Elliptical Insights: Understanding Statistical Methods through Elliptical Geometry

Michael Friendly
York University

Georges Monette
York University

John Fox
McMaster University

Draft v. 2.2, March 20, 2012

Abstract

Visual insights into a wide variety of statistical methods, for both didactic and data analytic purposes, can often be achieved through geometric diagrams and geometrically based statistical graphs. This paper extols and illustrates the virtues of the ellipse and her higher-dimensional cousins for both these purposes in a variety of contexts, including linear models, multivariate linear models, and mixed-effect models. We emphasize the strong relationships among statistical methods, matrix-algebraic solutions, and geometry that can often be easily understood in terms of ellipses.

Key words: added-variable plots; Bayesian estimation; concentration ellipse; data ellipse; discriminant analysis; Francis Galton; hypothesis-error plots; kissing ellipsoids; measurement error; mixed-effect models; multivariate meta-analysis; regression paradoxes; ridge regression; statistical geometry

1 Introduction

> Whatever relates to extent and quantity may be represented by geometrical figures. Statistical projections which speak to the senses without fatiguing the mind, possess the advantage of fixing the attention on a great number of important facts.
>
> Alexander von Humboldt (1811, p. ciii)

In the beginning (of modern statistical methods), there was the ellipse. As statistical methods progressed from bivariate to multivariate, the ellipse escaped the plane to a 3D ellipsoid, and then onwards to higher dimensions. This paper extols and illustrates the virtues of the ellipse and her higher-dimensional cousins for both didactic and data analytic purposes.

When Francis Galton (1886) first studied the relationship between heritable traits of parents and their offspring, he had a remarkable visual insight— contours of equal bivariate frequencies in the joint distribution seemed to form concentric shapes whose outlines were, to Galton, tolerably close to concentric ellipses differing only in scale.

Galton's goal was to to predict (or explain) how a characteristic, Y, (e.g., height) of children was related to that of their parents, X. To this end, he calculated summaries, $\mathrm{Ave}(Y \mid X)$, and, for symmetry, $\mathrm{Ave}(X \mid Y)$, and plotted these as lines of means on his diagram. Lo and behold, he had a second visual insight: the lines of means of $(Y \mid X)$ and $(X \mid Y)$ corresponded approximately to the locus of horizontal and vertical tangents to the concentric ellipses. To complete the picture, he added lines showing the major and minor axes of the family of ellipses, with the result shown in Figure 1.

It is not stretching the point too far to say that a large part of modern statistical methods descends from these visual insights:[1] correlation and regression (Pearson, 1896), the bivariate normal distribution, and prin-

[1] Pearson (1920, p. 37) later stated, "that Galton should have evolved all this from his observations is to my mind one of the most noteworthy scientific discoveries arising from pure analysis of observations."

cipal components (Pearson, 1901, Hotelling, 1933) all trace their ancestry to Galton's geometrical diagram.[2]

Basic geometry goes back at least to Euclid, but the properties of the ellipse and other conic sections may be traced to Apollonius of Perga (ca. 262 BC–ca. 190 BC), a Greek geometer and astronomer who gave the ellipse, parabola, and hyperbola their modern names. In a work popularly called the Conics (Boyer, 1991), he described the fundamental properties of ellipses (eccentricity, axes, principles of tangency, normals as minimum and maximum straight lines to the curve) with remarkable clarity nearly 2000 years before the development of analytic geometry by Descartes.

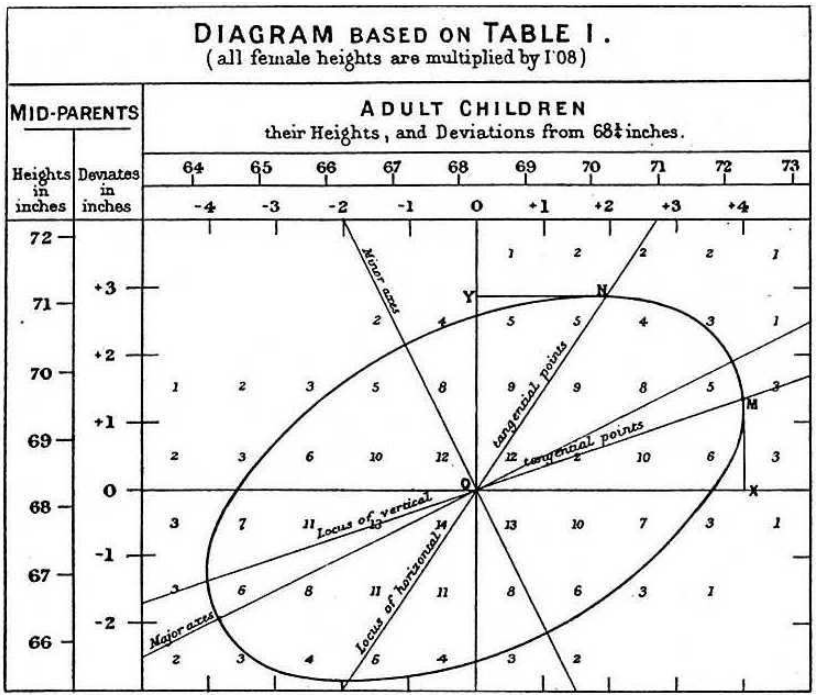

Figure 1: Galton's 1886 diagram, showing the relationship of height of children to the average of their parents' height. The diagram is essentially an overlay of a geometrical interpretation on a bivariate grouped frequency distribution, shown as numbers.

Over time, the ellipse would be called to duty to provide simple explanations of phenomena once thought complex. Most notable is Kepler's insight that the Copernican theory of the orbits of planets as concentric circles (which required notions of epicycles to account for observations) could be brought into alignment with the detailed observational data from Tycho Brahe and others by an exquisitely simple law: "The orbit of every planet is an ellipse with the sun at a focus." One century later, Isaac Newton was able to connect this elliptical geometry with astrophysics by deriving all three of Kepler's laws as simpler consequences of general laws of motion and universal gravitation.

This paper takes up the cause of the ellipse as a geometric form that can provide similar service to statistical understanding and data analysis. Indeed, it has been doing that since the time of Galton, but these graphic and geometric contributions have often been incidental and scattered in the literature (e.g., Bryant, 1984, Campbell and Atchley, 1981, Saville and Wood, 1991, Wickens, 1995). We focus here on visual insights through ellipses in the areas of linear models, multivariate linear models, and mixed-effect models.

[2]Well, not entirely. Auguste Bravais [1811–1863] (1846), an astronomer and physicist first introduced the mathematical theory of the bivariate normal distribution as a model for the joint frequency of errors in the geometric position of a point. Bravais derived the formula for level slices as concentric ellipses and had a rudimentary notion of correlation but did not appreciate this as a representation of data. Nonetheless, Pearson (1920) acknowledged Bravais's contribution, and the correlation coefficient is often called the Bravais-Pearson coefficient in France (Denis, 2001).

Our goal is to provide as comprehensive a treatment of this topic as possible in a single article together with online supplements.

The plan of this paper is as follows: Section 2 provides the minimal notation and properties of ellipsoids[3] necessary for the remainder of the paper. Due to length restrictions, other useful and important properties of geometric and statistical ellipsoids have been relegated to the (online?) Appendix. Section 3 describes the use of the *data ellipsoid* as a visual summary for multivariate data. In Section 4, we apply data ellipsoids and confidence ellipsoids for parameters in linear models to explain a wide range of phenomena, paradoxes and fallacies that are clarified by this geometric approach. This approach is extended to multivariate linear models in Section 5, primarily through the use of ellipsoids to portray hypothesis (H) and error (E) covariation in what we call *HE plots*. Finally, in Section 6, we discuss a diverse collection of statistical problems whose solutions can all be described in terms of "kissing ellipsoids."

2 Notation and basic results

There are various representations of an ellipse (or ellipsoid in three or more dimensions), both geometric and statistical. Some basic notation and properties are described below.

