Visualizing Tests for Equality of Covariance Matrices

Supplemental Appendix

Michael Friendly and Matthew Sigal

September 18, 2017

Contents

Introduction

This is the supplemental appendix to “Visualizing Tests for Equality of Covariance Matrices,” submitted to *The American Statistician.* It covers topics of interest that were considered too long or not sufficiently essential to include in the paper. Section 1 gives a brief, capsule summary of the HE plot framework for visualizing multivariate tests of means using Hypothesis-Error plots and related canonical discriminant plots. An example is presented using the `Wine` data. Section 2 presents a different approch to questions of multivariate homogeneity of variance using analyses of distances from a centroid to measure multivariate dispersion.

1 Visualizing mean differences: The HE plot framework

For univariate response models, there is a variety of effective graphical methods for visualizing differences among group means, ranging from simple dot charts (with error bars), to boxplots that provide more details, to effect plots (Fox, 2003) for complex designs.

In the MANOVA setting, a collection of such univariate plots for each of the responses can be useful, but they do not show how the responses vary *jointly*. In this section we introduce some *multivariate* graphical methods (HE plots and canonical views) for the comparison of means (for further details, see Fox, Friendly, & Monette, 2009; Friendly, 2007; and, Friendly & Sigal, 2014) that can also be usefully applied to the comparison of covariance matrices.

The essence of these methods is the Hypothesis–Error (HE) plot, and we call the collection of their applications the HE plot framework. The main ideas underlying this framework are:

(a) All hypothesis tests of model terms or linear hypotheses in the MLM correspond to statistics based on the eigenvalues λ_i of the sum of squares and products (SSP) matrix for the hypothesis (the $\boldsymbol{H}$ matrix) relative to the SSP matrix for error (the $\boldsymbol{E}$ matrix), i.e., the $s = \min(p, df_h)$ non-zero values in the eigenvalue decomposition, $\boldsymbol{H}\boldsymbol{E}^{-1}\boldsymbol{v}_i = \lambda_i \boldsymbol{v}_i$. They answer the question, "How big is $\boldsymbol{H}$ relative to $\boldsymbol{E}$?" and are analogous to univariate F statistics.

(b) The $\boldsymbol{H}$ matrix for any term t is essentially the $p \times p$ SSP matrix of the *fitted* values, $\widehat{\boldsymbol{Y}_t}^{\mathsf{T}}\widehat{\boldsymbol{Y}_t}$, for that term.[1] The $\boldsymbol{E}$ matrix is the SSP matrix of the *residuals* for the full model.

(c) The $\boldsymbol{H}$ and $\boldsymbol{E}$ matrices can therefore be visualized as the data ellipsoids of the fitted values and residuals. In the `heplot()` function, $\boldsymbol{H}$ and $\boldsymbol{E}$ can be scaled to show the size of each effect; alternatively, the $\boldsymbol{H}$ matrix can be scaled to provide a visual test of significance, in the sense that the $\boldsymbol{H}$ ellipsoid projects outside the $\boldsymbol{E}$ ellipsoid *iff* the effect is significant by Roy's maximum root test.

(d) For MANOVA designs, when there are more than a few response variables, all these effects may be readily viewed in the orthonormalized *canonical* space corresponding to the linear transformation $\boldsymbol{Y}\boldsymbol{T} \mapsto \boldsymbol{Y}^{\star} = \boldsymbol{Y}\boldsymbol{E}^{-1/2}\boldsymbol{V}$. This is the view that shows the maximal differences among means.

Overall, this framework is incredibly versatile and can be utilized to visualize and summarize the important aspects from a plethora of designs.

[1] This uses so-called "Type II" sums of squares, calculated according to the principle of marginality, testing each term after all others, except ignoring the term's higher-order relatives.

1.1 HE plot example: Wine data

This example illustrates the basic properties of Hypothesis-Error plots, using the Wine data analyzed in Section 4 of the main paper. Recall that this data set consists of 13 chemical properties of wine samples derived from three different cultivars of grapes grown in a region of Italy. The question is *how* do the cultivars differ in their means on these properties.

