

Visualizing Tests for Equality of Covariance Matrices

Michael Friendly and Matthew Sigal

Abstract

This paper explores a variety of topics related to the question of testing the equality of covariance matrices in multivariate linear models, particularly in the MANOVA setting. The main focus is on graphical methods that can be used to address the evaluation of this assumption. We introduce some extensions of data ellipsoids, hypothesis-error (HE) plots and canonical discriminant plots and demonstrate how they can be applied to the testing of equality of covariance matrices. Further, a simple plot of the components of Box's M test is proposed that shows *how* groups differ in covariance and also suggests other visualizations and alternative test statistics. A multivariate extension of Levene's test of homogeneity of variance is also discussed, which leads to other novel visualizations for equality of covariance matrices within the HE plot framework. These methods are implemented and freely available in the **heplots** and **candisc** packages for R. Examples from the paper are available in supplementary materials.

Keywords— Box M test; HE plots; MANOVA; graphics

1 Introduction

To make the preliminary test on variances is rather like putting to sea in a rowing boat to find out whether conditions are sufficiently calm for an ocean liner to leave port. — G. E. P. Box (1953)

This paper concerns the extension of tests of homogeneity of variance from the classical univariate ANOVA setting to the analogous multivariate (MANOVA) setting. Such tests are a routine but important aspect of data analysis, as particular violations can drastically impact model estimates (Lix & Keselman, 1996). In the multivariate context, the following questions and topics are of main interest here:

- **Visualization:** How can we visualize differences among group variances and covariance matrices, perhaps in a way that is analogous to what is done to visualize differences among group means? Multivariate linear models (MLMs) present additional challenges for data visualization because we often want to see the effects for a collection of response variables *simultaneously*, which pushes the boundaries of typical graphical displays. As will be illustrated, differences among covariance matrices can be comprised of spread in overall size (“scatter”) and shape (“orientation”). When there are more than a few response variables, what low-dimensional views can show the most interesting properties related to the equality of covariance matrices?
- **Other test statistics:** Test statistics for MANOVA and for equality of covariance matrices are based on properties of eigenvalues of various matrices. Available tests statistics for mean differences suggest alternatives for the question of equality of covariance matrices.

The following subsections provide a capsule summary of the issues in this topic. Most of the discussion is couched in terms of a one-way design for simplicity, but the same ideas can apply to two-way (and higher) designs, where a “group” factor is defined as the product combination (interaction) of two or more factor variables. When there are also numeric covariates, this topic can also be extended to the multivariate analysis of covariance (MANCOVA) setting. This can be accomplished by applying these techniques to the residuals from predictions by the covariates alone.

1.1 Homogeneity of Variance in Univariate ANOVA

In classical (Gaussian) univariate ANOVA models, the main interest is typically on tests of mean differences in a response y according to one or more factors. The validity of the typical F test, however, relies on the assumption of *homogeneity of variance*: all groups have the same (or similar) variance,

$$\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_g^2.$$

It turns out that the F test for differences in means is relatively robust to violation of this assumption (Harwell, Rubinstein, Hayes, & Olds, 1992), as long as the group sizes are roughly equal.¹

A variety of classical test statistics for homogeneity of variance are available, including Hartley’s F_{max} (Hartley, 1950), Cochran’s C (Cochran, 1941), and Bartlett’s test (Bartlett, 1937), but these have been found to have terrible statistical properties (Rogan & Keselman, 1977), which prompted Box’s famous quote.

Levene (1960) introduced a different form of test, based on the simple idea that when variances are equal across groups, the average absolute values of differences between the observations and group means will also be equal, i.e., substituting an L_1 norm for the L_2 norm of variance. In a one-way design, this is equivalent to a test of group differences in the means of the auxiliary variable $z_{ij} = |y_{ij} - \bar{y}_i|$.

More robust versions of this test were proposed by Brown & Forsythe (1974). These tests substitute the group mean by either the group median or a trimmed mean in the ANOVA of the absolute deviations, and should be almost always preferred to Levene’s version (which unfortunately was adopted as the default in some software such as SPSS). See Conover, Johnson, & Johnson (1981) for an early review and Gastwirth, Gel, & Miao (2009) for a general discussion of these tests. In what follows, we refer to this class of tests as “Levene-type” tests and suggest a multivariate extension described below (Section 5).

1.2 Homogeneity of variance in MANOVA

In the MANOVA context, the main emphasis, of course, is on differences among mean vectors, testing

$$H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \cdots = \boldsymbol{\mu}_g .$$

However, the standard test statistics (Wilks’ Lambda, Hotelling-Lawley trace, Pillai-Bartlett trace, Roy’s maximum root) rely upon the analogous assumption that the within-group covariance matrices for all groups are equal,

$$\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \cdots = \boldsymbol{\Sigma}_g .$$

In the multivariate setting, there has been considerable attention to the sensitivity of these tests to both non-normality and lack of equality of covariance matrices, largely through simulation studies (e.g., Finch & French, 2013; Hakstian, Roed, & Lind, 1979). Most of these have been conducted in the simple case of two-group designs (where Hotelling’s T^2 is the

¹If group sizes are greatly unequal **and** homogeneity of variance is violated, then the F statistic is too liberal (p values too large) when large sample variances are associated with small group sizes. Conversely, the F statistic is too conservative if large variances are associated with large group sizes.

equivalent of all the standard tests) or in one-way designs. A classic study in this area is Olson (1974), that recommended:

for protection against nonnormality and heterogeneity of covariance matrices, the largest-root test should be avoided, while the Pillai-Bartlett trace test may be recommended as the most robust of the MANOVA tests, with adequate power to detect true differences in a variety of situations (p. 894).

We mention in passing that, with a burgeoning interest in robust methods over the last few decades, there have been a variety of proposals for how to conduct robust tests of differences on mean vectors, mostly in the one-way MANOVA setting (e.g., Aelst & Willems, 2011; Todorov & Filzmoser, 2010). Generally speaking, these involve using more robust alternatives for mean vectors (medians, trimmed means, rank-based methods) and for covariance matrices (e.g., minimum covariance determinant (MCD) and minimum volume ellipsoid (MVE)).

Yet, there has not been as much attention paid to the second-order problem of assessing equality of covariance matrices. Box’s M test, described below, remains the main procedure readily available in statistical software for this problem. The properties and alternatives to Box’s test have not been widely studied (some exceptions are O’Brien, 1992, and @TikuBalakrishnan:1984).

However, beyond issues of robustness, the question of equality of covariance matrices is often of general interest itself. For instance, variability is often an important issue in studies of strict equivalence in laboratories comparing across multiple patient measurements and in other applied contexts (see Gastwirth et al., 2009 for other exemplars). Moreover the outcome of such tests often have important consequences for the details of a main method of analysis. Just as the Welsh t -test (Welch, 1947) is now commonly used and reported for a two-group test of differences in means under unequal variances, a preliminary test of equality of covariance matrices is often used in discriminant analysis to decide whether linear (LDA) or quadratic discriminant analysis (QDA) should be applied in a given problem. In such cases, the data at hand should inform the choice of statistical analysis to utilize.

1.3 Assessing heterogeneity of covariance matrices: Box’s M test

Box (1949) proposed the following likelihood-ratio test (LRT) statistic for testing the hypothesis of equal covariance matrices,

$$M = (N - g) \ln |\mathbf{S}_p| - \sum_{i=1}^g (n_i - 1) \ln |\mathbf{S}_i|, \quad (1)$$

where $N = \sum n_i$ is the total sample size and $\mathbf{S}_p = (N - g)^{-1} \sum_{i=1}^g (n_i - 1) \mathbf{S}_i$ is the pooled covariance matrix. M can thus be thought of as a ratio of the determinant of the pooled \mathbf{S}_p

to the geometric mean of the determinants of the separate \mathbf{S}_i .

In practice, there are various transformations of the value of M to yield a test statistic with an approximately known distribution (Timm, 1975). Roughly speaking, when each $n_i > 20$, a χ^2 approximation is often used; otherwise an F approximation is known to be more accurate.

Asymptotically, $-2 \ln(M)$ has a χ^2 distribution. The χ^2 approximation due to Box (1949, 1950) is that

$$X^2 = -2(1 - c_1) \ln(M) \sim \chi_{df}^2$$

with $df = (g - 1)p(p + 1)/2$ degrees of freedom, and a bias correction constant:

$$c_1 = \left(\sum_i \frac{1}{n_i - 1} - \frac{1}{N - g} \right) \frac{2p^2 + 3p - 1}{6(p + 1)(g - 1)}.$$

In this form, Bartlett’s test for equality of variances in the univariate case is the special case when there is only one response variable, so Bartlett’s test is sometimes used as univariate follow-up to determine which response variables show heterogeneity of variance.

Yet, like its univariate counterpart, Box’s test is well-known to be highly sensitive to violation of (multivariate) normality and the presence of outliers. For example, Tiku & Balakrishnan (1984) concluded from simulation studies that the normal-theory LRT provides poor control of Type I error under even modest departures from normality. O’Brien (1992) proposed some robust alternatives, and showed that Box’s normal theory approximation suffered both in controlling the null size of the test and in power. Zhang & Boos (1992) also carried out simulation studies with similar conclusions and used bootstrap methods to obtain corrected critical values.

