horizontally, reflecting a strong loading on dimension 1, while the Raven is aligned more vertically, and is more strongly associated with dimension 2. For the predictors, *ss* is almost perfectly aligned with dimension 1, while the other variables load on both. Similar to the biplots that were previously discussed, this plot provides a compact 2D view for a complex multivariate problem.

5. Recent Extensions

The general goal of this work has been to extend data visualization methods for univariate response models to their counterparts for MLMs. Towards this end, we describe some recent efforts to address: The interpretation of model coefficients, the use of such plots in robust analyses, and in examining data for influential cases.

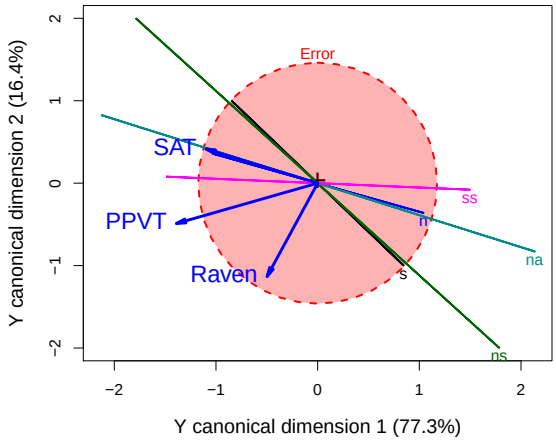

FIGURE 9: Generalized canonical HE plot, visualizing the two canonical dimensions of the Rohwer dataset.

5.1. Coefficient Plots for MLMs

In presentations and articles, it is commonplace to present results of fitted models in tables of estimated coefficients and their standard errors. This practice, while convenient, is now often deprecated (Gelman, Pasarica & Dodhia 2002, Kastellec & Leoni 2007) in favor of plots of point estimates and confidence intervals that, when carefully done, can communicate findings more clearly.

The left panel of Figure 10 gives an example of a univariate coefficient plot for the results of the linear model for SAT in the Rohwer data, showing 68% and 95% confidence intervals set at $\beta \pm \{1, 2\} se_\beta$. The corresponding MLM would require three such plots, one for each response variable and could not indicate multivariate confidence regions for joint hypothesis tests.

However, a simple generalization of this idea is the *multivariate* coefficient plot, which is illustrated in bivariate form in the right panel of Figure 10. To simplify the plot, this shows only joint 68% confidence ellipses, corresponding to a multivariate version of $\boldsymbol{\beta} \pm se_{\boldsymbol{\beta}}$. This has the property that a joint, multivariate test of $\boldsymbol{H}_0 : \boldsymbol{\beta} = \mathbf{0}$ is rejected when the confidence ellipse does not cover the origin (as shown by shading). See Friendly et al. (2013) for more details.

5.2. Robust MLMs

All calculations and test statistics for classical, normal-theory linear models are based on standard, first and second moment summaries, such as mean vectors and covariance matrices. It is well-known that these can be distorted by multivariate outliers, particularly in smaller samples.

In principle, such effects in standard multivariate analyses can often be countered by using robust mean and covariance estimates, such as simple multivariate trim-

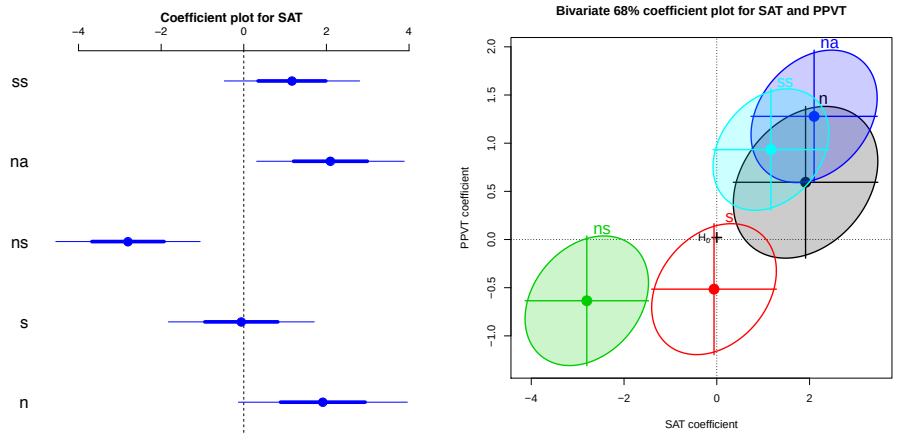

FIGURE 10: Coefficient plots for the Rohwer data. Left: Plot of coefficients in the univariate linear model for SAT. Right: Bivariate coefficient plot for SAT and PPVT. 68% confidence ellipsoids that exclude the origin, corresponding to $\boldsymbol{H}_0 : \boldsymbol{\beta} = \mathbf{0}$, are shaded.