The standard MANOVA of differences in means is conducted as follows and shows a highly significant overall multivariate test of differences among regions on the response variables jointly.

```
Wine.mod <- lm(as.matrix(Wine[, -1]) ~ Cultivar, data=Wine)
Anova(Wine.mod, test="Roy")
```

```
Type II MANOVA Tests: Roy test statistic
         Df test stat approx F num Df den Df    Pr(>F)
Cultivar  2    9.0817   114.57     13    164 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Follow-up univariate F tests for each of the responses show that the grape varieties differ substantially on all of these measures. However, we want to understand the *pattern* of these differences, individually and jointly. The boxplots of the data shown in Fig. 7 of the paper show the univariate differences, but don't help to understand the correlations of these differences.

```
pairs(Wine.mod, variables=1:5, fill=TRUE, fill.alpha=.1)
```

An HE plot for the first five response variables is shown in Figure 1 here as a scatterplot matrix for all pairs. In this plot, in each panel the co-variation of the means (the $\boldsymbol{H}$ matrix) is shown by the blue ellipse. The eccentricity of the $\boldsymbol{H}$ ellipse indicates the correlation of the means. For example, means on Alcohol are positively correlated with those on Ash and Mg. The red ellipse for the $\boldsymbol{E}$ matrix shows the size and correlation of the residuals. As noted above, this plot has the property that the overall multivariate test of `Cultivar` differences is significant (by Roy's test) if the $\boldsymbol{H}$ ellipse projects anywhere outside the $\boldsymbol{E}$ ellipse. It can be seen that the group means are well-separated on all pairs of variables, but the pattern of these is somewhat complex, even for just five of the responses.

1.2 Canonical views

There are too many response variables to see them all at once in variable space, but with $g = 3$ groups, the MANOVA has a very simple 2D representation in the canonical space.

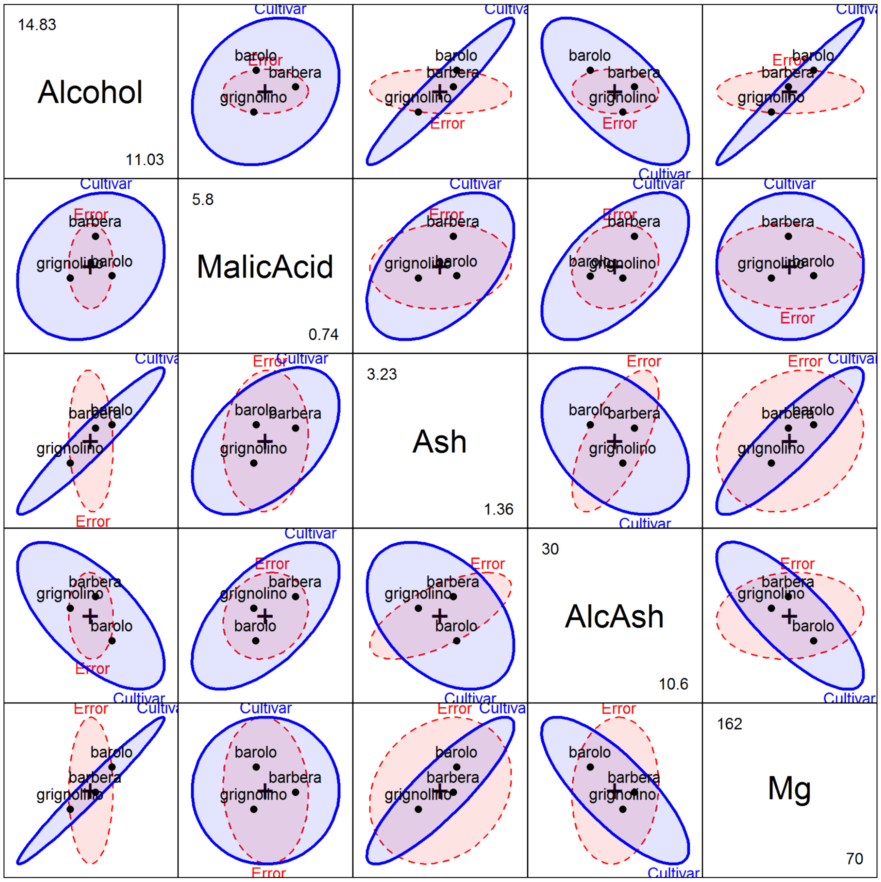

Figure 1: Pairwise HE plot for the first five response variables in the Wine data. The group means for each pair are labeled.

This is effectively a projection of the data into the space of the linear combinations of the responses that accounts for the greatest differences among the group means.