1.4 Visualizing heterogeneity

The goal of this article is to use the above background as a platform for discussing approaches to visualizing and testing the heterogeneity of covariance matrices in multivariate designs. While researchers often rely on a single number to determine if their data have met a particular threshold, such compression will often obscure interesting information, particularly when a test concludes that differences exist, and one is left to wonder “why?”. It is within this context where, again, visualizations often reign supreme. In fact, we find it somewhat surprising that this issue has not been addressed before graphically in any systematic way.

In this article, we propose three visualization-based approaches to questions of heterogeneity of covariance in MANOVA designs: (a) direct visualization of the information in the \mathbf{S}_i and \mathbf{S}_p using *data ellipsoids* to show size and shape as minimal schematic summaries; (b) a simple dotplot of the components of Box’s M test: the log determinants of the \mathbf{S}_i together with that of the pooled \mathbf{S}_p . Extensions of these simple plots raise the question of whether

measures of heterogeneity other than that captured in Box’s test might also be useful; and, (c) the connection between Levene-type tests and an ANOVA (of centered absolute differences) suggests a parallel with a multivariate extension of Levene-type tests and a MANOVA. We explore this with a version of Hypothesis-Error (HE) plots we have found useful for visualizing mean differences in MANOVA designs.

Accordingly, the following sections introduce and apply our conceptual framework for general graphical methods for visualizing data in relation to MLM-related questions and their applications. This is based on the simple ideas that: (a) a *data ellipsoid* provides a visual summary of location and scatter of a multivariate sample; (b) these can be combined in various ways to give visual tests of group differences in means and covariance matrices; and, (c) when there are more than just a few response variables, a reduced-rank (canonical) transformation provides an appealing way to visualize these effects in an optimal low-dimensional approximation.

Section 4 introduces some novel visualizations of the components related to Box’s test, which in turn suggest other possible test statistics that deserve further study. Section 5 describes a multivariate generalization of Levene’s test within the HE plot framework that yields attractive and useful displays.

A different graphical approach to the main question is to consider multivariate dispersion in terms of distances of the points from their centroids; this is illustrated in the Supplementary Materials. These methods are all implemented in R (R Core Team, 2015), principally in the `heplots` and `candisc` packages.²

2 Visualizing covariance matrices

Before diving into details and statistical tests, it is useful to see how to visualize covariance matrices themselves. We do this using the graphical analog of minimally sufficient statistics $(\bar{\mathbf{y}}_i, \mathbf{S}_i)$ for the MANOVA problem— a *minimally sufficient graphical display*. This graphical principle has been called *visual thinning* (Friendly, 2007): reducing a graphical display to the essentials of what you want to see by relying upon statistics that most efficiently capture the parameters of interest. In multivariate displays, this usually means replacing data points by well-chosen visual summaries.

²The complete R code for our examples is provided in the Supplementary Materials, and are hosted online at https://mattsigal.github.io/eqcov_supp/

2.1 Data ellipsoids

The essential idea (Dempster, 1969; Friendly, Monette, & Fox, 2013) is that for a p -dimensional sample, $\mathbf{Y}_{n \times p}$, the $p \times p$ covariance matrix \mathbf{S} can be represented by the p -dimensional *concentration* or *data ellipsoid*, \mathcal{E}_c of size (“radius”) c . This is defined as the set of all points \mathbf{y} satisfying

$$\mathcal{E}_c(\bar{\mathbf{y}}, \mathbf{S}) := \{\mathbf{y} : (\mathbf{y} - \bar{\mathbf{y}})^\top \mathbf{S}^{-1} (\mathbf{y} - \bar{\mathbf{y}}) \leq c^2\}. \quad (2)$$

It is readily seen that the quadratic form in Eqn. (2) corresponds to the set of points whose squared Mahalanobis distances $D_M^2(\mathbf{y}) = (\mathbf{y} - \bar{\mathbf{y}})^\top \mathbf{S}^{-1} (\mathbf{y} - \bar{\mathbf{y}})$, from the centroid of the sample, $\bar{\mathbf{y}} = (\bar{y}_1, \bar{y}_2, \dots, \bar{y}_p)^\top$, are less than or equal to c^2 .

When the variables are multivariate normal, the data ellipsoid approximates a contour of constant density in their joint distribution. In this case $D_M^2(\mathbf{y})$ has a large-sample χ_p^2 distribution, or, in finite samples, approximately $[p(n-1)/(n-p)]F_{p, n-p}$. Hence, in the bivariate case, taking $c^2 = \chi_2^2(0.95) = 5.99 \approx 6$ encloses approximately 95% of the data points under normal theory. A 68% coverage data ellipse with $c^2 = \chi_2^2(0.68) = 2.28$ gives a bivariate analog of the standard $\bar{x} \pm 1s_x$ and $\bar{y} \pm 1s_y$ intervals. See Friendly et al. (2013) for properties of data ellipsoids and their use to interpret a wide variety of problems and applications in multivariate linear models.

In practice, p -dimensional data ellipsoids can be viewed in variable space via 2D or 3D projections, or for all p variables, in a pairwise scatterplot matrix of 2D projections. Alternatively, they can be viewed in the space of any linear transformation $\mathbf{Y}\mathbf{T} \mapsto \mathbf{Y}^*$, where the principal components transformation provides useful views in low-D projections accounting for maximal total variance.

2.2 Simple example: Iris data

It is easiest to illustrate these ideas using the well-known Iris data set (Anderson, 1935), which pertains to four measures (sepal width and height, and petal width and height) of three species of iris flowers from the Gaspé Peninsula. One approach to visualizing within group variability is to begin with an enhanced scatterplot that adds a standard (68%) data ellipse for each group. Then imagine taking away the data points (and other enhancements) leaving only the data ellipses, and add the corresponding data ellipse for the pooled sample variance covariance matrix \mathbf{S}_p . This gives a visual summary of group means and of the within-group covariance, and is shown in the right panel of Figure 1. In this plot the variances and covariances look similar for the Versicolor and Virginca groups, but the Setosa group differs by exhibiting a higher correlation between sepal length and width and a smaller variance on sepal length.

Finally, we center all the ellipses at the origin in order to focus *only* on size and shape of the within-group covariances, so that these can be *directly* compared visually.³ For these two variables, we can now see that the covariance of *Virginica* is nearly identical to S_p , while *Versicolor* has somewhat greater variance on sepal length.⁴

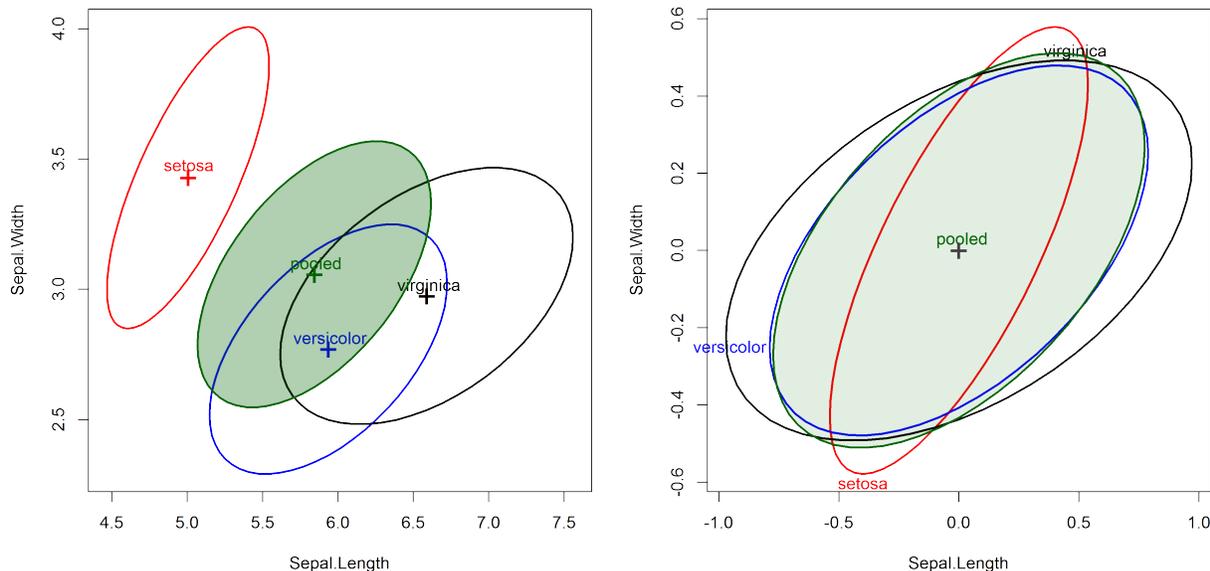


Figure 1: Covariance ellipses for the Iris data. Left: separate groups and the pooled within-group covariance matrix; right: all covariance matrices centered at the origin.

This method becomes particularly useful when we look at the data ellipses for all pairs of variables in scatterplot matrix format. As in the right panel of Figure 1, we center these ellipsoids at the origin. The display in Figure 2 shows only size (variance) and shape (correlation) differences, which speak directly to the question of homogeneity of covariance matrices.

It can now be seen precisely *how* the covariance matrix for Setosa differs from those of the other species. The within group correlations differ for all pairs of variables, and as well, the variances are noticeably smaller for petal width and petal length. In addition, while *Versicolor* and *Virginica* have similar shapes, close to that of the pooled covariance matrix, in most panels (particularly for petal width), *Virginica* exhibits greater variance.

³This example seems at first glance to be a special case, because all variables are measured in the same units. However, the units do not matter in most of our plots because the axis ranges are taken from the data and scale units are not equated. The plots in Figure 1 would look identical except for tick labels if we transformed sepal length from centimeters to inches.