ming (Gnanadesikan & Kettenring 1972) or the high-breakdown bound minimum volume ellipsoid (MVE) and minimum covariance determinant (MCD) methods (Rousseeuw & Leroy 1987, Rousseeuw & Van Driessen 1999). Often, these robust methods supply weights that can be used to "robustify" other multivariate methods and visualization techniques, but this integration in applied software is still quite spotty, inference for the general MLM is weak and associated graphical methods are limited.

In the `heplots` package (Fox, Friendly & Monette 2013), we extend normal-theory HE plots to robust equivalents via similar use of weighting and robust covariance estimation, using a simple iteratively reweighted least squares approach. Other approaches do provide high breakdown bound estimates, and will be implemented in the future. As an example, we have reestimated the MANOVA model with the Romano-British pottery data utilizing the `robmlm` function in the `heplots` package. Fitting is done by iterated re-weighted least squares, using weights based on the Mahalanobis squared distances of the current residuals from the origin, and a scaling (covariance) matrix, and its visual output is shown in Figure 11.

In this example, the figure on the left illustrates changes in observation weight. It clearly shows that three data points from Llanedyrn and one from Ashley Rails are atypical and given weights close to zero. The right figure demonstrates the difference between the classical and robust estimates. It is interesting to note that there is only minimal change in the mean ellipse when using the robust estimation, but there is a substantial difference in the error covariance ellipse. This is a quick representation that visually demonstrates how effects become more visible using robust methods.

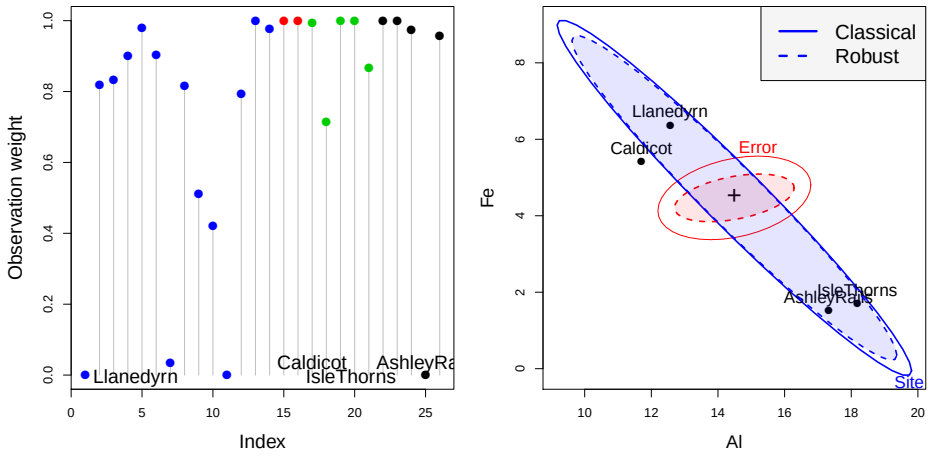

FIGURE 11: Robust MLMs. Left: Observation weights estimates by reweighted LS. Right: HE plot, overlaying the classical and robust estimates of H and E.

5.3. Influence Diagnostics for MLMs

A wide variety of influence diagnostics (e.g., Cook's distance, DFFITS, Hat values) and associated plots (leverage–influence plots) for detecting influential observations in univariate models have been available for a long time (Cook & Weisberg 1982). As well, a general theory of influence diagnostics for MLMs (Barrett & Ling 1992) is available to support these measures, but there is no available software implementation allowing these methods to be used. Our `mvinfluence` package is an initial implementation of some of these ideas (Friendly 2012).

The generalization of influence measures to multivariate response models is relatively straight-forward, and also extend to case deletion diagnostics for subsets (I) of size $m > 1$. For example, the multivariate analog of Hat values, that measure *leverage* of observations in terms of the predictors is

$$H_I = \boldsymbol{X}_I (\boldsymbol{X}^\mathsf{T}\boldsymbol{X})^{-1} \boldsymbol{X}_I^\mathsf{T},$$

where $\boldsymbol{X}_I$ refers to the rows of $\boldsymbol{X}$ in subset I.