```r
Wine.can <- candisc(Wine.mod)
Wine.can
```

```
Canonical Discriminant Analysis for Cultivar:

   CanRsq Eigenvalue Difference Percent Cumulative
1 0.90081     9.0817     4.9533  68.748     68.748
2 0.80501     4.1285     4.9533  31.252    100.000

Test of H0: The canonical correlations in the
current row and all that follow are zero

  LR test stat approx F numDF denDF   Pr(> F)
```

```
1     0.019341   77.620    26   326 < 2.2e-16 ***
2     0.194990   56.422    12   164 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The two canonical correlations are both large and highly significant. A 2D canonical plot (Figure 2) shows the canonical scores together with data ellipses for the scores and vectors for the structure coefficients. The second canonical dimension separates Grinolino wines from the other two regions. The first canonical dimension accounts for 69% of the group mean differences, and the groups are very well separated on this dimension. Longer vectors correspond to response variables that are better discriminated among the groups on this dimension. A guess at the interpretation of this dimension is that it contrasts those wines high in Malic acid and non-flavor phenols vs. those with high phenols and OD ratios.

```
plot(Wine.can, ellipse=TRUE, var.lwd=2)
```

We think of such canonical plots as a multivariate juicer for MANOVA problems: they show the data in the low-D view that extracts the most flavor regarding group differences, and also show the relations of the response variables to this space.

As well, note that in this canonical view, the data ellipses for the canonical scores for the groups have apparently different shapes, which speaks to the question of heterogeneity of covariance matrices. This is explored in the Section 5 of the paper in relation to our multivariate generalization of Levene-type tests.

2 Distance based approaches

If we think of a covariance matrix as representing multivariate *dispersion* of points around a multivariate *centroid*, then a reasonable alternative to methods based on the covariance matrices themselves is an approach based on distances. A set of multivariate samples, $\boldsymbol{y}_{ij}, i = 1, \ldots, g; j = 1, \ldots n_i$ can be said to be equally dispersed if an "average" distance, $\Delta(\boldsymbol{y}_{ij}, \widetilde{\boldsymbol{y}}_i)$ around a group "centroid" $\widetilde{\boldsymbol{y}}_i$ does not differ substantially across the groups.

Equality of multivariate dispersion is not exactly the same as equality of covariance matrices (it doesn't account for orientation), but it has the advantage of transforming the p-dimensional problem into one based on an ANOVA of distances, in a way similar to what Levene-type tests do for variances.

This idea admits a variety of different forms:

- **distances** can be defined as standard Euclidean distances $D(\boldsymbol{y}) = [(\boldsymbol{y} - \bar{\boldsymbol{y}})^\mathsf{T} (\boldsymbol{y} - \bar{\boldsymbol{y}})]^{1/2}$ in an orthogonal $\mathbb{R}^p$ variable space, Mahalanobis distances, $D_M(\boldsymbol{y}) = [(\boldsymbol{y} - \bar{\boldsymbol{y}})^\mathsf{T} \boldsymbol{S}^{-1} (\boldsymbol{y} - \bar{\boldsymbol{y}})]^{1/2}$, which take the variances and covariances in $\boldsymbol{S}$ into account, or a wide variety of