⁴Such plots are produced by the `covEllipses()` function in the `heplots` package.

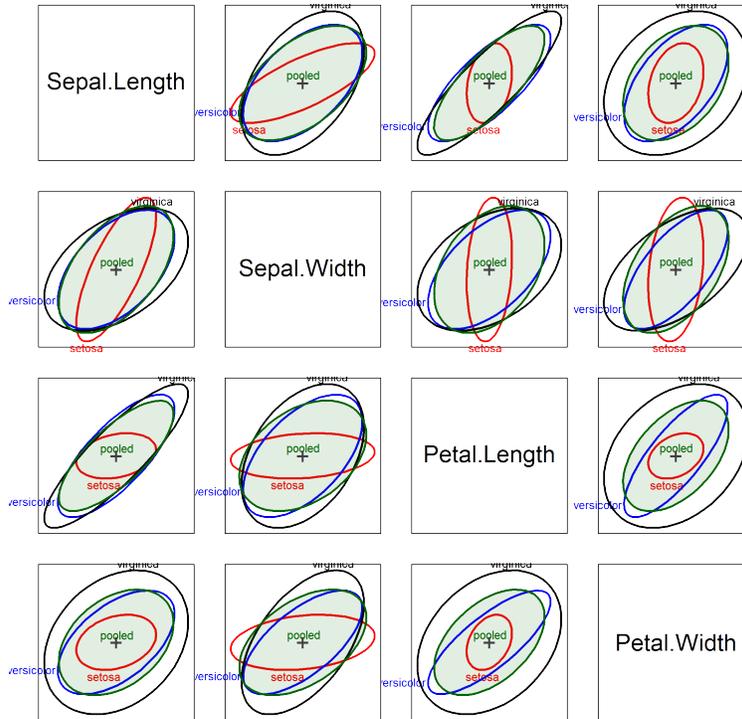


Figure 2: Pairwise data ellipses for the Iris data, centered at the origin. This view makes it easy to compare the variances and covariances for all pairs of variables.

2.2.1 More general models

In these plots, the centered views correspond to an analysis of the covariance matrices among the *residuals* from the MLM predicting the four responses from the species variable. Consequently, the same ideas apply in more general models. For example, in a MANCOVA setting, the model may include one or more quantitative covariates.⁵ The analyses suggested above could then be applied to the residuals from this model. Likewise, in a two-way MANOVA design, with factors A and B , we could treat the combinations of these factors as the “group” variable and view the pairwise data ellipses.⁶

2.3 Low-rank views

With $p > 3$ response variables, a simple alternative to the pairwise 2D projections shown in Figure 2 is the projection into the principal component space accounting for the greatest

⁵For instance, in R notation, `mod1 <- lm(cbind(y1, y2, y3) ~ Group + x1 + x2)`, where three response variables are being predicted by the grouping factor and two covariates.

⁶For example, we could estimate such a model within R using `mod2 <- lm(cbind(y1, y2, y3) ~ A:B)` and then generate the pairwise covariance data ellipses with `covEllipses(residuals(mod2), variables=1:3)`.

amounts of total variance in the data. For the Iris data, a simple PCA of the covariance matrix shows that nearly 98% of total variance in the data is accounted for in the first two dimensions.

Figure 3 shows the plots of the covariance ellipsoids for the first two principal component scores, uncentered (left panel) and centered (right panel). The dominant PC1 (92% of total variance) essentially orders the species by a measure of overall size of their sepals and petals. In the centered view, it can again be seen how *Setosa* differs in covariance from the other two species, and that while *Virginica* and *Versicolor* both have similar shapes to the pooled covariance matrix, *Versicolor* has somewhat greater variance on PC1.

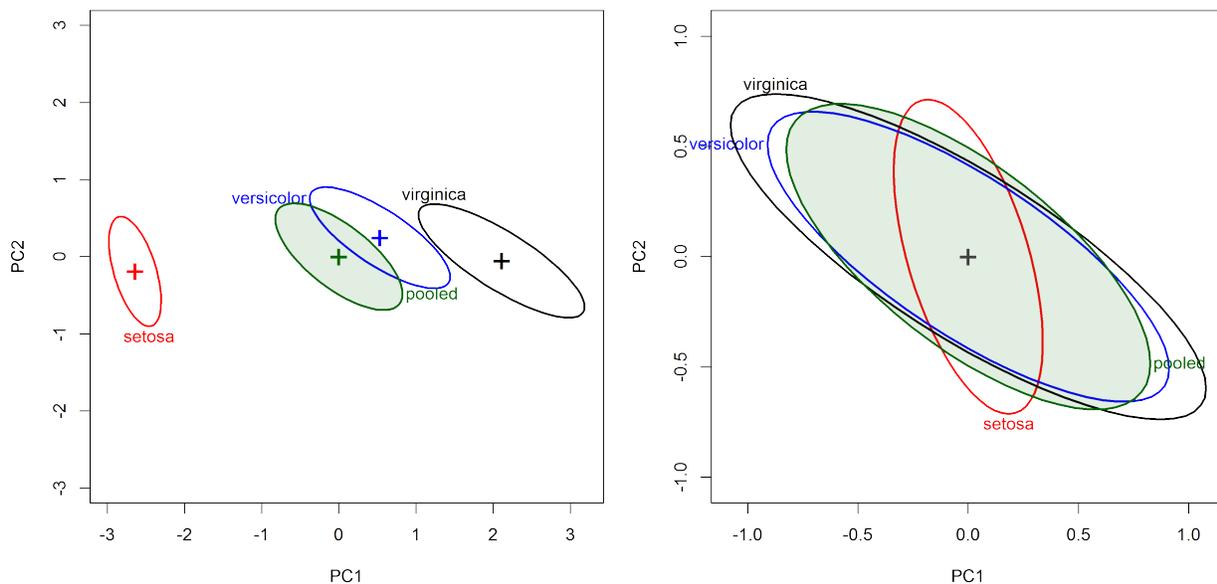


Figure 3: Covariance ellipsoids for the first two principal components of the iris data. Left: Uncentered, showing group means on the principal components; right: centered at the origin.

We note that PCA is focused on a low-rank approximation to account for *total variance* of the data. In the MANOVA context, the main question concerns between-group variance (differences among means) relative to within-group variance. For this question, views in canonical space provide the same advantages, as is described in the Supplemental Materials.

2.3.1 Small dimensions can matter

For the Iris data, the first two principal components account for 98% of total variance, so we might think we are done here. Yet, it turns out that in a variety of multivariate contexts small dimensions can matter. For example, Friendly & Kwan (2009) showed that problems of multicollinearity in regression models could be readily viewed as near singularities that exist in the space of the *smallest* principal component dimensions, but cannot be seen in the

larger dimensions. Similarly, multivariate outliers often do not appear in bivariate views of the data in variable space, but can stand out like sore thumbs in the space of the smallest PCA dimensions.⁷

As we will see, Box’s M test, because it is a (linear) function of all the eigenvalues of the between and within group covariance matrices, is also subject to the influence of the smaller dimensions, where differences among \mathbf{S}_i and of \mathbf{S}_p can lurk.

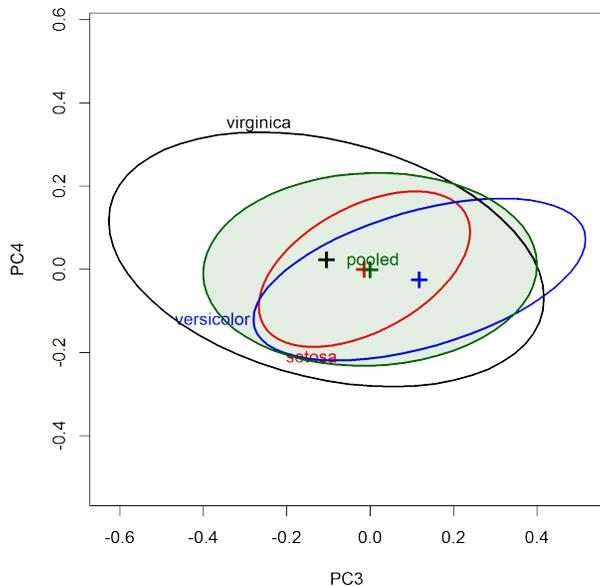


Figure 4: Covariance ellipsoids for the last two principal components.

Figure 4 shows the covariance ellipsoids in (PC3, PC4) space. Even though these dimensions contribute little to total variance, there are more pronounced differences in the within-group shapes (correlations) relative to the pooled covariance. The total sample covariance (ignoring Species) is of course uncorrelated in all principal component dimensions.

3 Other examples

In what follows, it will be instructive to use two other empirical examples to illustrate our graphical methods: (a) one where it turns out that there are important differences among group means but little evidence for heterogeneity of covariances; (b) another where there are

⁷A simple yet powerful demonstration: Generate triples, (x_1, x_2, x_3) as $U[0, 1]$ and scale each set to unit sum, so all points lie on the simplex $x_1 + x_2 + x_3 = 1$. Then add a few outliers within a unit sphere of radius $r \leq 0.05$ centered at the origin. The outliers will not stand out in any univariate or bivariate plots along the coordinate axes, but will be dramatic when viewed along the third principal component, which is orthogonal to the plane of the simplex.

both differences in means and heterogeneity, but the number of response variables is large, which makes understanding these effects more difficult.