2.1 Geometrical ellipsoids

We refer to the common notion of a bounded ellipsoid (with non-empty interior) in the p-dimensional space $\mathbb{R}^p$ as a *proper ellipsoid*. An origin-centered proper ellipsoid may be defined by the quadratic form

$$\mathcal{E} := \{\boldsymbol{x} : \boldsymbol{x}^\mathsf{T} \boldsymbol{C} \boldsymbol{x} \leq 1\} \ , \tag{1}$$

where equality in Eqn. (1) gives the boundary, $\boldsymbol{x} = (x_1, x_2, \ldots, x_p)^\mathsf{T}$ is a vector referring to the coordinate axes and $\boldsymbol{C}$ is a symmetric positive definite $p \times p$ matrix. If $\boldsymbol{C}$ is only positive semi-definite, then the ellipsoid will be *improper*, having the shape of a cylinder with elliptical cross-sections and unbounded in the direction of the null space of $\boldsymbol{C}$. To extend the definition to *singular* (sometimes known as "degenerate") ellipsoids, we turn to a definition that is equivalent to Eqn. (1) for proper ellipsoids. Let $\mathcal{S}$ denote the unit sphere in $\mathbb{R}^p$,

$$\mathcal{S} := \{\boldsymbol{x} : \boldsymbol{x}^\mathsf{T} \boldsymbol{x} = 1\} \ , \tag{2}$$

and let

$$\mathcal{E} := \boldsymbol{A}\mathcal{S} \ , \tag{3}$$

where $\boldsymbol{A}$ is a non-singular $p \times p$ matrix. Then $\mathcal{E}$ is a proper ellipsoid that could be defined using Eqn. (1) with $\boldsymbol{C} = \left(\boldsymbol{A}^\mathsf{T} \boldsymbol{A}\right)^{-1}$. We obtain singular ellipsoids by allowing $\boldsymbol{A}$ to be any matrix, not necessarily non-singular nor even square. A more general representation of ellipsoids based on the singular value decomposition (SVD) of $\boldsymbol{C}$ is given in Appendix A.1. Some useful properties of geometric ellipsoids are described in Appendix A.2.

2.2 Statistical ellipsoids

In statistical applications, $\boldsymbol{C}$ will often be the inverse of a covariance matrix (or a sum of squares and cross-products matrix), and the ellipsoid will be centered at the means of variables, or at estimates of parameters under some model. Hence, we will also use the following notation:

For a positive definite matrix $\boldsymbol{\Sigma}$ we use $\mathcal{E}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ to denote the ellipsoid

$$\mathcal{E} := \{\boldsymbol{x} : (\boldsymbol{x} - \boldsymbol{\mu})^\mathsf{T} \boldsymbol{\Sigma}^{-1} (x - \boldsymbol{\mu}) = 1\} \ . \tag{4}$$

[3] As in this paragraph, we generally use the term "ellipsoid" as to refer to "ellipse or ellipsoid" where dimensionality doesn't matter, or context is clear.

When $\boldsymbol{\Sigma}$ is the covariance matrix of a multivariate vector $\boldsymbol{x}$ with eigenvalues $\lambda_1 > \lambda_2 > \ldots$, the following properties represent the "size" of the ellipsoid in $\mathbb{R}^p$:

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

For testing hypotheses for parameters of multivariate linear models, these different senses of "size" correspond (with suitable transformations) to (a) Wilks's Λ, (b) the Hotelling-Lawley trace, (c) the Pillai trace, and (d) Roy's maximum root tests, as we describe below in Section 5.

Note that every non-negative definite matrix $\boldsymbol{W}$ can be factored as $\boldsymbol{W} = \boldsymbol{A}\boldsymbol{A}^\mathsf{T}$, and the matrix $\boldsymbol{A}$ can always be selected so that it is square. $\boldsymbol{A}$ will be non-singular if and only if $\boldsymbol{W}$ is non-singular. A computational definition of an ellipsoid that can be used for all non-negative definite matrices and that corresponds to the previous definition in the case of positive-definite matrices is

$$\mathcal{E}(\boldsymbol{\mu}, \boldsymbol{W}) = \boldsymbol{\mu} + \boldsymbol{A}\mathcal{S} \ , \tag{5}$$

where $\mathcal{S}$ is a unit sphere of conformable dimension and $\boldsymbol{\mu}$ is the centroid of the ellipsoid. One convenient choice of $\boldsymbol{A}$ is the Choleski square root, $\boldsymbol{W}^{1/2}$, as we describe in Appendix A.3. Thus, for some results below, a convenient notation in terms of $\boldsymbol{W}$ is

$$\mathcal{E}(\boldsymbol{\mu}, \boldsymbol{W}) = \boldsymbol{\mu} \oplus \sqrt{\boldsymbol{W}} = \boldsymbol{\mu} \oplus \boldsymbol{W}^{1/2} \ , \tag{6}$$

where $\oplus$ emphasizes that the ellipsoid is a scaling and rotation of the unit sphere followed by translation to a center at $\boldsymbol{\mu}$ and $\sqrt{\boldsymbol{W}} = \boldsymbol{W}^{1/2} = \boldsymbol{A}$. This representation is not unique, however: $\boldsymbol{\mu} \oplus \boldsymbol{B} = \boldsymbol{\nu} \oplus \boldsymbol{C}$ (i.e., they generate the same ellipsoid) *iff* $\boldsymbol{\mu} = \boldsymbol{\nu}$ and $\boldsymbol{B}\boldsymbol{B}^\mathsf{T} = \boldsymbol{C}\boldsymbol{C}^\mathsf{T}$. From this result, it is readily seen that under a linear transformation given by a matrix $\boldsymbol{L}$ the image of the ellipse is

$$\boldsymbol{L}[(\mathcal{E}(\boldsymbol{\mu}, \boldsymbol{W}))] = \mathcal{E}(\boldsymbol{L}\boldsymbol{\mu}, \boldsymbol{L}\boldsymbol{W}\boldsymbol{L}^\mathsf{T}) = \boldsymbol{L}\boldsymbol{\mu} \oplus \sqrt{\boldsymbol{L}\boldsymbol{W}\boldsymbol{L}^\mathsf{T}} = \boldsymbol{L}\boldsymbol{\mu} \oplus \boldsymbol{L}\sqrt{\boldsymbol{W}} \ .$$

3 The data ellipse and ellipsoid

The *data ellipse* (Monette, 1990) (or *concentration ellipse*, Dempster, 1969, Ch. 7) provides a remarkably simple and effective display for viewing and understanding bivariate *marginal* 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. Under classical (Gaussian) assumptions, the data ellipse provides a sufficient visual summary, as we describe below.

It is historically appropriate to illustrate the data ellipse and describe its properties using Galton's (1886, Table I) data, from which he drew Figure 1 as a conceptual diagram,[4] shown in Figure 2, where the frequency at each point is represented by a sunflower symbol. We also overlay the 40%, 68% and 95% data ellipses, as described below.

In Figure 2, the ellipses have the mean vector $(\bar{x}, \bar{y})$ as their center; the lengths of arms of the central cross show the standard deviations of the variables, which correspond to the shadows of the 40% ellipse. In addition, the correlation coefficient can be visually represented as the fraction of a vertical tangent line from $\bar{y}$ to the top of the ellipse that is below the regression line $\widehat{y}|x$, shown by the arrow labeled 'r.' Finally, as Galton noted, the regression line for $\widehat{y} \,|\, x$ (or $\widehat{x} \,|\, y$) can be visually estimated as the locus of the points of vertical (or horizontal) tangents with the family of concentric ellipses. See Monette (1990, Figs. 5.1–5.2) and Friendly (1991, p. 183) for illustrations and further discussion of the properties of the data ellipse.

[4] These data are reproduced in Stigler (1986, Table 8.2, p. 286)

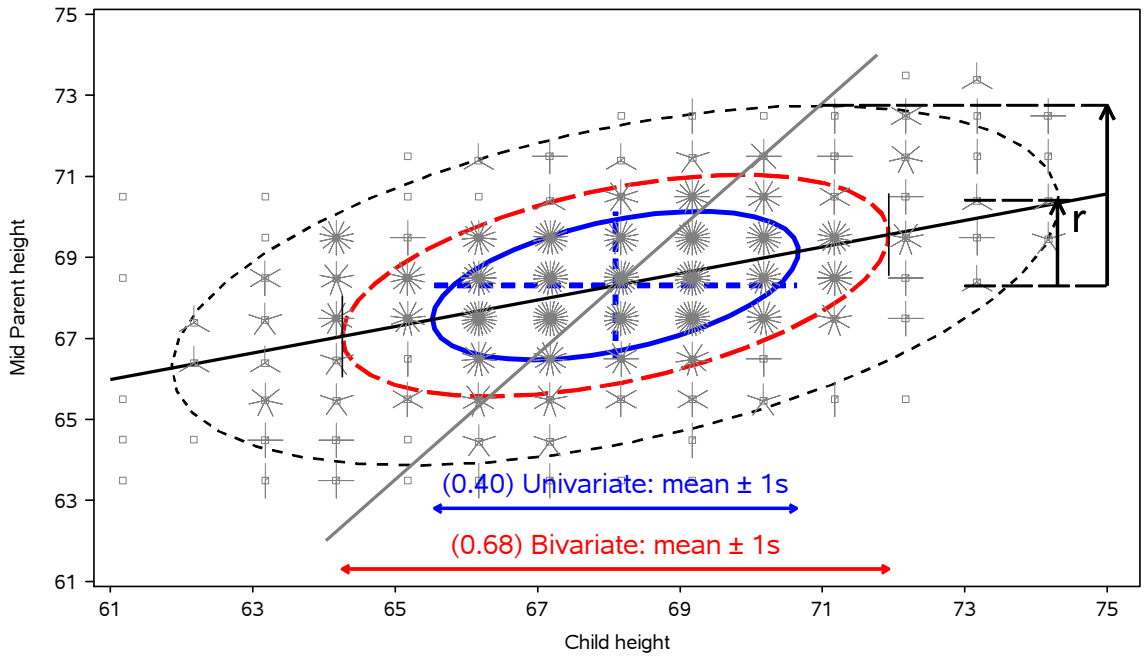

Figure 2: Sunflower plot of Galton's data on heights of parents and their children (in.), with 40%, 68% and 95% data ellipses and the regression lines of y on x (black) and x on y (grey). The ratio of the vertical to the regression line (labeled 'r') to the vertical to the top of the ellipse gives a visual estimate of the correlation (r=0.46, here). Shadows (projections) on the coordinate axes give standard intervals, $\bar{x} \pm s_x$ and $\bar{y} \pm s_y$, with various coverage properties. Plotting children's height on the abscissa follows Galton.