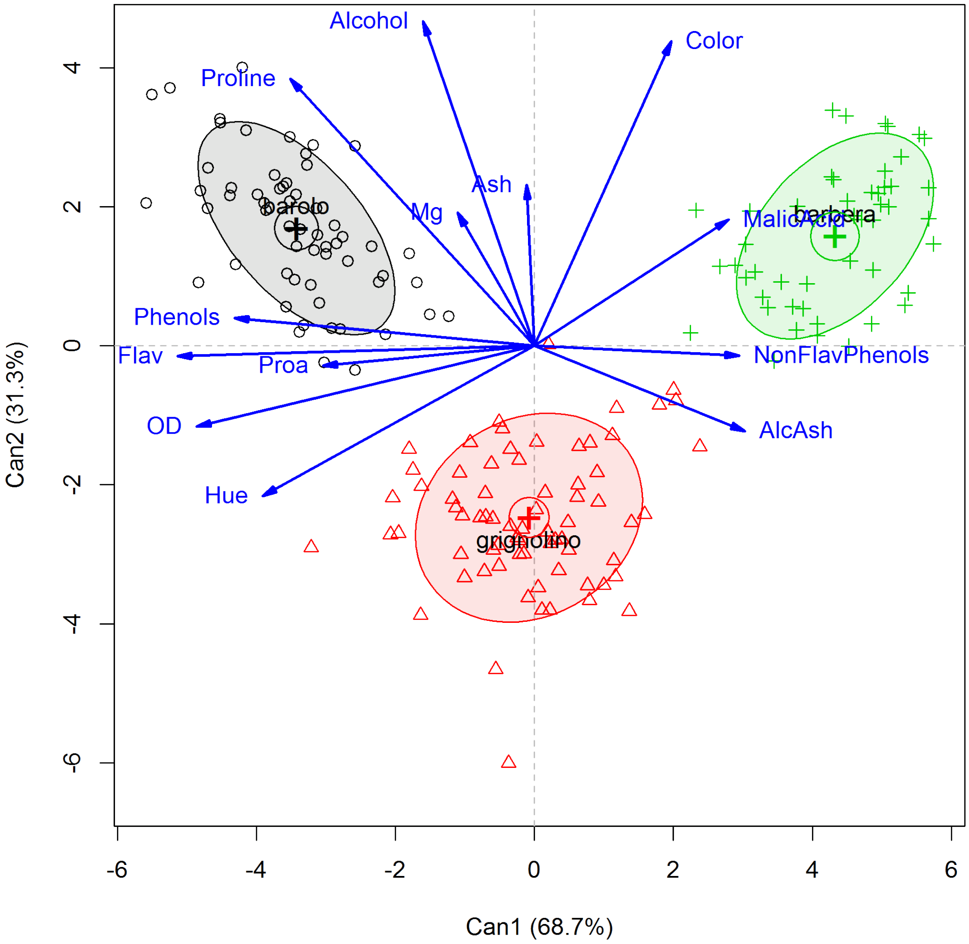

Figure 2: 2D canonical discriminant plot for the Wine data. Variable vectors show the structure coefficients (correlations) of the response variables with the canonical dimensions.

other distance measures (e.g., Manhattan, L_1 or taxi-cab distance, based on the sum of absolute differences).

- **centroids** can be defined as multivariate means, medians, trimmed means, "spatial medians", and so forth.
- **average** can be defined as the arithmetic average, a root-mean-square or robust mean.
- **tests** for differences in dispersion can be based on a standard ANOVA, or a non-parametric permutation test.

Several versions of this distance based approach to multivariate dispersion were suggested by Anderson (2006; extending Van Valen, 1978). Her methods (designed for biological/ecological applications) take this idea a bit further, in that it starts with the $N \times N$ pairwise distance measures among all observations $\boldsymbol{y}_{ij}$, carries out a metric multidimensional scaling (MDS, equivalent to a PCA of the doubly-centered distance matrix), and then applies the distance to centroid calculations to the principal coordinates. This is largely to accommodate a variety of non-Euclidean dissimilarity measures of interest in ecological studies.

These methods have been implemented in the `betadisper()` function in the `vegan` package (Oksanen et al., 2016). The F-test provided by the `anova()` method is analogous to a Levene-type test based on average distance from the group centroids. The package also provides a permutation test, `permutest()` that assesses the p-value of the F-statistic relative to the permutation distribution under random assignment to groups. The examples below are merely illustrative of this approach and the graphical methods they provide.