3.1 Skulls data

For comparison with what we have seen for the Iris data, the `Skulls` data set provides an example where there are also substantial differences among the means of groups, but little evidence for heterogeneity of their covariance matrices.

The data concern four physical measurements of size and shape made on 150 Egyptian skulls from five epochs ranging from 4000 BC to 150 AD. The measures are: maximal breadth (`mb`), basibregmatic height (`bh`), basialveolar length (`bl`), and nasal height (`nh`) of each skull. See http://www.redwoods.edu/instruct/agarwin/anth_6_measurements.htm for the formal definitions of these measures, and Figure 5 for a diagram of what they pertain to. The question of interest in this analysis is whether and how these measurements changed over time. Systematic changes over time in means and/or covariances is of interest because it could indicate interbreeding among migrant populations (or the influence of other factors, such as diet).

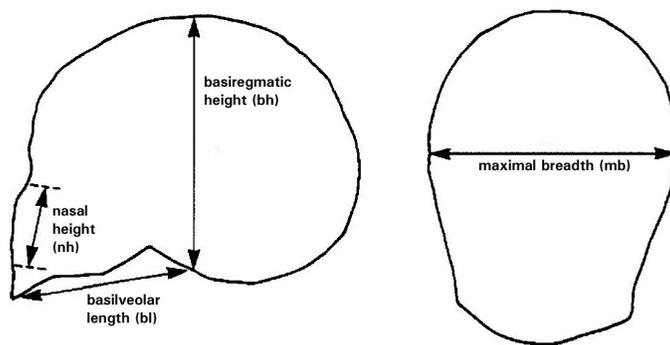


Figure 5: Diagram of the skull measurements for the Egyptian skulls data set. Maximal breadth and basibregmatic height are the basic measures of “size” of a skull. Basialveolar length and nasal height are important anthropometric measures of skull “shape”.

A MANOVA of this data set shows a highly significant effect of the `epoch` factor (Pillai trace = 0.3533, approx. $F(16, 434.45) = 3.512, p < 0.000001$).⁸ Treating `epoch` as an ordered factor yields an even strong test for linear trend in the means over time, and all non-linear trends are effectively null. The conclusion so far is that for these measures of skull size and shape, there are approximately systematic changes over time.

Figure 6 shows the centered covariance ellipsoids for all epochs and for the pooled data.⁹ For

⁸This can be conducted as a MLM in R as follows: `lm(cbind(mb, bh, bl, nh) ~ epoch, data=Skulls)`.

⁹Such figures can be generated using the `covEllipses()` function from the `heplots` package.

the most part, these are all coincident, indicating equal covariance matrices. Only for the variable basialveolar length does any epoch differ perceptibly, where it has slightly greater variance in the earliest epoch (4000BC). It can also be seen that these four measures are relatively uncorrelated within each epoch.

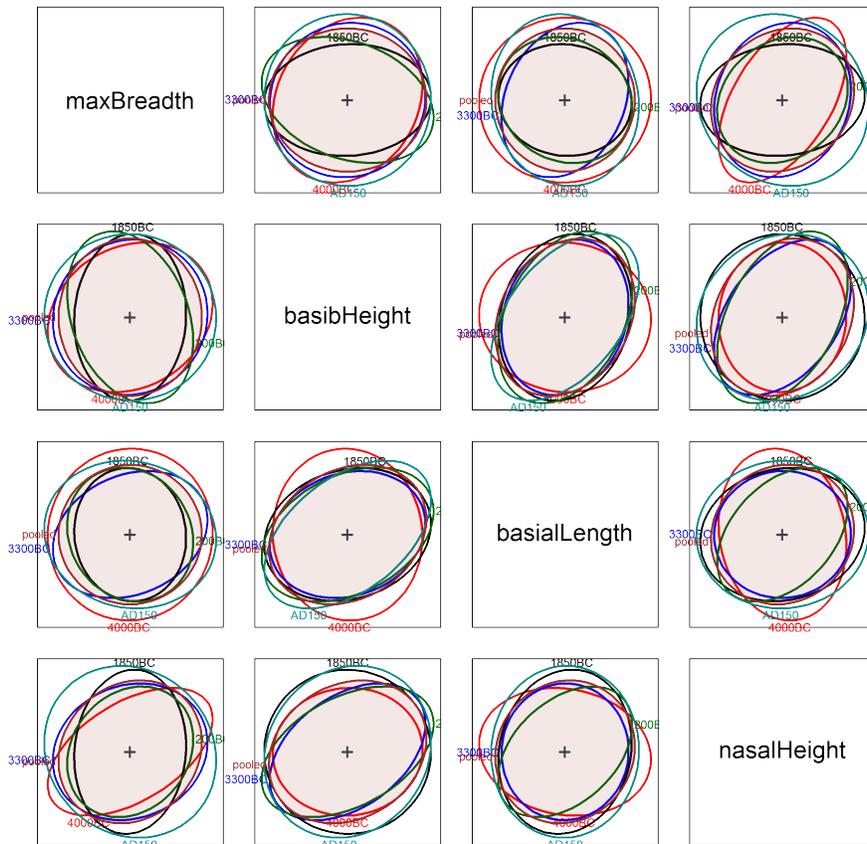


Figure 6: Pairwise data ellipses for the Skulls data, centered at the grand mean. Those for the pooled data are shaded.

3.2 Wine data

The **Wine** data¹⁰ is a classic in the machine learning literature as a high-D classification problem, but is also of interest for examples of MANOVA and discriminant analysis. These data are the results of a chemical analysis of wines grown in the same region in Italy but derived from three different cultivars of grapes: Barolo, Grignolino, and Barbera. The analysis determined the quantities of 13 constituents found in each of the three types of wines. The total sample size is $N = 178$, but the data are unbalanced ($n_i = 59, 71, 48$).

¹⁰This data set is contained in the **candisc** package, and is originally from the UCI Machine Learning Repository (<http://archive.ics.uci.edu/ml/datasets/Wine>).

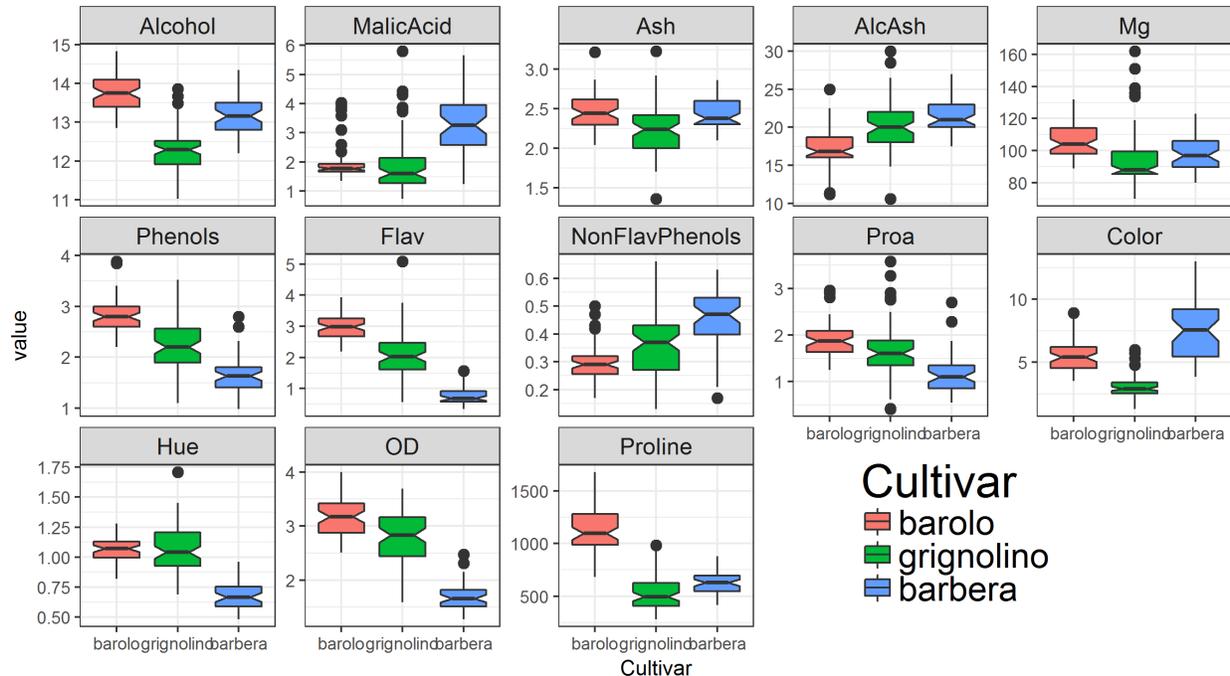


Figure 7: Boxplots of the distributions of the Wine variables, by Cultivar. How do the means differ? How do the variances differ?

By way of introduction to this data set, the set of boxplots in Figure 7 for all of the response variables by `Cultivar` gives a useful overview. It is easy to see that most of the variables differ substantially among the cultivars, but the pattern of differences in means or medians is complex across the variables. There is also a substantial number of outliers for some of the variables, particularly `MalicAcid` and `Proa`. It is much harder to characterize how the wines differ in variance, though differences on some variables appear pronounced (e.g., `MalicAcid`, `Flav`, `Color`).