More formally (Dempster, 1969, Monette, 1990), for a p-dimensional sample, $\boldsymbol{Y}_{n \times p}$, we recognize the quadratic form in Eqn. (4) as corresponding to the squared Mahalanobis distance, $D^2_M(\boldsymbol{y}) = (\boldsymbol{y} - \bar{\boldsymbol{y}})^\mathsf{T} \boldsymbol{S}^{-1} (\boldsymbol{y} - \bar{\boldsymbol{y}})$, of the point $\boldsymbol{y} = (y_1, y_2, \ldots, y_p)^\mathsf{T}$ from the centroid of the sample, $\bar{\boldsymbol{y}} = (\bar{y}_1, \bar{y}_2, \ldots, \bar{y}_p)^\mathsf{T}$. Thus, we use a more explicit notation to define the *data ellipsoid* $\mathcal{E}_c$ of size ("radius") c as the set of all points $\boldsymbol{y}$ with $D^2_M(\boldsymbol{y})$ less than or equal to c^2,

$$\mathcal{E}_c(\boldsymbol{y}; \boldsymbol{S}, \bar{\boldsymbol{y}}) := \{\boldsymbol{y} : (\boldsymbol{y} - \bar{\boldsymbol{y}})^\mathsf{T} \boldsymbol{S}^{-1} (\boldsymbol{y} - \bar{\boldsymbol{y}}) \leq c^2\} \ , \tag{7}$$

where $\boldsymbol{S} = (n-1)^{-1} \sum_{i=1}^n (\boldsymbol{y}_i - \bar{\boldsymbol{y}})^\mathsf{T} (\boldsymbol{y}_i - \bar{\boldsymbol{y}})$ is the sample covariance matrix. In the computational notation of Eqn. (6), the boundary of the data ellipsoid of radius c is thus

$$\mathcal{E}_c(\bar{\boldsymbol{y}}, \boldsymbol{S}) = \bar{\boldsymbol{y}} \oplus c \boldsymbol{S}^{1/2} \ . \tag{8}$$

Many properties of the data ellipsoid hold regardless of the joint distribution of the variables; but if the variables are multivariate normal, then the data ellipsoid approximates a contour of constant density in their joint distribution. In this case $D^2_M(x, y)$ has a large-sample χ^2_p 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. Other radii also have useful interpretations:

- In Figure 2, we demonstrate that $c^2 = \chi^2_2(0.40) \approx 1$ gives a data ellipse of 40% coverage with the property that its projection on either axis corresponds to a standard interval, $\bar{x} \pm 1 s_x$ and $\bar{y} \pm 1 s_y$. The same property of univariate coverage pertains to any linear combination of x and y.
- By analogy with a univariate sample, a 68% coverage data ellipse with $c^2 = \chi^2_2(0.68) = 2.28$ gives a bivariate analog of the standard $\bar{x} \pm 1 s_x$ and $\bar{y} \pm 1 s_y$ intervals. The univariate shadows, or those of

any linear combination, then correspond to standard Scheffé intervals taking "fishing" (simultaneous interfence) in a $p = 2$-dimensional space into account.

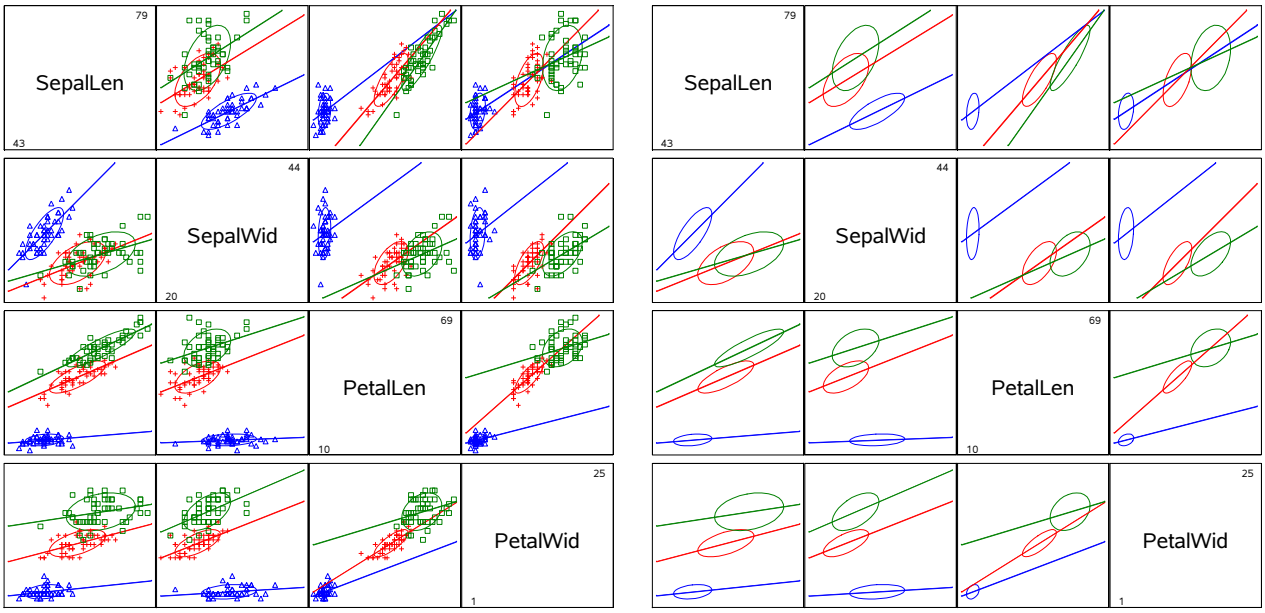

Figure 3: Scatterplot matrices of Anderson's iris data: (a) showing data, separate 68% data ellipses, and regression lines for each species; (b) showing only ellipses and regression lines. Key– *Iris setosa*: blue, △; *Iris versicolor*: red, +; *Iris virginca*: green, □.

As useful as the data ellipse might be for a single, unstructured sample, its value as a visual summary increases with the complexity of the data. For example, Figure 3 shows scatterplot matrices of all pairwise plots of the variables from Edgar Anderson's 1935 classic data on three species of iris flowers found in the Gaspé Peninsula, later used by Fisher (1936) in his development of discriminant analysis. The data ellipses show clearly that the means, variances, correlations, and regression slopes differ systematically across the three iris species in all pairwise plots. We emphasize that the ellipses serve as sufficient visual summaries of the important statistical properties (first and second moments)[5] by removing the data points from the plots in the version at the right.

4 Linear models: data ellipses and confidence ellipses

Here we consider how ellipses help to visualize relationships among variables in connection with linear models (regression, ANOVA). We begin with views in the space of the variables (data space) and progress to related views in the space of model parameters ($\boldsymbol{\beta}$ space).

4.1 Simple linear regression

Various aspects of the standard data ellipse of radius 1 illuminate many properties of simple linear regression, as shown in Figure 4. These properties are also useful in more complex contexts.

- One-half of the widths of the vertical and horizontal projections (dotted black lines) give the standard deviations s_x and s_y respectively.

[5]We recognize that a normal-theory summary (first and second moments), shown visually or numerically, can be distorted by multivariate outliers, particularly in smaller samples. In what follows, robust covariance estimates can, in principle, be substituted for the classical, normal-theory estimates in all cases. To save space, we don't explore these possibilities further here.