The multivariate version of Cook's distance can be represented as the standardized distance between the estimated coefficients $\boldsymbol{B}$ using all cases and $\boldsymbol{B}_{(I)}$ estimated omitting subset I.

$$D_I = [vec(\boldsymbol{B} - \boldsymbol{B}_{(I)})]^\mathsf{T} \, [\boldsymbol{S}^{-1} \otimes (\boldsymbol{X}^\mathsf{T}\boldsymbol{X})] \, [vec(\boldsymbol{B} - \boldsymbol{B}_{(I)})]. \tag{5}$$

A wide variety of other influence measures is also defined (e.g., DFFITS, COVRATIO), but all can be expressed in the general form

$$\text{Influence}_I = \text{Leverage}_I \times \text{Residual}_I \tag{6}$$

for some function of leverage and the multivariate residuals corresponding to subset I.

We illustrate these methods using the MLM for the Rohwer data using only the low SES group in Figure 12. The left panel is a bubble plot of Cook's D_I against leverage, H_I for subsets of size $m = 1$, also showing D_I by the area of the bubble symbol. It can be seen that observation 5 is highly influential, but four other cases also have Cook's distances greater than the nominal cutoff for identifying "large" values.

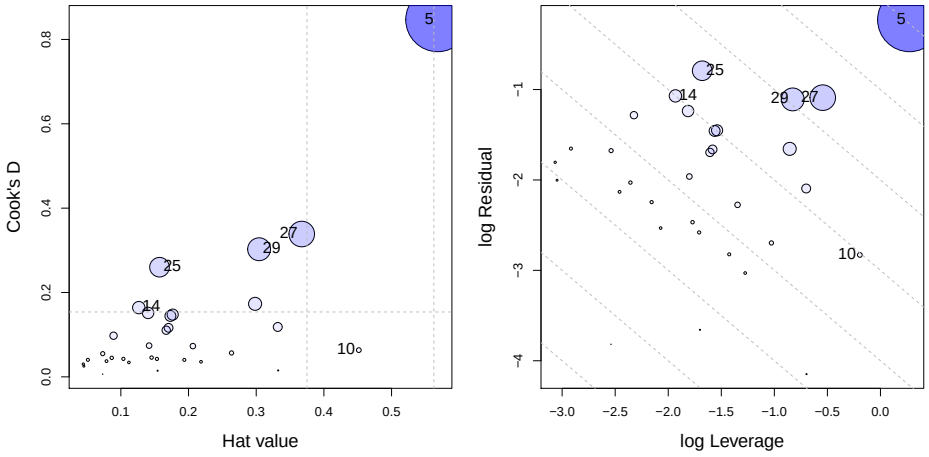

FIGURE 12: Influence plots for MLMs, using Rohwer data Left: Generalized Cook's D vs. Hat value. Dashed lines in the plot indicate conventional cutoffs for "large" values. Right: Leverage–Residual LR plot. Dashed lines represent contours of constant influence.

The right panel shows a novel form of influence plot (the LR plot) suggested originally by McCulloch & Meeter (1983). From (6), it follows that a plot of log(Leverage) vs. log(Residual) will have contours of constant influence along lines with slope $= -1$. This plot simplifies the interpretation of influence plots by placing all observations in relation to their influence away from the origin.

6. Summary and Conclusion

The classical univariate general linear model is the cornerstone for the development of much of modern statistical theory and practice. A great deal of the applied usefulness of this methodology stems from development of a wide range of visualization methods allowing researchers to see and understand their data, diagnose possible problems with assumptions and model specification and communicate their results effectively.

As we have argued, the extension of this model to multivariate responses is well-developed in theory and increasingly prevalent in applied research. Yet, the analogous advancement of visualization methods for MLMs has lagged behind. In

this paper we have set out a framework for filling these gaps, based on the following general ideas:

Data ellipsoids provide a simple visual summary of bivariate relations under classical (Gaussian) assumptions.

- They highlight important differences among groups (means, variances, correlations) in MANOVA designs;
- They can be embedded in scatterplot matrix format to show all pairwise, bivariate relations;
- They extend easily to 3D visualizations, and can be modified to use robust estimators.

HE plots provide a visual summary of multivariate hypothesis tests for all MLM models.

- They show H ellipsoids with group means for MANOVA factors and 1 df_h ellipsoids for quantitative predictors in ways that facilitate interpretation of multivariate effects.
- They can be embedded in an HE plot matrix to show all pairwise views.

Dimension-reduction techniques provide low-dimensional (2D or 3D), approximate visual summaries for high-D data.

- The biplot shows multivariate observations and variable vectors in the low-D view that accounts for the maximum variance in the data.
- Canonical HE plots are similar, but show the dimensions that account for maximal discrimination among groups or maximal canonical correlation.