4 Visualizing Box's M test

The covariance ellipse plots we have seen in earlier examples (e.g., Figure 1 and Figure 6) are useful schematic summaries, but in some cases, a simpler visual summary might be more useful. Eqn. (1) suggests that the simplest visualization might focus on the components of Box's M test, for example, a dot plot of the log determinants of the covariance matrices \mathbf{S}_i together with that of the pooled \mathbf{S}_p . To the extent that the covariance matrices are all equal, so too should the values on which Box's test are based.

An important virtue of these plots is that they can show *how* the groups differ from each other, and from the pooled covariance matrix on the scalar measure $\ln |\mathbf{S}|$. In this way, they can suggest more specific questions or hypotheses regarding the equality of covariance matrices,

analogous to the use of contrasts and linear hypotheses for testing differences among group mean vectors.

Such plots are far more useful with **confidence intervals** around the $\ln|\mathbf{S}_i|$ and $\ln|\mathbf{S}_p|$. Recently, Cai, Liang, & Zhou (2016) have suggested an asymptotic, central-limit theorem approximation to the distribution of $\log|\mathbf{S}|$.¹¹

To illustrate, Box’s M test gives an approximate chi-square, $\chi^2(20) = 140.94, p < 2.2 \times 10^{-16}$ for the Iris data, while the Skulls data gives $\chi^2(40) = 45.67, p = 0.248$. The corresponding plots for these tests are shown in Figure 8. For the Iris data (left), *Setosa* stands out having a substantially smaller covariance matrix (by $\log|\mathbf{S}|$) than the other species. The intervals for *Versicolor* and *Virginica* overlap with that for $\ln|\mathbf{S}_p|$, but seem to differ from each other.

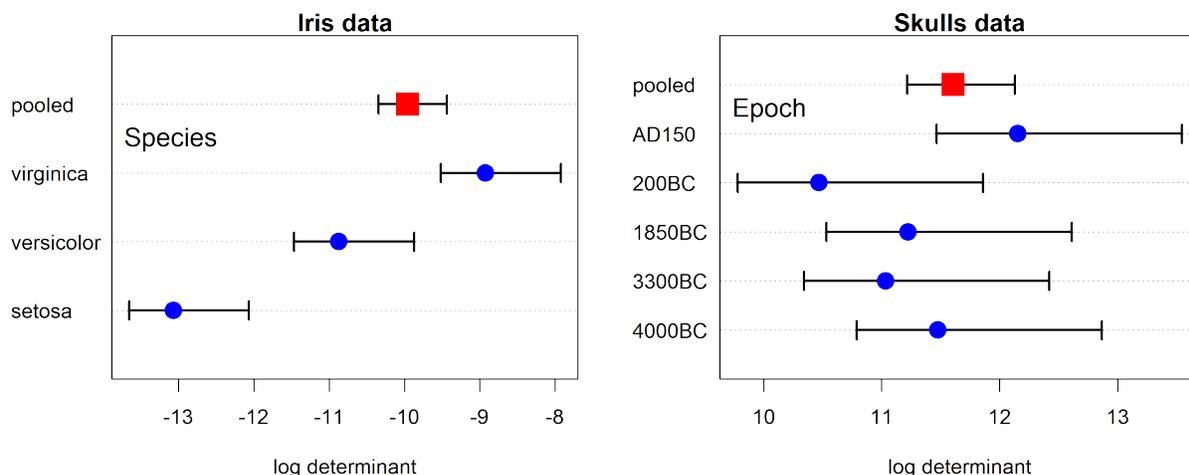


Figure 8: Plots of log determinants of the components of Box’s M test with asymptotic 95% confidence intervals. Left: The Iris data shows substantial heterogeneity; right: the Skulls data shows some small differences, but no evidence for heterogeneity.

In contrast, for the `Skulls` data (Section 3.1), the plot of the log determinants in Figure 8 (right) shows that the 95% confidence intervals for the $\ln|\mathbf{S}_i|$ all overlap with each other and with that for the pooled $\ln|\mathbf{S}_p|$.

Although these differences among covariance matrices are not significant by Box’s M test, we can use this example to illustrate how such plots can suggest scientifically meaningful hypotheses regarding the equality of covariance matrices, analogous to what we are accustomed to doing with tests for mean differences. Assume that changes in variances and covariances of such skull measurements is of interest, and that we were able to obtain a sample 10 times as large from each epoch. We might then try to interpret the general decrease in $\ln|\mathbf{S}_i|$ from the earliest epoch to 200BC. Were skulls becoming more homogeneous over time? But, what happened in the 150 AD sample?

¹¹Box’s M test is calculated by the function `boxM()` in `heplots`. These plots are produced by the `plot()` method for "boxM" objects.

For completeness, and use below, Figure 9 shows the same type of plot for the Wine data. Box's M test is overwhelmingly significant, $\chi^2(182) = 684.2, p < 2.2 \times 10^{-16}$, and the figure shows why: the $\ln|\mathbf{S}_i|$ for Barbera and Barolo differ substantially from that of the pooled $\ln|\mathbf{S}_p|$ and from that of Grignolino, which is closest to that of the pooled covariance matrix. But the Wine data has 13 response variables; perhaps there is a way to understand this further.

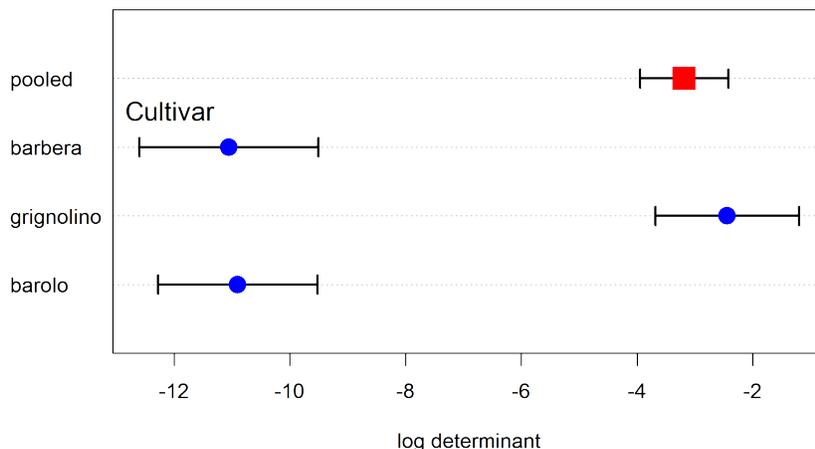


Figure 9: Plot of log determinants of the covariance matrices for the Wine data.

4.1 Eigenvalue plots

Having made some progress with visualizing the components of Box's M test, it is natural to ask if other plots or other test statistics can address these relationships in a more nuanced manner. In the MLM, the various test statistics (Wilks' Λ , Hotelling-Lawley and Pillai Trace criteria, Roy's maximum root test) are all functions of the eigenvalues of a hypothesis matrix \mathbf{H} relative to an error matrix \mathbf{E} . So too, all reasonable test statistics for equality of covariance matrices are functions of the eigenvalues of the \mathbf{S}_i and \mathbf{S}_p .

Another sensible plot is therefore an analog of a *scree plot* of eigenvalue versus dimension number, with separate curves for each matrix, similar to their use in exploratory factor analysis.

Figure 10 shows such a plot for the Wine data. This dataset is comprised of 13 response variables, so Box's test is based on 13 eigenvalues. To preserve resolution, we show eigenvalues on the log scale, and in two separate panels.

It can be seen that the eigenvalues on the largest dimensions do not differ very much across the groups. However, they differ progressively more among the groups on the dimensions with small eigenvalues. Again (as in Figure 9), Grignolino (indicated as group 2) is rather

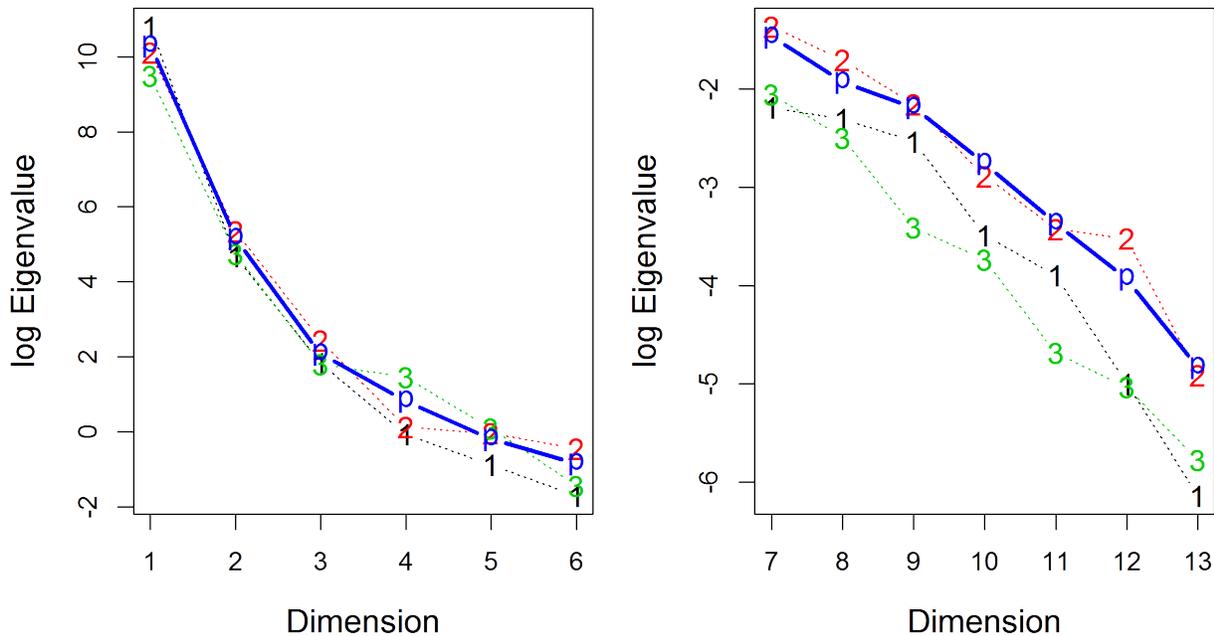


Figure 10: Scree plots of log eigenvalues of the covariance matrices for the Wine data. Those for the pooled covariance matrix are marked ‘p’.

similar to the result for the pooled covariance matrix, while the Barolo (group 1) and Barbera (group 3) wines differ. This demonstrates that Box’s M test is indeed sensitive to differences among the smaller eigenvalues.