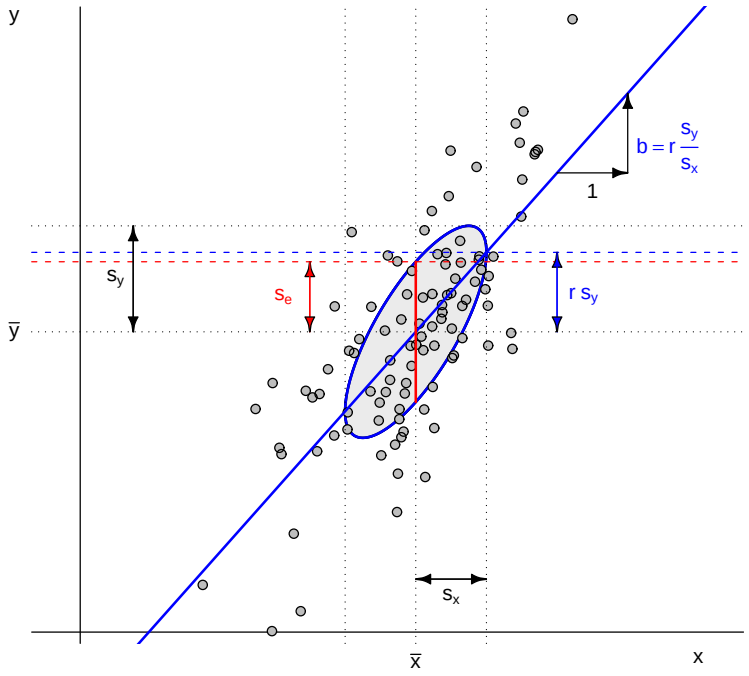

Figure 4: Annotated standard data ellipse showing standard deviations of x and y, residual standard deviation (s_e), slope (b), and correlation (r).

- Because any line through the center of the ellipse $(\bar{x}, \bar{y})$, corresponds to some linear combination, $mx + ny$, the half-width of the corresponding tangent lines gives the standard deviation of this linear combination.
- The standard deviation of the residuals, s_e can be visualized as the half-width of the vertical (red) line at $x = \bar{x}$.
- The vertical distance between the mean of y and the points where the ellipse has vertical tangents is rs_y. (As a fraction of s_y, this distance is $r = 0.75$ in the figure.)
- The (blue) regression line of y on x passes through the points of vertical tangency. Similarly, the regression of x on y (not shown) passes through the points of horizontal tangency.

4.2 Visualizing a confidence interval for the slope

A visual approximation to a 95% confidence interval for the slope, and thus a visual test of $H_0 : \beta = 0$ can be seen in Figure 5. From the formula for a 95% confidence interval, $\mathrm{CI}_{.95}(\beta) = b \pm t^{0.975}_{n-2} \times \mathrm{SE}(b)$, we can take $t^{0.975}_{n-2} \approx 2$ and $\mathrm{SE}(b) \approx \frac{1}{\sqrt{n}} \left(\frac{s_e}{s_x}\right)$, leading to

$$\mathrm{CI}_{.95}(\beta) \approx b \pm \frac{2}{\sqrt{n}} \times \left(\frac{s_e}{s_x}\right) \quad . \tag{9}$$

To show this visually, the left panel of Figure 5 displays the standard data ellipse surrounded by the "regression parallelogram," formed with the vertical tangent lines and the tangent lines parallel to the regression line. This corresponds to the conjugate axes of the ellipse induced by the Choleski factor of S_{yx} as shown in

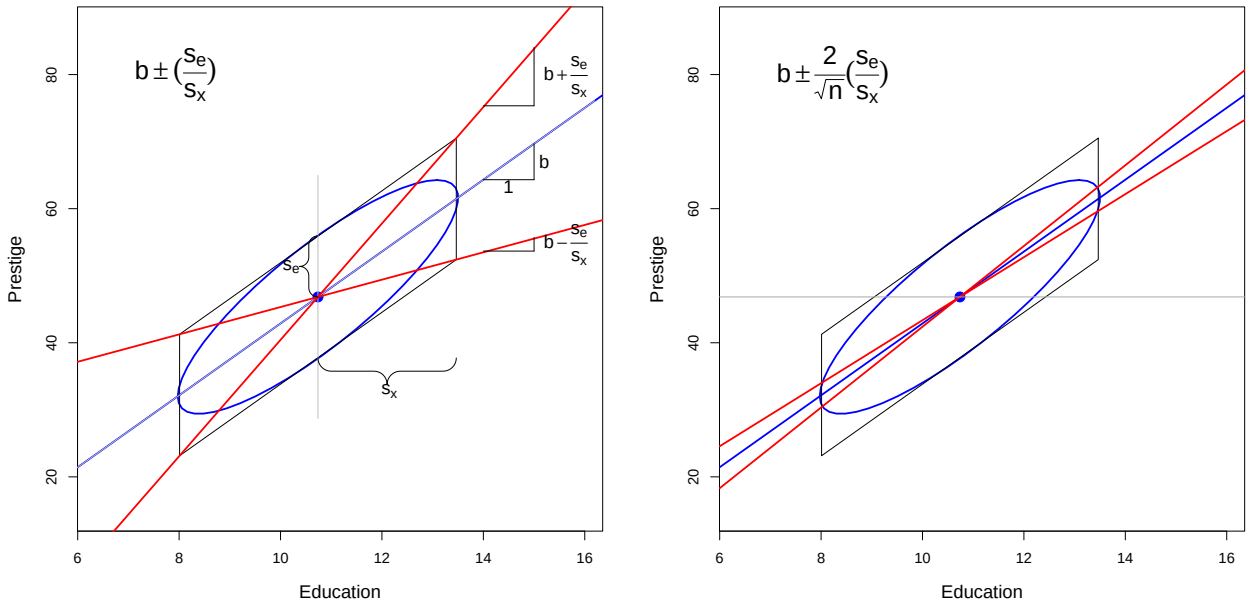

Figure 5: Visual 95% confidence interval for the slope in linear regression. Left: Standard data ellipse surrounded by the regression parallelogram. Right: Shrinking the diagonal lines by a factor of $2/\sqrt{n}$, gives the approximate 95% confidence interval for β.

Figure A.3 in Appendix A.3. Simple algebra demonstrates that the diagonal lines through this parallelogram have slopes of

$$b \pm \frac{s_e}{s_x}$$

So, to obtain a visual estimate of the 95% confidence interval for β (*not*, we note, the 95% CI for the regression line), we need only shrink the diagonal lines of the regression parallelogram toward the regression line by a factor of $2/\sqrt{n}$, giving the red lines in the right panel of Figure 5. In the data used for this example, $n = 102$, so the factor is approximately 0.2 here.[6] Now consider the horizontal line through the center of the data ellipse. If this line is outside the envelope of the confidence lines, as it is in Figure 5, we can reject $H_0 : \beta = 0$ via this simple visual approximation.

4.3 Simpson's paradox, marginal and conditional relationships

Because it provides a visual summary of means, variances, and correlations, the data ellipse is ideally suited as a tool for illustrating and explicating various phenomena that occur in the analysis of linear models. One class of simple, but important, examples concerns the difference between the marginal relationship between variables, ignoring some important factor or covariate, and the conditional relationship, adjusting (controlling) for that factor or covariate.

Simpson's paradox (Simpson, 1951) occurs when the marginal and conditional relationships differ in direction. This may be seen in the plots of Sepal length against Sepal width for the iris data shown in Figure 6. Ignoring iris species, the marginal, total-sample correlation is slightly negative as seen in panel (a). The individual-sample ellipses in panel (b) show that the conditional, within-species correlations are all positive, with approximately equal regression slopes. The group means have a negative relationship, accounting for the negative marginal correlation.

[6]The data are for the rated prestige and average years of education of 102 Canadian occupations circa 1970; see Fox and Suschnigg (1989).