2.1 Example: Iris data

It is easiest to illustrate this approach using the Iris data. Here, we calculate Euclidean distances among all observations and use `betadisper()` to carry out analysis of these distances by Species.

```
library(vegan)
dst <- dist(iris[,1:4])
iris.bd <- betadisper(dst, iris$Species)
```

Both the `anova()` and `permutest()` methods indicate that there are significant differences in the dispersion by Species.

```
anova(iris.bd)
```

```
Analysis of Variance Table

Response: Distances
           Df  Sum Sq Mean Sq F value  Pr(>F)
Groups      2  2.9092 1.45458  10.748 4.4e-05 ***
Residuals 147 19.8941 0.13533
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

It is easy to see this in the principal coordinates plot of the first two MDS dimensions. The plot method (Figure 3) highlights the distances of each observation to the group centroid and also draws a 68% data ellipse. The configuration of the group means is similar to that of the PCA (Fig. 3 of the paper, left panel), however the orientation of the within-group data ellipses are in the opposite direction here. (The orientation of MDS axes is arbitrary.)

```
labs <- paste("Dimension", 1:4, "(",
              round(100*iris.bd$eig / sum(iris.bd$eig), 2), "%)")
```

```
plot(iris.bd, cex=1, pch=15:17,
     main="Iris data: MDS coordinates", cex.lab=1.25,
     xlab=labs[1], ylab=labs[2],
     hull=FALSE, ellipse=TRUE, conf=0.68, lwd=2)
```

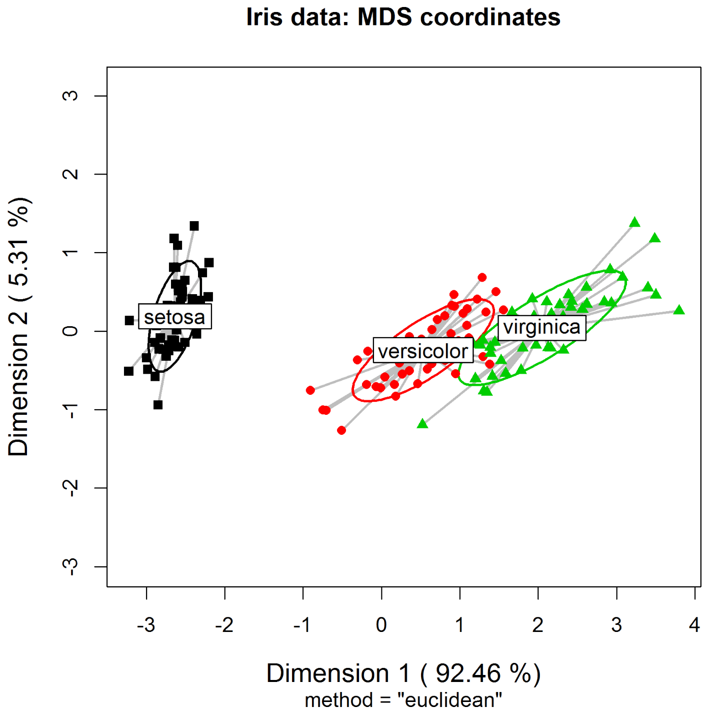

Figure 3: Principal coordinates plot of the MDS solution for the Iris data. Ellipses are 68% data ellipses for the coordinate scores.

The boxplot method (Figure 4) applied to the distances confirms what we can see from the ANOVA tests and the plot in principal coordinates space: There are substantial differences in dispersion among the species, particularly for *Setosa* versus the other two species, which do not differ from each other.