4.2 Other test statistics

As we saw above (Section 2), the question of equality of covariance matrices can be expressed in terms of the similarity in size and shape of the data ellipses for the individual group \mathbf{S}_i relative to that of \mathbf{S}_p . Box’s M test uses just one possible function to describe this size: the logs of their determinants.

When Σ is the covariance matrix of a multivariate vector \mathbf{y} with eigenvalues $\lambda_1 \geq \lambda_2 \geq \dots \lambda_p$, the properties shown in Table 1 represent methods of describing the size and shape of the ellipsoid in \mathbb{R}^p . More general theory and statistical applications of the geometry of ellipsoids is given by Friendly et al. (2013).

Hence, for a sample covariance matrix \mathbf{S} , $|\mathbf{S}|$ is a measure of generalized variance and $\ln |\mathbf{S}|$ is a measure of average variance across the p dimensions.

The "boxM" methods in `heplots` can compute and plot all of the functions of the eigenvalues in Table 1. The results are shown in Figure 11.

Table 1: Statistical and geometrical properties of “size” of an ellipsoid

Size	Conceptual formula	Geometry	Function
(a) Generalized variance:	$\det(\mathbf{\Sigma}) = \prod_i \lambda_i$	area, (hyper)volume	geometric mean
(b) Average variance:	$\text{tr}(\mathbf{\Sigma}) = \sum_i \lambda_i$	linear sum	arithmetic mean
(c) Average precision:	$1/\text{tr}(\mathbf{\Sigma}^{-1}) = 1/\sum_i(1/\lambda_i)$		harmonic mean
(d) Maximal variance:	λ_1	maximum dimension	supremum

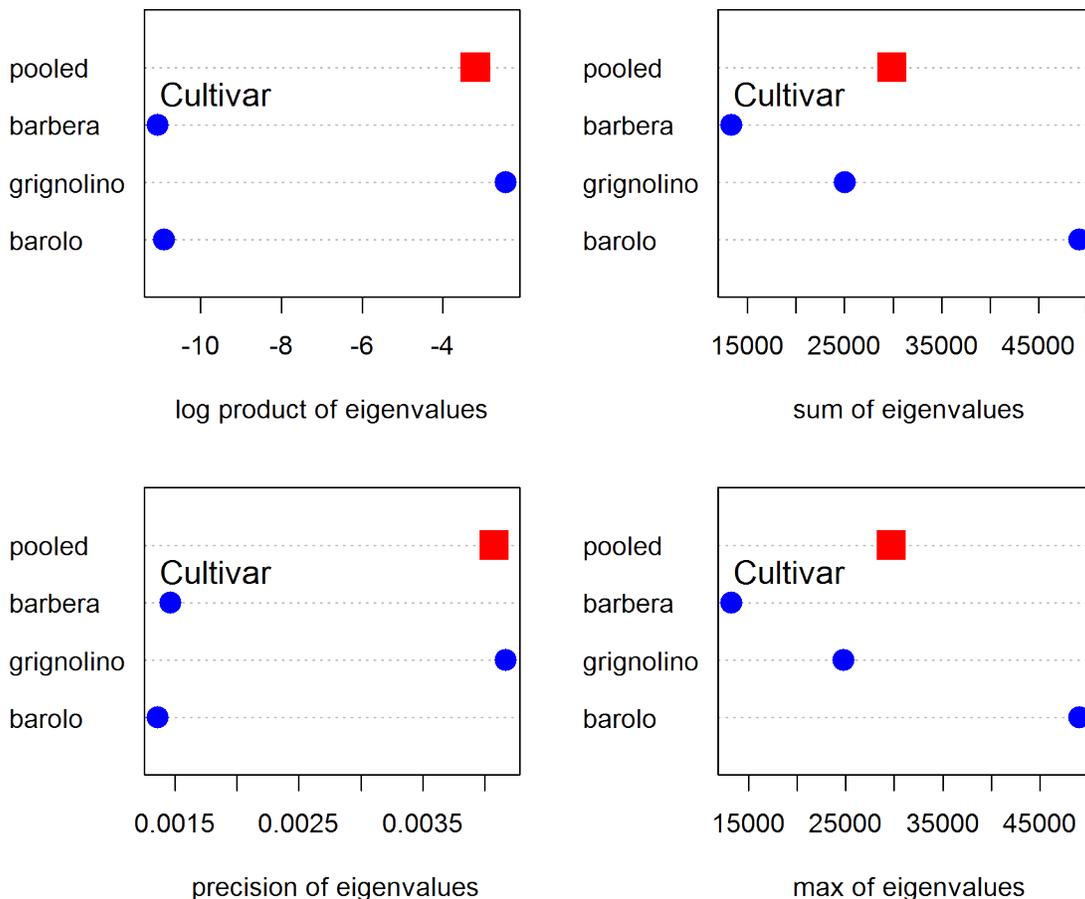


Figure 11: Plot of eigenvalue statistics of the covariance matrices for the Wine data.

Except for the absence of error bars, the plot for log product in Figure 11 (upper left panel) is the same as that in Figure 9. In principle, it is possible to add such confidence intervals for all these measures through the use of bootstrapping, but this has not yet been implemented.

For this data set, the pattern of points in the plot for Box’s M is also more or less the same as that for the precision measure. The plots for the sum of and maximum eigenvalue are also similar to each other, but differ from those of the two measures in the left column of Figure 11. The main point is that these are not *all the same*, so different functions reflect

different patterns of the eigenvalues.¹²

The sum of eigenvalues is similar in form to the Hotelling-Lawley trace criterion used in MLM test and the maximum eigenvalue is similar to Roy’s test, and all these test statistics have convenient F approximations. Plots such as Figure 11 and those we have tried for other examples suggest that it would be useful to develop analogous tests for homogeneity of covariance matrices, with possibly different properties of error rates and power against specified alternatives. We do not pursue this topic here, but leave it open for further research.

5 Multivariate generalization of Levene-type tests

In the MANOVA setting, a multivariate version of Levene’s test (and robust Brown-Forsythe extensions) for the standard model $\mathbf{y}_{ij} = \boldsymbol{\mu}_i + \boldsymbol{\epsilon}_{ij}$ considers instead the multivariate analog substituting \mathbf{z}_{ij} for \mathbf{y}_{ij} , where $\mathbf{z}_{ij} = |\mathbf{y}_{ij} - \tilde{\mathbf{y}}_i|$ are the deviations from a “central” value (mean, median, trimmed mean) $\tilde{\mathbf{y}}_i$ for group $i, i = 1, 2, \dots, g; j = 1, 2, \dots, p$. In matrix terms, this can be expressed as the linear model,

$$\mathbf{Z}_{N \times p} = \mathbf{X}_{(N \times g)} \mathbf{B}_{(g \times p)} + \mathbf{U}_{(N \times p)}$$

where \mathbf{X} is design matrix for the groups, typically composed of an intercept and $g - 1$ contrasts or dummy variables, \mathbf{B} is the matrix of parameters, and \mathbf{U} is the matrix of residuals.

The same multivariate tests (Wilks’ Λ , Pillai trace, etc.) used to test differences in mean vectors \mathbf{y} can then be used to test for differences in the means of the centered deviations, \mathbf{z} . For example, this test applied to the Wine data gives Pillai trace = 0.7924, approx. $F(26, 328) = 8.278, p < 0.000001$. The overall multivariate test is highly significant, leading us to conclude that the covariance matrices differ, as we saw numerically in Box’s M test and in Figure 9. But the question is, *where* and *how* do they differ in relation to the response variables?

For multivariate tests of means, Hypothesis-Error (HE) plots and related canonical HE plots provide an ideal method to visualize the patterns of group mean differences in multivariate space. It would take us too far afield to describe these methods here; we give a brief summary and examples using the Wine data in the Supplemental Appendix. The essential ideas are that: (a) variation in means is captured by a matrix \mathbf{H} of sum of squares and products of the fitted values relative to an analogous matrix \mathbf{E} for error; (b) these can be visualized as suitably scaled data ellipsoids for \mathbf{H} and \mathbf{E} in variable space or in the canonical space of the linear combinations accounting for greatest differences in group means.