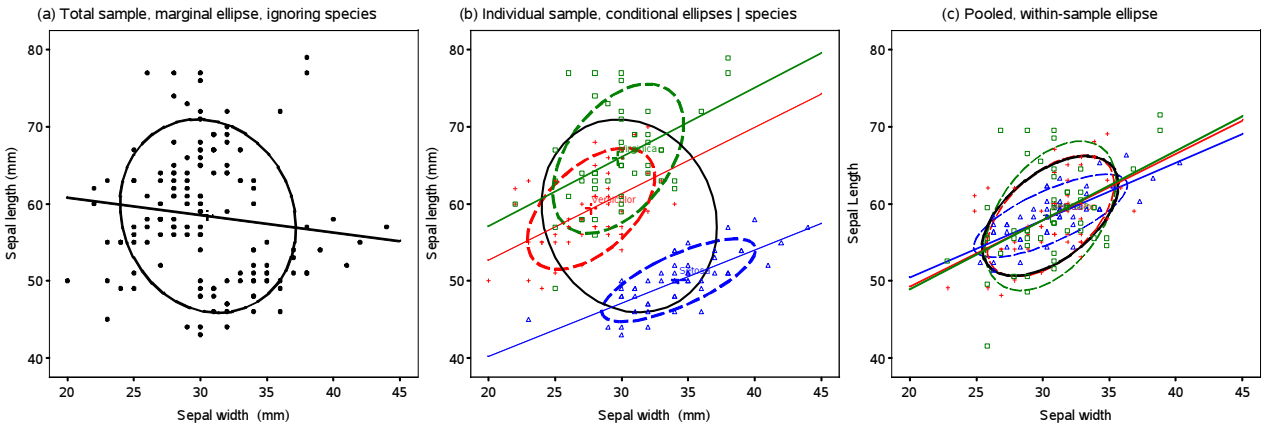

Figure 6: Marginal (a), conditional (b), and pooled within-sample (c) relationships of Sepal length and Sepal width in the iris data. Total-sample data ellipses are shown as black, solid curves; individual-group data and ellipses are shown with colors and dashed lines

A correct analysis of the (conditional) relationship between these variables, controlling or adjusting for mean differences among species, is based on the pooled within-sample covariance matrix,

$$\boldsymbol{S}_{\text{within}} = (N-g)^{-1} \sum_{i=1}^{g} \sum_{j=1}^{n_i} (\boldsymbol{y}_{ij} - \bar{\boldsymbol{y}}_{i\cdot})(\boldsymbol{y}_{ij} - \bar{\boldsymbol{y}}_{i\cdot})^{\mathsf{T}} = (N-g)^{-1} \sum_{i=1}^{g} (n_i - 1)\boldsymbol{S}_i \ , \tag{10}$$

where $N = \sum n_i$, and the result is shown in panel (c) of Figure 6. In this graph, the data for *each* species were first transformed to deviations from the species means on both variables and then translated back to the grand means.

In a more general context, $\boldsymbol{S}_{\text{within}}$ appears as the $\boldsymbol{E}$ matrix in a multivariate linear model, adjusting or controlling for all fitted effects (factors and covariates). For essentially correlational analyses (principal components, factor analysis, etc.), similar displays can be used to show how multi-sample analyses can be compromised by substantial group mean differences, and corrected by analysis of the pooled within-sample covariance matrix, or by including important group variables in the model. Moreover, display of the the individual within-group data ellipses can show visually how well the assumption of equal covariance matrices, $\Sigma_1 = \Sigma_2 = \cdots = \Sigma_g$, is satisfied in the data, for the two variables displayed.

4.4 Other paradoxes and fallacies

Data ellipses can also be used to visualize and understand other paradoxes and fallacies that occur with linear models. We consider situations in which there is a principal relationship between variables y and x of interest, but (as in the preceding subsection) the data are stratified in g samples by a factor ("group") that might correspond to different subpopulations (e.g., men and women, age groups), different spatial regions (e.g., states), different points in time, or some combination of the above.

In some cases, group may be unknown, or may not have been included in the model, so we can only estimate the marginal association between y and x, giving a slope β_{marginal} and correlation r_{marginal}. In other cases, we may not have individual data, but only aggregate group data, $(\bar{y}_i, \bar{x}_i), i = 1, \ldots, g$, from which we can estimate the between-groups ("ecological") association, with slope β_{between} and correlation r_{between}. When all data are available and the model is an ANCOVA model of the form $y \sim x + \text{group}$, we can estimate a common conditional, within-group slope, β_{within}, or, with the model $y \sim x + x \times \text{group}$, the separate within-group slopes, β_i.

Figure 7 illustrates these estimates in a simulation of five groups, with $n_i = 10$, means $\bar{x}_i = 2\,i + \mathcal{U}(-0.4, 0.4)$ and $\bar{y}_i = \bar{x}_i + \mathcal{N}(0, 0.5^2)$, so that $r_{\text{between}} \approx 0.95$. Here $\mathcal{U}(a, b)$ represents the uniform

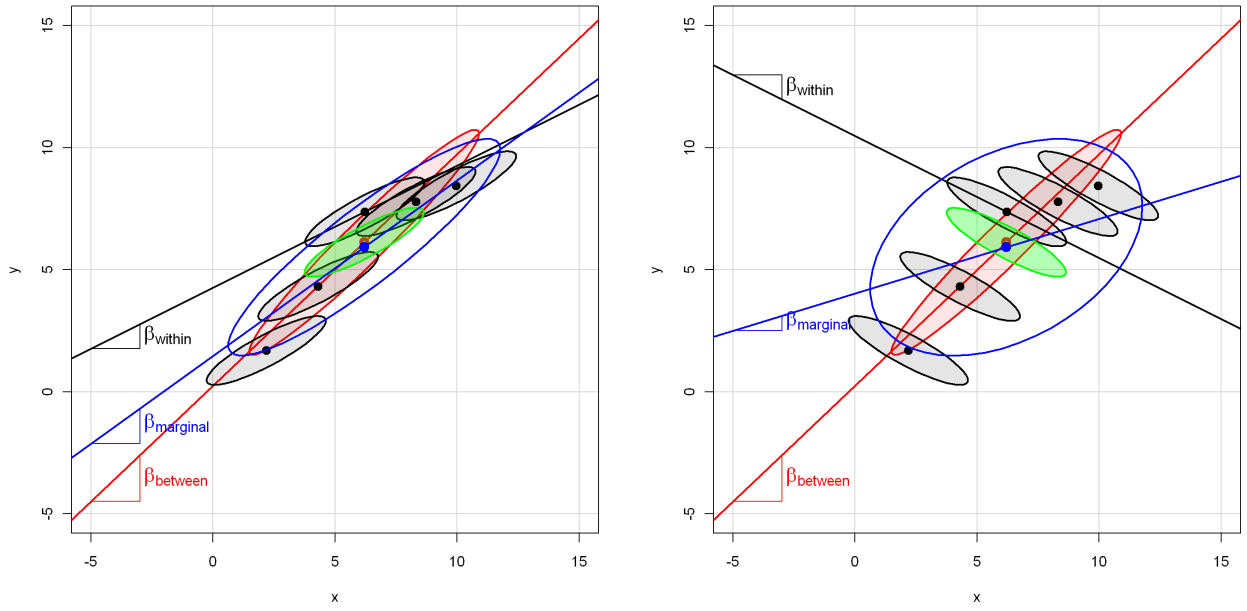

Figure 7: Paradoxes and fallacies: between (ecological), within (conditional) and whole-sample (marginal) associations. In both panels, the five groups have the same group means, and $\text{Var}(x) = 6$ and $\text{Var}(y) = 2$ within each group. The within-group correlation is $r = +0.87$ in all groups in the left panel, and is $r = -0.87$ in the right panel. The green ellipse shows the average within-group data ellipse.

distribution between a and b, and $\mathcal{N}(\mu, \sigma^2)$ represents the normal distribution with mean μ and variance σ^2. For simplicity, we have set the within-group covariance matrices to be identical in all groups, with $\text{Var}(x) = 6$, $\text{Var}(y) = 2$,and $\text{Cov}(x, y) = \pm 3$ in the left and right panels, respectively, giving $r_{\text{within}} = \pm 0.87$.

In the left panel, the conditional, within-group slope is smaller than the ecological, between-group slope, reflecting the smaller within-group than between-group correlation. In general, however, it can be shown that

$$\boldsymbol{\beta}_{\text{marginal}} \in [\boldsymbol{\beta}_{\text{within}}, \boldsymbol{\beta}_{\text{between}}] \ ,$$

which is also evident in the right panel, where the within-group slope is negative. This result follows from the fact that the marginal data ellipse for the total sample has a shape that is a convex combination (weighted average) of the average within-group covariance of (x, y), shown by the green ellipse in Figure 7, and the covariance of the means $(\bar{x}_i, \bar{y}_i)$, shown by the red between-group ellipse. In fact, the between and within data ellipses in Figure 7 are just (a scaling of) the $\boldsymbol{H}$ and $\boldsymbol{E}$ ellipses in an hypothesis-error (HE) plot for the MANOVA model, $(x, y) \sim$ group, as will be developed in Section 5. See Figure 8 for a visual demonstration, using the same data as in Figure 7.

The right panels of Figure 7 and Figure 8 provide a prototypical illustration of Simpson's paradox, where β_{within} and β_{marginal} can have opposite signs. Underlying this is a more general *marginal fallacy* (requiring only substantively different estimates, but not necessarily different signs), that can occur when some important factor or covariate is unmeasured or has been ignored. The fallacy consists of estimating the unconditional or marginal relationship (β_{marginal}) and believing that it reflects the conditional relationship, or that those pesky "other" variables will somehow average out. In practice, the marginal fallacy probably occurs most often when one views a scatterplot matrix of $(y, x_1, x_2, \dots)$ and believes that the slopes of relationships in the separate panels reflect the pairwise conditional relationships with other variables controlled. In a regression context, the antidote to the marginal fallacy is the added-variable plot (described in Section 4.8), which displays the conditional relationship between the response and a predictor directly, controlling for all other predictors.