In the popular children's TV show *Sesame Street*, it was common to sign off with "Today's show has been brought to you by the letter E", where they might have featured elephants, eagles, emus, and ellipses. In a similar vein and as a coda to this paper, we also remark that this approach has been provided by the beautiful and useful connections that exist among aspects of statistical models, matrix algebra, and geometry (Friendly et al. 2013). There are ellipsoids everywhere and almost all the properties of multivariate tests and dimension-reduction techniques can be understood in terms of eigenvalue decompositions. This framework provides opportunities for further extensions.

R code for some of the figures in this paper are included in the Appendix and related examples in the web supplement for a conference presentation, `http://www.datavis.ca/papers/ssc2013/`.

Acknowledgments

This work was supported by Grant OGP0138748 from the National Sciences and Engineering Research Council of Canada to Michael Friendly. We are grateful to John Fox, whose work on R packages (`car`, `heplots`) provided the necessary infrastructure for much of this work.

[Recibido: mayo de 2014 — Aceptado: noviembre de 2014]

References

Barrett, B. E. & Ling, R. F. (1992), 'General classes of influence measures for multivariate regression', *Journal of the American Statistical Association* **87**(417), 184–191.

Cook, R. D. & Weisberg, S. (1982), *Residuals and Influence in Regression*, Chapman and Hall, New York.

Dempster, A. P. (1969), *Elements of Continuous Multivariate Analysis*, Addison-Wesley, Reading, MA.

Fisher, R. A. (1936), 'The use of multiple measurements in taxonomic problems', *Annals of Eugenics* **8**, 379–388.

Fox, J., Friendly, M. & Monette, G. (2013), *heplots: Visualizing Tests in Multivariate Linear Models*. R package version 1.0-11.
*http://CRAN.R-project.org/package=heplots

Fox, J., Friendly, M. & Weisberg, S. (2013), 'Hypothesis tests for multivariate linear models using the *car* package', *The R Journal* **5**(1), 39–52.

Friendly, M. (2007), 'HE plots for multivariate general linear models', *Journal of Computational and Graphical Statistics* **16**(2), 421–444. doi: 10.1198/106186007X208407.
*http://www.math.yorku.ca/SCS/Papers/jcgs-heplots.pdf

Friendly, M. (2010), 'HE plots for repeated measure designs', *Journal of Statistical Software* **37**(4), 1–37.
*http://www.jstatsoft.org/v37/i04

Friendly, M. (2012), *mvinfluence: Influence Measures and Diagnostic Plots for Multivariate Linear Models*. R package version 0.6.
*http://CRAN.R-project.org/package=mvinfluence

Friendly, M. & Fox, J. (2013), *candisc: Visualizing Generalized Canonical Discriminant and Canonical Correlation Analysis*. R package version 0.6-5.
*http://CRAN.R-project.org/package=candisc

Friendly, M., Monette, G. & Fox, J. (2013), 'Elliptical insights: Understanding statistical methods through elliptical geometry', *Statistical Science* **28**(1), 1–39. doi: http://dx.doi.org/10.1214/12-STS402.
*http://datavis.ca/papers/ellipses.pdf

Gabriel, K. R. (1971), 'The biplot graphic display of matrices with application to principal components analysis', *Biometrics* **58**(3), 453–467.

Gabriel, K. R. (1981), Biplot display of multivariate matrices for inspection of data and diagnosis, *in* V. Barnett, ed., 'Interpreting Multivariate Data', John Wiley and Sons, London, chapter 8, pp. 147–173.

Gelman, A., Pasarica, C. & Dodhia, R. (2002), 'Let's practice what we teach: Turning tables into graphs', *The American Statistician* **56**(2), 121–130.

Gittins, R. (1985), *Canonical Analysis: A Review with Applications in Ecology*, Springer-Verlag, Berlin.

Gnanadesikan, R. & Kettenring, J. R. (1972), 'Robust estimates, residuals, and outlier detection with multiresponse data', *Biometrics* **28**, 81–124.

Gower, J., Lubbe, S. & Roux, N. (2011), *Understanding Biplots*, Wiley. *http://books.google.ca/books?id=66gQCi5JOKYC

Kastellec, J. P. & Leoni, E. L. (2007), 'Using graphs instead of tables in political science', *Perspectives on Politics* **5**(04), 755–771. doi: 0.1017/S1537592707072209.