The fact that this extension of the Levene tests can be treated as a MANOVA of the deviations

¹²In analogous examples for other data sets we also see different patterns over measures, but these are not the same as in this example.

in \mathbf{Z} suggests that HE plots can also be used to visualize the differences in variances and covariances across groups, as we illustrate in the remainder of this section. Note that in this formulation,

- The overall multivariate test by the General Linear Test of $H_0 : \mathbf{LB} = \mathbf{0}$ gives rise to an \mathbf{H} matrix against the hypothesis, here of equal absolute deviations from the centering values, and an \mathbf{E} matrix, expressing within group sum of squares and cross products of these deviations.
- The diagonal elements of \mathbf{H} are exactly the same as would be used in univariate Levene-type tests. The off-diagonal entries reflect the covariances of the absolute deviations.
- The corresponding \mathbf{E} matrix expresses the within-group variations and their covariances in the absolute deviations.
- The univariate versions of this test simply use the diagonal elements of the \mathbf{H} and \mathbf{E} matrices, i.e., $F_i = \frac{h_{ii}/df_h}{e_{ii}/df_e}$, with $df_h = g - 1$ and $df_e = N - g$.
- HE plots in variable space show how the means of group variabilities differ in relation to their within-group variances and covariances.
- Comparable plots in canonical space show how means of (absolute) deviations differ in a space that accounts for the maximal differences among groups.

A similar test was proposed by O'Brien (1992), but there is a subtle difference between our proposal and his. O'Brien applied MANOVA to the matrix $\mathbf{W}_{N \times p^2}$ whose (i, j) -th row is the vectorized version of the $p \times p$ outer product, $[(\mathbf{y}_{ij} - \tilde{\mathbf{y}}_i)(\mathbf{y}_{ij} - \tilde{\mathbf{y}}_i)^\top]$. This \mathbf{W} directly reflects the contributions of each \mathbf{y}_{ij} to the variances and covariances of each group, but it is not amenable to visual display. In our proposal, the covariances are reflected in the cross-products of the the absolute deviations from the central values calculated in the \mathbf{H} and \mathbf{E} matrices and enjoy the same visualizations for scatter as for means.

5.1 HE plots for multivariate Levene-type tests

To return to the Wine dataset, a pairwise HE plot of all 13 responses in variable space is too complex to show here with any resolution. Instead, Figure 12 shows these for four selected variables. The means for the groups in the \mathbf{H} ellipses reflect the variances of the two responses across the wine cultivars. A positive orientation of the ellipse in a given panel means that the groups with larger variance on one variable also have larger variance on the other, as is true for the variables Flav and OD. The overall nature of these relations for all variables is more easily seen through a canonical HE plot, which is considered next.

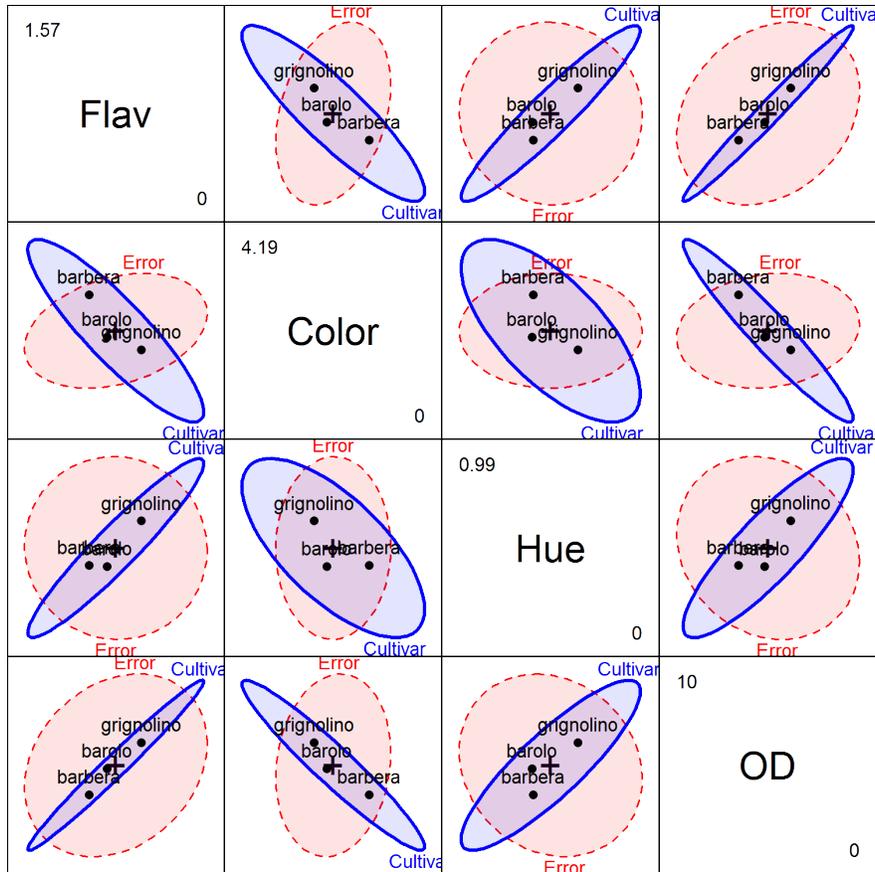


Figure 12: HE plot for selected variables in the multivariate extension of Levene's test showing significant differences in group scatter.

5.2 Canonical views

A canonical discriminant analysis of the MLM (`winedev.mod`) based on median-centered deviations was carried out (`candisc(winedev.mod)`). Because there are $g = 3$ groups, only two canonical dimensions are needed to account for *all* group differences in the means of the absolute deviations; this provides a compact summary of heterogeneity in variances and covariances among the wine cultivars.

The two squared canonical correlations are $\rho_1^2 = 0.528$ and $\rho_2^2 = 0.265$, accounting for respectively 75.6% and 24.4% of the between-group differences in scatter. The standard step-down likelihood ratio tests of these gave $F(26, 326) = 8.74, p < 2.2 \times 10^{-16}$ and $F(12, 164) = 4.92, p < 6.37 \times 10^{-7}$ respectively, both highly significant.

The 2D canonical plot is shown in Figure 13. The horizontal dimension (Can1) shows that Barbera wines differ most in variability from the other two cultivars. The variable `Color` stands out in this plot as the only variable on which Barbera wines are have greater variance than the rest; this can also be seen in the boxplot for this variable (Figure 7). The second

dimension mainly distinguishes Barolo wines from the other two cultivars. The variables Proline, Malic Acid and Non-flavor phenols contribute most strongly to this dimension.

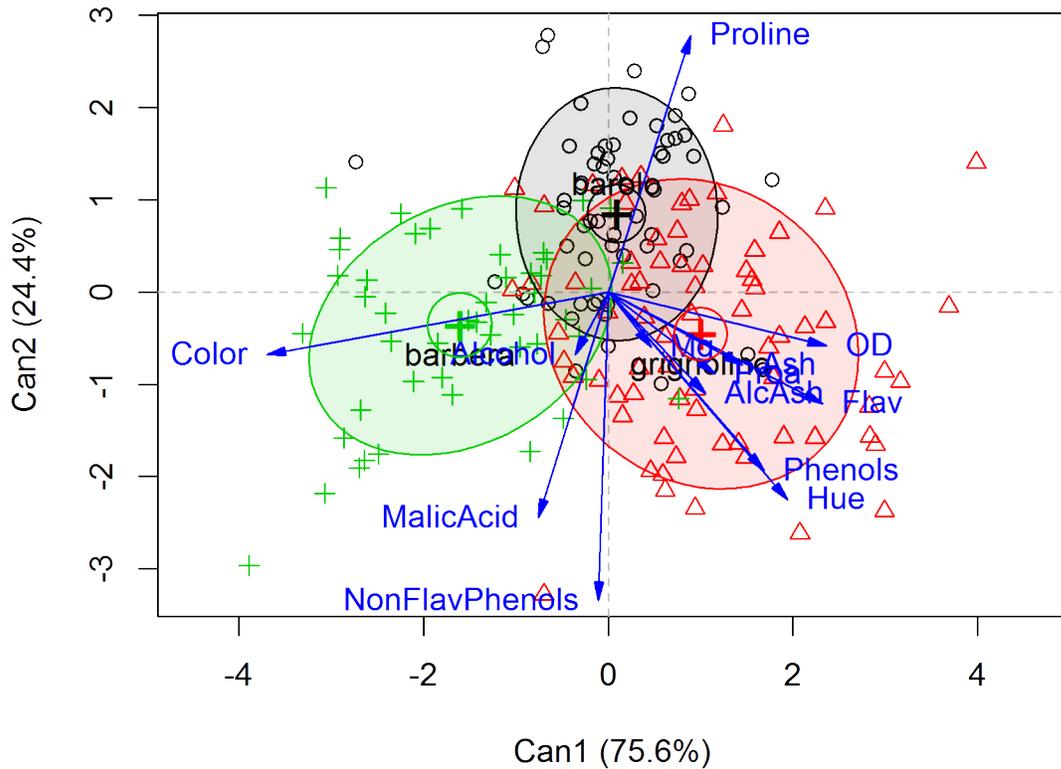


Figure 13: 2D canonical discriminant plot for differences in covariance matrices in the Wine data. The canonical scores (points) for the samples are summarized by their data ellipses. Variable vectors shown in this plot show the correlations of the responses with the two dimensions.

We admit that HE plots themselves are somewhat novel. This extension to tests for differences in variances takes a bit more thought to get used to. However, for this relatively complex problem, the canonical plot in Figure 13 provides a relatively simple description.