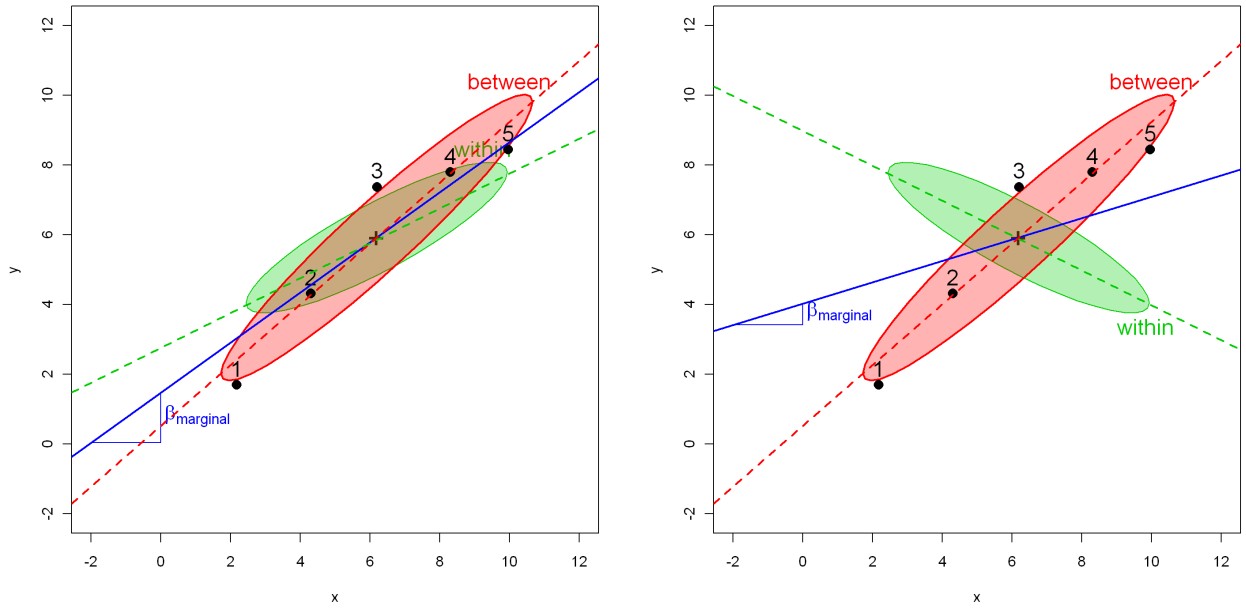

Figure 8: Visual demonstration that $\boldsymbol{\beta}_{\text{marginal}}$ lies between $\boldsymbol{\beta}_{\text{within}}$ and $\boldsymbol{\beta}_{\text{between}}$. Each panel shows an HE plot for the MANOVA model $(x, y) \sim$ group, in which the within and between ellipses are identical to those in Figure 7, except for scale.

The right panels of Figure 7 and Figure 8 also illustrate Robinson's paradox (Robinson, 1950), where β_{within} and β_{between} can have opposite signs.[7] The more general *ecological fallacy* (e.g., Lichtman, 1974, Kramer, 1983) is to draw conclusions from aggregated data, estimating β_{between} or r_{between}, believing that they reflect relationships at the individual level, estimating β_{within} or r_{within}. Perhaps the earliest instance of this was André-Michel Guerry's (1833) use of thematic maps of France depicting rates of literacy, crime, suicide, and other "moral statistics" by department to argue about the relationships of these moral variables as if they reflected individual behavior.[8] As can be seen in Figure 7, the ecological fallacy can often be resolved by accounting for some confounding variable(s) that vary between groups.

Finally, there are situations where only a subset of the relevant data are available (e.g., one group in Figure 7), or when the relevant data are available only at the individual level, so that only the conditional relationship, β_{within} can be estimated. The *atomistic fallacy* (also called the *fallacy of composition* or the *individualistic fallacy*), e.g., Alker (1969), Riley (1963), is the inverse to the ecological fallacy, and consists of believing that one can draw conclusions about the ecological relationship, β_{between}, from the conditional one.

The atomistic fallacy occurs most often in the context of multilevel models (Diez-Roux, 1998) where it is desired to draw inferences regarding variability of higher-level units (states, countries) from data collected from lower-level units. For example, imagine that the right panel of Figure 7 depicts the negative relationship of mortality from heart disease (y) with individual income (x) for individuals within countries. It would be fallacious to infer that the same slope (or even its sign) applies to a between-country analysis of heart disease mortality vs. GNP per capita. A positive value of β_{between} in this context might result from the fact that, across countries, higher GNP per capita is associated with less healthy diet (more fast food, red meat, larger

[7] William Robinson (1950) examined the relationship between literacy rate and percentage of foreign-born immigrants in the U.S. states from the 1930 Census. He showed that there was a surprising positive correlation, $r_{\text{between}} = 0.526$ at the state level, suggesting that foreign birth was associated with greater literacy; at the individual level, the correlation r_{within} was -0.118, suggesting the opposite. An explanation for the paradox was that immigrants tended to settle in regions of greater than average literacy.

[8] Guerry was certainly aware of the logical problem of ecological inference, at least in general terms (Friendly, 2007a), and carried out several side-analyses to examine potential confounding variables.

portions), leading to increased heart disease.

4.5 Leverage, influence, *and* precision

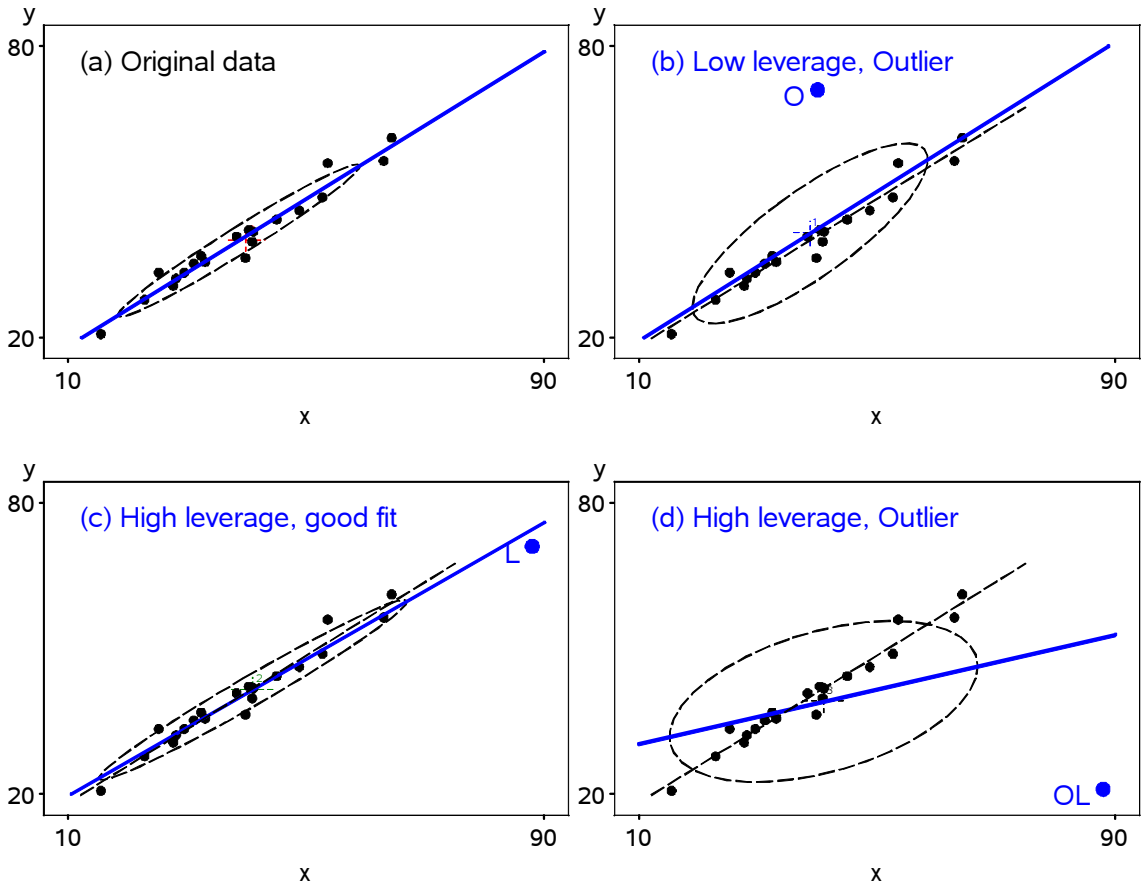

Figure 9: Leverage-Influence quartet with data ellipses. (a) Original data; (b) adding one low-leverage outlier (O); (c) adding one "good" leverage point (L); (d) adding one "bad" leverage point (OL). In panels (b)–(d) the dashed black line is the fitted line for the original data, while the thick solid blue line reflects the regression including the additional point. The data ellipses show the effect of the additional point on precision.