```
boxplot(iris.bd, xlab="Species", notch=TRUE, col=c("gray", "red", "green"))
```

2.2 Example: Skulls data

For the Skulls data, we saw earlier that there were no differences among covariance matrices in any plots or analyses. Here, we calculate Euclidean distances among all observations and use betadisper() to carry out analysis of these distances by Epoch.

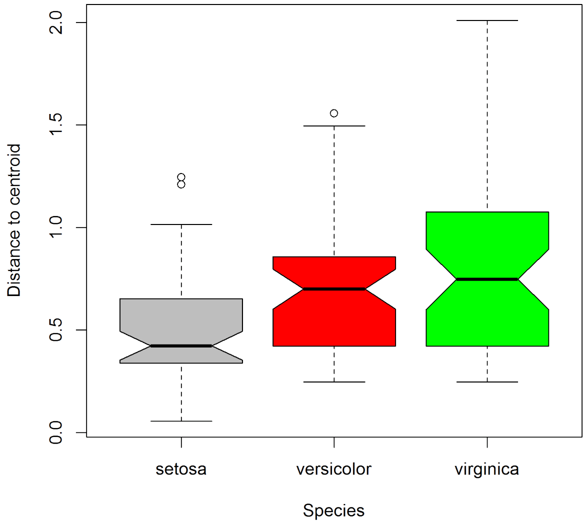

Figure 4: Boxplot of distances to the group centroid for the Iris data. Notches indicate 95% confidence intervals for group medians.

```
dst <- dist(Skulls[, -1])
skulls.bd <- betadisper(dst, Skulls$epoch)
```

Both the `anova()` and `permutest()` method indicate that there are non-significant differences in the dispersion by Epoch.

```
anova(skulls.bd)
```

```
Analysis of Variance Table

Response: Distances
           Df  Sum Sq Mean Sq F value Pr(>F)
Groups      4   35.43  8.8565  0.8211 0.5137
Residuals 145 1563.96 10.7859
```

Then, a simple display of the differences among groups in their dispersion is a boxplot of the distances to the centroid (Figure 5). The pattern of the medians seen here is similar to that shown in the plot for Box's M test in Fig. 8 (right).

```
boxplot(skulls.bd, notch=TRUE, col="lightblue", xlab="Epoch")
```

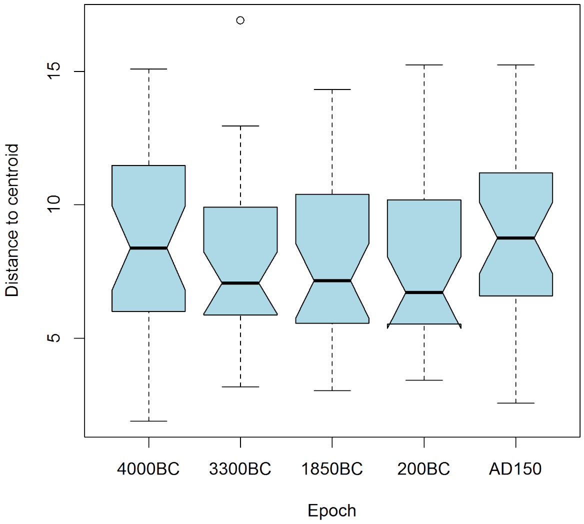

Figure 5: Boxplot of distances to the group centroid for the Skulls data

References

Anderson, M. J. (2006). Distance-based tests for homogeneity of multivariate dispersions. *Biometrics*, *62*(1), 245–253. http://doi.org/10.1111/j.1541-0420.2005.00440.x

Fox, J. (2003). Effect displays in R for generalized linear models. *Journal of Statistical Software*, *8*(15), 1–27.

Fox, J., Friendly, M., & Monette, G. (2009). Visualizing hypothesis tests in multivariate linear models: The *heplots* package for R. *Computational Statistics*, *24*(2), 233–246. Retrieved from http://dx.doi.org/10.1007/s00180-008-0120-1

Friendly, M. (2007). HE plots for multivariate general linear models. *Journal of Computational and Graphical Statistics*, *16*(2), 421–444. http://doi.org/10.1198/106186007X208407

Friendly, M., & Sigal, M. (2014). Recent advances in visualizing multivariate linear models. *Revista Colombiana de Estadistica*, *37*(2), 261–283. http://doi.org/http://dx.doi.org/10.15446/rce.v37n2spe.47934

Oksanen, J., Blanchet, F. G., Friendly, M., Kindt, R., Legendre, P., McGlinn, D., … Wagner, H. (2016). *Vegan: Community ecology package*. Retrieved from https://CRAN.R-project.org/package=vegan

Van Valen, L. (1978). The statistics of variation. *Evolutionary Theory*, *4*, 33–43.