6 Concluding Remarks

The idea for this paper arose from discussion in a graduate course on multivariate data analysis in connection with the topic of the assumptions of MANOVA (independence, multivariate normality of residuals, homogeneity of covariance matrices), diagnostic tests for these, and

remedies when they are violated. The instructor (the first author) had presented Box’s M test, with the caveat of the opening quotation by George Box that the test was not robust against mild-to-moderate non-normality, while the Pillai-Bartlett trace test for mean differences does possess this robustness.

One student commented that this was akin to recommending a screening test for lung cancer that was more sensitive to influenza. Another student asked, “Well, if I run Box’s test and it shows significance, how can I decide if my MANOVA results are valid? How can I determine which groups differ in covariance matrices and on which variables?”

These seemed to be perfectly reasonable questions, and while we were aware of the modifications that might account for non-normality (O’Brien, 1992; Tiku & Balakrishnan, 1984), these are not typically available in standard software and therefore their theory is cold comfort to applied researchers. More importantly, we were struck with how little information is provided by the result of such a significance test, regardless of its Type I error and power properties. What was lacking in the null hypothesis p -value was any *insight* into the nature of group differences in covariance matrices.

The approach we have outlined here stems from the mantras of exploratory data analysis (EDA) in a multivariate setting:

The greatest value of a picture is when it forces us to notice what we never expected to see (Tukey, 1977, p. vi).

The purpose of [data] display is comparison (recognition of phenomena), not numbers (Tukey, 1990).

We began with the idea of displaying covariance matrices directly in terms of data ellipsoids that serve as normal theory minimally sufficient visual summaries. We mentioned, but did not illustrate, that classical estimates of \mathbf{S} could be replaced by robust MVE and MCD alternatives in the presence of multivariate outliers.

The simple dot plot (Figure 9) of the components of Box’s M test answers the first question: How do the groups differ in covariance matrices? It does more than this, however, because it suggests that *other* functions of the eigenvalues of the \mathbf{S}_i and \mathbf{S}_p might provide alternative measures of homogeneity of covariance matrices (Figure 11), and that their distribution across the orthogonal dimensions of within-group variation provides valuable insight into the properties of these measures and statistical tests.

The HE plot framework we have described provides some answers to the second question: If the multivariate responses do differ among groups in their variances and covariances, which responses contribute to this and how do they differ? We formulated an extension of the univariate Levene–Brown–Forsythe test to the MANOVA setting, with the property that HE plots and canonical discriminant HE plots for the question of mean differences apply directly to the question of homogeneity of covariance matrices. Pairwise HE plots in variable space

show which responses differ in scatter, and their projection into canonical space (Figure 13) provides a convenient 2D representation, often with quite a simple interpretation.

Finally, we do not fear that this paper might attract the opprobrium that we “flung data onto many canvases to see what stuck.” To the contrary, we were struck by the relative degree of consistency of the resulting plots across these various visual approaches. When covariance ellipsoids showed differences among groups in variable space or principal components space, we could understand how these differences were reflected in the Box’s M plots and in our HE plots for the extension of Levene-type tests. All of this is much more satisfying than a p -value, robust or otherwise.

Supplementary Materials

A Supplemental Appendix describes the HE plot framework for tests of mean differences and gives some further examples. The R packages `candisc` and `heplots` are freely available from the Comprehensive R Archive Network, <http://cran.us.r-project.org/>. Complete R scripts for the Iris, Wine, and Skulls examples are available at https://mattsigal.github.io/eqcov_supp/.

References

- Aelst, S. V., & Willems, G. (2011). Robust and efficient one-way MANOVA tests. *Journal of the American Statistical Association*, *106*(494), 706–718. <http://doi.org/10.1198/jasa.2011.tm09748>
- Anderson, E. (1935). The irises of the Gaspé peninsula. *Bulletin of the American Iris Society*, *35*, 2–5.
- Bartlett, M. S. (1937). Properties of sufficiency and statistical tests. *Proceedings of the Royal Society of London. Series A*, *160*(901), 268–282. <http://doi.org/10.2307/96803>
- Box, G. E. P. (1949). A general distribution theory for a class of likelihood criteria. *Biometrika*, *36*(3-4), 317–346. <http://doi.org/10.1093/biomet/36.3-4.317>
- Box, G. E. P. (1950). Problems in the analysis of growth and wear curves. *Biometrics*, *6*, 362–389.
- Box, G. E. P. (1953). Non-normality and tests on variances. *Biometrika*, *40*(3/4), 318–335. <http://doi.org/10.2307/2333350>
- Brown, M. B., & Forsythe, A. B. (1974). Robust tests for equality of variances. *Journal of the American Statistical Association*, *69*(346), 364–367. <http://doi.org/10.1080/01621459>.

1974.10482955

Cai, T. T., Liang, T., & Zhou, H. H. (2016). Law of log determinant of sample covariance matrix and optimal estimation of differential entropy for high-dimensional gaussian distributions. Preprint; to appear in *JMA*. Retrieved from <http://www.stat.yale.edu/~hz68/Covariance-Determinant.pdf>

Cochran, W. G. (1941). The distribution of the largest of a set of estimated variances as a fraction of their total. *Annals of Eugenics*, *11*(1), 47–52. <http://doi.org/10.1111/j.1469-1809.1941.tb02271.x>

Conover, W. J., Johnson, M. E., & Johnson, M. M. (1981). A comparative study of tests for homogeneity of variances, with applications to the outer continental shelf bidding data. *Technometrics*, *23*(4), 351–361. <http://doi.org/10.1080/00401706.1981.10487680>

Dempster, A. P. (1969). *Elements of continuous multivariate analysis*. Reading, MA: Addison-Wesley.

Finch, H., & French, B. (2013). A monte carlo comparison of robust MANOVA test statistics. *Journal of Modern Applied Statistical Methods*, *12*(2), 35–81. <http://doi.org/10.22237/jmasm/1383278580>

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., & Kwan, E. (2009). Where's Waldo: Visualizing collinearity diagnostics. *The American Statistician*, *63*(1), 56–65.

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

Gastwirth, J. L., Gel, Y. R., & Miao, W. (2009). The impact of Levene's test of equality of variances on statistical theory and practice. *Statistical Science*, *24*(3), 343–360. <http://doi.org/10.1214/09-STS301>

Hakstian, A. R., Roed, J. C., & Lind, J. C. (1979). Two-sample t-2 procedure and the assumption of homogeneous covariance matrices. *Psychological Bulletin*, *86*(6), 1255–1263. <http://doi.org/10.1037/0033-2909.86.6.1255>

Hartley, H. O. (1950). The use of range in analysis of variance. *Biometrika*, *37*(3–4), 271–280. <http://doi.org/10.1093/biomet/37.3-4.271>

Harwell, M. R., Rubinstein, E. N., Hayes, W. S., & Olds, C. C. (1992). Summarizing monte carlo results in methodological research: The one- and two-factor fixed effects ANOVA cases. *Journal of Educational and Behavioral Statistics*, *17*(4), 315–339. <http://doi.org/10.3102/>

10769986017004315

Levene, H. (1960). Robust tests for equality of variances. In I. Olkin, S. G. Ghurye, W. Hoeffding, W. G. Madow, & H. B. Mann (Eds.), *Contributions to probability and statistics: Essays in honor of Harold Hotelling* (pp. 278–292). Stanford, Calif: Stanford University Press.

Lix, J., L.M. Keselman, & Keselman, H. (1996). Consequences of assumption violations revisited: A quantitative review of alternatives to the one-way analysis of variance F test. *Review of Educational Research*, *66*(4), 579–619. <http://doi.org/10.3102/00346543066004579>

Olson, C. L. (1974). Comparative robustness of six tests in multivariate analysis of variance. *Journal of the American Statistical Association*, *69*(348), 894–908. <http://doi.org/10.1080/01621459.1974.10480224>

O'Brien, P. C. (1992). Robust procedures for testing equality of covariance matrices. *Biometrics*, *48*(3), 819–827. Retrieved from <http://www.jstor.org/stable/2532347>

R Core Team. (2015). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from <http://www.R-project.org/>

Rogan, J. C., & Keselman, H. J. (1977). Is the ANOVA f-test robust to variance heterogeneity when sample sizes are equal?: An investigation via a coefficient of variation. *American Educational Research Journal*, *14*(4), 493–498. <http://doi.org/10.3102/00028312014004493>

Tiku, M., & Balakrishnan, N. (1984). Testing equality of population variances the robust way. *Communications in Statistics - Theory and Methods*, *13*(17), 2143–2159. <http://doi.org/10.1080/03610928408828818>

Timm, N. H. (1975). *Multivariate analysis with applications in education and psychology*. Belmont, CA: Wadsworth (Brooks/Cole).

Todorov, V., & Filzmoser, P. (2010). Robust statistic for the one-way MANOVA. *Computational Statistics & Data Analysis*, *54*(1), 37–48. <http://doi.org/10.1016/j.csda.2009.08.015>

Tukey, J. W. (1977). *Exploratory data analysis*. Reading, MA: Addison Wesley.

Tukey, J. W. (1990). Data-based graphics: Visual display in the decades to come. *Statistical Science*, *5*(3), 327–339.

Welch, B. L. (1947). The generalization of “student’s” problem when several different population variances are involved. *Biometrika*, *34*(1–2), 28–35. <http://doi.org/10.1093/biomet/34.1-2.28>

Zhang, J., & Boos, D. D. (1992). Bootstrap critical values for testing homogeneity of covariance matrices. *Journal of the American Statistical Association*, *87*(418), 425–429. Retrieved from <http://www.jstor.org/stable/2290273>