The topic of leverage and influence in regression is often introduced with graphs similar to Figure 9, what we call the "leverage-influence quartet." In these graphs, a bivariate sample of $n = 20$ points was first generated with $x \sim \mathcal{N}(40, 10^2)$ and $y \sim 10 + 0.75x + \mathcal{N}(0, 2.5^2)$. Then, in each of panels (b)–(d) a single point was added at the locations shown, to represent, respectively, a low-leverage point with a large residual,[9] a high-leverage point with small residual (a "good" leverage point), and a high-leverage point with large residual (a "bad" leverage point). The goal is to visualize how leverage [$\propto (x - \bar{x})^2$] and residual ($y - \hat{y}^*$) (where $\hat{y}_i^*$ is the fitted value for observation i, computed on the basis of an auxiliary regression in which observation i is deleted) combine to produce influential points—those that affect the estimates of $\boldsymbol{\beta} = (\beta_0, \beta_1)^\mathsf{T}$.

The "standard" version of this graph shows *only* the fitted regression lines for each panel. So, for the moment, ignore the data ellipses in the plots. The canonical, first-moment-only, story behind the standard

[9] In this context, a residual is "large" when the point in question deviates substantially from the regression line for the rest of the data—what is sometimes termed a "deleted residual"; see below.

version is that the points added in panels (b) and (c) are not harmful—the fitted line does not change very much when these additional points are included. Only the bad leverage point, "OL," in panel (d) is harmful.

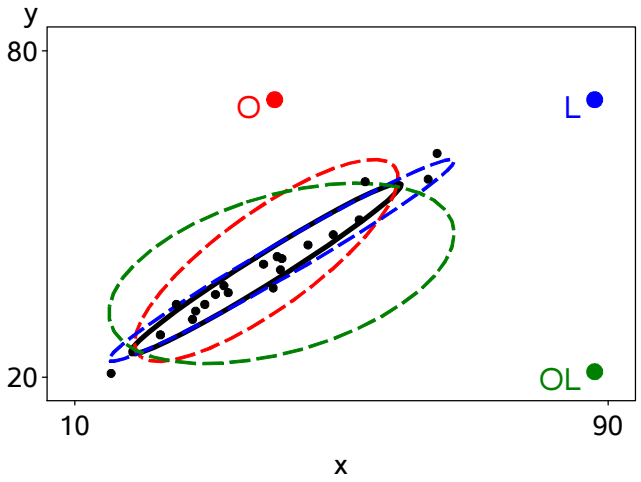

Figure 10: Data ellipses in the Leverage-Influence quartet. This graph overlays the data ellipses and additional points from the four panels of Figure 9. It can be seen that only the OL point affects the slope, while the O and L points affect precision of the estimates in opposite directions.

Adding the data ellipses to each panel immediately makes it clear that there is a second-moment part to the the story—the effect of unusual points on the *precision* of our estimates of $\boldsymbol{\beta}$. Now, we see *directly* that there is a big difference in impact between the low-leverage outlier [panel (b)] and the high-leverage, small-residual case [panel (c)], even though their effect on coefficient estimates is negligible. In panel (b), the single outlier inflates the estimate of residual variance (the size of the vertical slice of the data ellipse at $\bar{x}$).

To make the added value of the data ellipse more apparent, we overlay the data ellipses from Figure 9 in a single graph, shown in Figure 10, to allow direct comparison. Because you now know that regression lines can be visually estimated as the locus of vertical tangents, we suppress these lines in the plot to focus on precision. Here, we can also see why the high-leverage point "L" [added in panel (c) of Figure 9] is called a "good leverage point." By increasing the standard deviation of x, it makes the data ellipse somewhat more elongated, giving increased precision of our estimates of $\boldsymbol{\beta}$.

Whether a "good" leverage point is *really* good depends upon our faith in the regression model (and in the point), and may be regarded either as increasing the precision of $\hat{\boldsymbol{\beta}}$ or providing an illusion of precision. In either case, the data ellipse for the modified data shows the effect on precision directly.

4.6 Ellipsoids in data space and $\boldsymbol{\beta}$ space

It is most common to look at data and fitted models in "data space," where axes correspond to variables, points represent observations, and fitted models are plotted as lines (or planes) in this space. As we've suggested, data ellipsoids provide informative summaries of relationships in data space. For linear models, particularly regression models with quantitative predictors, there is another space—"$\boldsymbol{\beta}$ space"—that provides deeper views of models and the relationships among them. In $\boldsymbol{\beta}$ space, the axes pertain to coefficients and points are models (true, hypothesized, fitted) whose coordinates represent values of parameters.

In the sense described below, data space and $\boldsymbol{\beta}$ space are *dual* to each other. In simple linear regression, for example, each line in data space corresponds to a point in $\boldsymbol{\beta}$ space, the set of points on any line in $\boldsymbol{\beta}$ space corresponds to a pencil of lines through a given point in data space, and the proposition that every pair of points defines a line in one space corresponds to the proposition that every two lines intersect in a point in the other space.

Moreover, ellipsoids in these spaces are dual and inversely related to each other. In data space, joint confidence intervals for the mean vector or joint prediction regions for the data are given by the ellipsoids $(\bar{x}_1, \bar{x}_2)^\mathsf{T} \oplus c\sqrt{\boldsymbol{S}}$. In the dual $\boldsymbol{\beta}$ space, joint confidence regions for the parameters are given by ellipsoids of the form $\widehat{\boldsymbol{\beta}} \oplus c\sqrt{\boldsymbol{S}^{-1}}$. We illustrate these relationships in the example below.

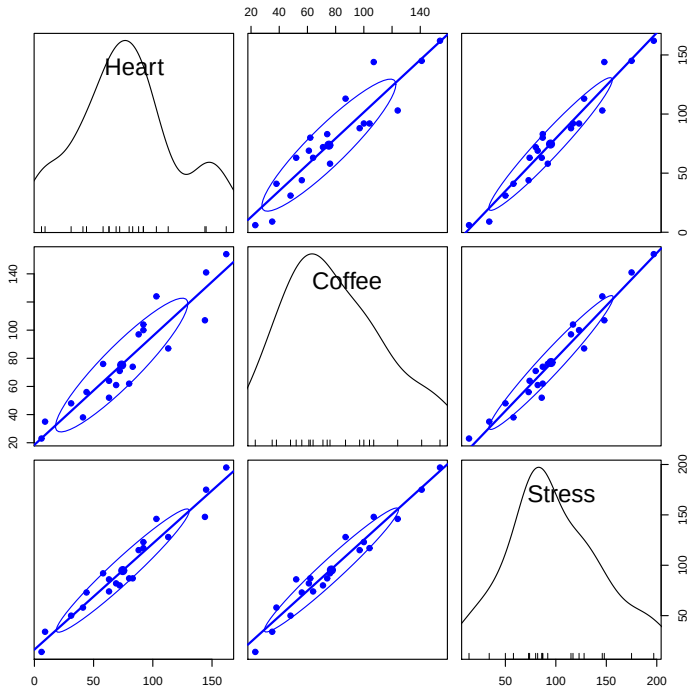

Figure 11: Scatterplot matrix, showing the pairwise relationships among Heart (y), Coffee (x_1), and Stress (x_2), with linear regression lines and 68% data ellipses for the marginal bivariate relationships.

Figure 11 shows a scatterplot matrix among the variables Heart (y), an index of cardiac damage, Coffee (x_1), a measure of daily coffee consumption, and Stress (x_2), a measure of occupational stress, in a contrived sample of $n = 20$. For the sake of the example we assume that the main goal is to determine whether or not coffee is good or bad for your heart, and stress represents one potential confounding variable among others (age, smoking, etc.) that might be useful to control statistically.

The plot in Figure 11 shows only the marginal relationship between each pair of variables. The marginal message seems to be that coffee is bad for your heart, stress is bad for your heart and coffee consumption is also related to occupational stress. Yet, when we fit both variables together, we obtain the following results, suggesting that coffee is good for you (the coefficient for coffee is now negative, though non-significant). How can this be?