McCulloch, C. E. & Meeter, D. (1983), 'Discussion of outliers... by R. J. Beckman and R. D. Cook', *Technometrics* **25**, 152–155.

Monette, G. (1990), Geometry of multiple regression and interactive 3-D graphics, *in* J. Fox & S. Long, eds, 'Modern Methods of Data Analysis', Sage Publications, Beverly Hills, CA, chapter 5, pp. 209–256.

R Core Team (2013), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria. *http://www.R-project.org/

Rousseeuw, P. & Leroy, A. (1987), *Robust Regression and Outlier Detection*, John Wiley and Sons, New York.

Rousseeuw, P. & Van Driessen, K. (1999), 'A fast algorithm for the minimum covariance determinant estimator', *Technometrics* **41**, 212–223.

Timm, N. H. (1975), *Multivariate Analysis with Applications in Education and Psychology*, Wadsworth (Brooks/Cole), Belmont, CA.

Tubb, A., Parker, A. & Nickless, G. (1980), 'The analysis of Romano-British pottery by atomic absorption spectrophotometry', *Archaeometry* **22**, 153–171.

Appendix. Supplement to Friendly & Sigal

Sample R Code to Reproduce Figures 3 and 4

Hypothesis-Error (HE) Plots

```
library(heplots)   # load heplots library
data(iris)         # load data
```

Conduct the MLM:

```
mod <- lm(cbind(Sepal.Length, Sepal.Width, Petal.Length, Petal.Width) ~ Species,
          data = iris)
```

HE Plot with Effect-Size scaling:

```
heplot(mod,
       xlab = "Petal Length in cm.", ylab = "Sepal length in cm.",
       size = "effect.size",
       fill = TRUE,  # shade ellipses
       fill.alpha = c(0.3, 0.1), # set transparency for E and H ellipses
       ylim = c(1.5, 4.75), xlim = c(2, 10))
```

HE Plot with Signicance scaling:

```
heplot(mod,
       xlab = "Petal Length in cm.", ylab = "Sepal length in cm.",
       size = "evidence",
       fill = TRUE,
       fill.alpha = c(0.3, 0.1),
       ylim = c(1.5, 4.75), xlim = c(2, 10))
```

Pairs Plot can be used to show each projection of response variables:

```
pairs(mod,
      fill = TRUE,
      fill.alpha = c(0.3, 0.1))
```

Sample R Code to Reproduce Figure 5 and 6

The MANCOVA Model:

```
rohwer.mod <- lm(cbind(SAT, PPVT, Raven) ~ SES + n + s + ns + na + ss,
                 data = Rohwer)
```

Colours to be used in the graphic:

```
col <- c("red", "black", "gray", "cyan", "magenta", "brown", "green", "blue")
```

Figure 5

Add ellipse to test all 5 regressors:

```
heplot(rohwer.mod,
       hypotheses = list("Regr" = c("n", "s", "ns", "na", "ss")),
       xlab = "Student Achievement Test",
       ylab = "Peabody Picture Vocabulary Test",
       cex.lab = 1.25, cex = 1.25,
       col = col, fill = TRUE, fill.alpha = 0.1)
```

Figure 6: Heterogenous Regressions

Fit both models:

```
rohwer.ses1 <- lm(cbind(SAT, PPVT, Raven) ~ n + s + ns + na + ss,
                  data = Rohwer, subset = SES == "Hi")
rohwer.ses2 <- lm(cbind(SAT, PPVT, Raven) ~ n + s + ns + na + ss,
                  data = Rohwer, subset = SES == "Lo")
```

Low SES students:

```
heplot(rohwer.ses2, col = c("red", rep("black",5), "blue"),
     hypotheses = list("B=0, Low SES" = c("n", "s", "ns", "na", "ss")),
       level = 0.5, cex = 1.25,
     fill = c(TRUE, FALSE), fill.alpha = 0.05,
     xlim = c(-15, 110), ylim = c(40,110),
     xlab = "Student Achievement Test",
     ylab = "Peabody Picture Vocabulary Test",
     label.pos = c(1, rep(NULL, 5), 1))
```

High SES students:

```
heplot(rohwer.ses1, col = c("red", rep("black", 5), "blue"),
      hypotheses = list("B=0, High SES" = c("n", "s", "ns", "na", "ss")),
      level = 0.5, cex = 1.25,
      add = TRUE,  # place both plots on same graphic
      error = TRUE, # error ellipse is not drawn by default with add = TRUE
      fill = c(TRUE, FALSE), fill.alpha = 0.05,
      xlim = c(-15, 110), ylim = c(40,110))
```