	Estimate ($\widehat{\beta}$)	Std. Error	t value	Pr($>\|t\|$)
Intercept	-7.7943	5.7927	-1.35	0.1961
Coffee	-0.4091	0.2918	-1.40	0.1789
Stress	1.1993	0.2244	5.34	0.0001

Figure 12 shows the relationship between the predictors in data space and how this translates into joint and individual confidence intervals for the coefficients in $\boldsymbol{\beta}$ space. The left panel is the same as the corresponding (Coffee, Stress) panel in Figure 11, but with a standard (40%) data ellipse. The right panel shows the joint

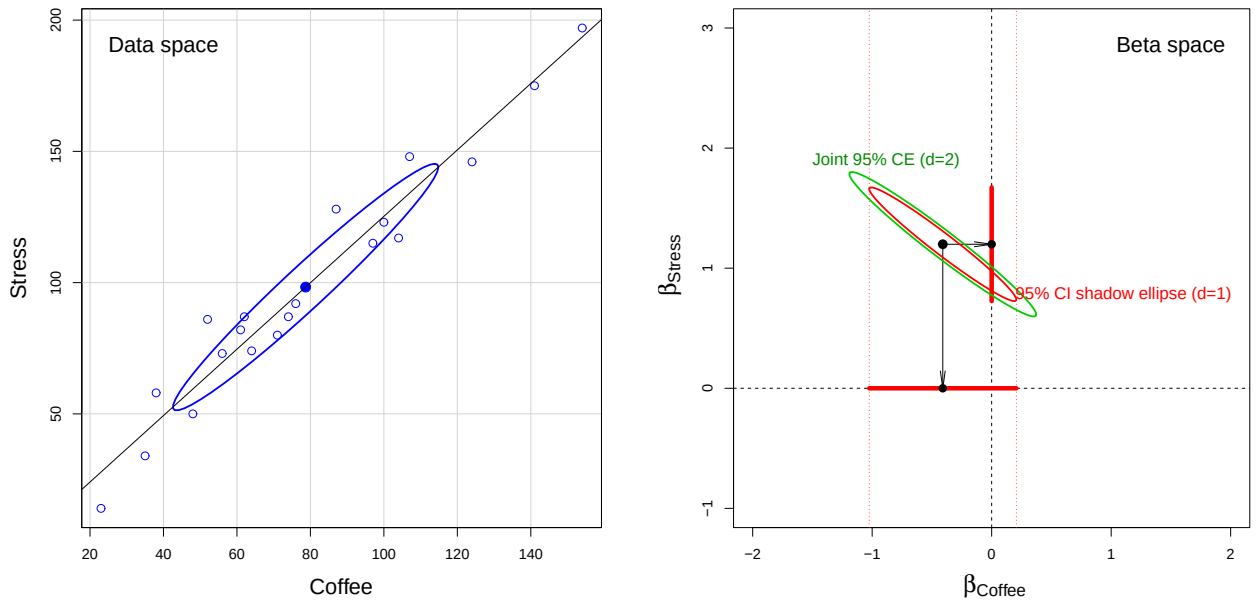

Figure 12: Data space and $\boldsymbol{\beta}$ space representations of Coffee and Stress. Left: Standard (40%) data ellipse. Right: Joint 95% confidence ellipse (green) for $(\beta_{\text{Coffee}}, \beta_{\text{Stress}})$, CI ellipse (red) with 95% univariate shadows.

95% confidence region and the individual 95% confidence intervals in $\boldsymbol{\beta}$ space, determined as

$$\widehat{\boldsymbol{\beta}} \oplus \sqrt{d F^{.95}_{q,\nu}} \times s_e \times \boldsymbol{S}_X^{-1/2}$$

where d is the number of dimensions for which we want coverage, ν is the residual degrees of freedom for s_e, and $\boldsymbol{S}_X$ is the covariance matrix of the predictors.

Thus, the green ellipse in Figure 12 is the ellipse of joint 95% coverage, using the factor $\sqrt{2F^{.95}_{2,\nu}}$ and covering the true values of $(\beta_{\text{Stress}}, \beta_{\text{Coffee}})$ in 95% of samples. Moreover:

- Any *joint* hypothesis (e.g., $H_0 : \beta_{\text{Stress}} = 1, \beta_{\text{Coffee}} = 1$) can be tested visually, simply by observing whether the hypothesized point, $(1, 1)$ here, lies inside or outside the joint confidence ellipse.
- The shadows of this ellipse on the horizontal and vertical axes give Scheffé joint 95% confidence intervals for the parameters, with protection for simultaneous inference ("fishing") in a 2-dimensional space.
- Similarly, using the factor $\sqrt{F^{1-\alpha/d}_{1,\nu}} = t^{1-\alpha/2d}_{\nu}$ would give an ellipse whose 1D shadows are $1 - \alpha$ Bonferroni confidence intervals for d posterior hypotheses.

Visual hypothesis tests and $d = 1$ confidence intervals for the parameters *separately* are obtained from the red ellipse in Figure 12, which is scaled by $\sqrt{F^{.95}_{1,\nu}} = t^{.975}_{\nu}$. We call this the "confidence-interval generating ellipse" (or, more compactly, the "confidence-interval ellipse"). The shadows of the confidence-interval ellipse on the axes (thick red lines) give the corresponding individual 95% confidence intervals, which are equivalent to the (partial, Type III) t-tests for each coefficient given in the standard multiple regression output shown above. Thus, controlling for Stress, the confidence interval for the slope for Coffee includes 0, so we cannot reject the hypothesis that $\beta_{\text{Coffee}} = 0$ in the multiple regression model, as we saw above in the numerical output. On the other hand, the interval for the slope for Stress excludes the origin, so we reject the null hypothesis that $\beta_{\text{Stress}} = 0$, controlling for Coffee consumption.

Finally, consider the relationship between the data ellipse and the confidence ellipse. These have exactly the same shape, but the confidence ellipse is exactly a 90^o rotation and rescaling of the data ellipse. In direc-

tions in data space where the data ellipse is wide—where we have more information about the relationship between Coffee and Stress—the confidence ellipse is narrow, reflecting greater precision of the estimates of coefficients. Conversely, where the data ellipse is narrow (less information), the confidence ellipse is wide (less precision). See Figure A.2 for the underlying geometry.

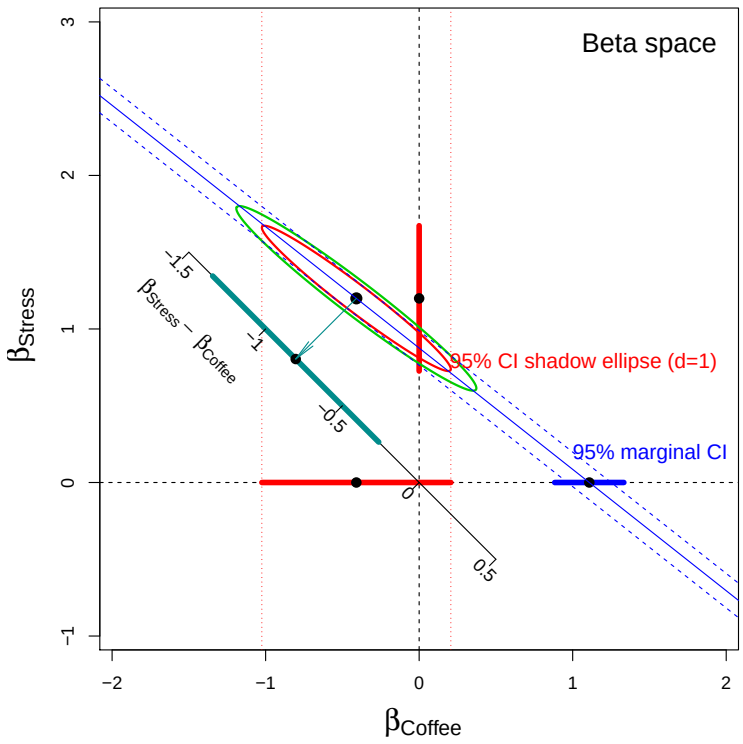

Figure 13: Joint 95% confidence ellipse for $(\beta_{\text{Coffee}}, \beta_{\text{Stress}})$, together with the 1D marginal confidence interval for β_{Coffee} ignoring Stress (thick blue line), and a visual confidence interval for $\beta_{\text{Stress}} - \beta_{\text{Coffee}} = 0$ (dark cyan).

The virtues of the confidence ellipse for visualizing hypothesis tests and interval estimates do not end here. Say we wanted to test the hypothesis that Coffee was unrelated to Heart damage in the *simple* regression ignoring Stress. The (Heart, Coffee) panel in Figure 11 showed the strong marginal relationship between the variables. This can be seen in Figure 13 as the oblique projection of the confidence ellipse to the horizontal axis where $\beta_{\text{Stress}} = 0$. The estimated slope for Coffee in the simple regression is exactly the oblique shadow of the center of the ellipse $(\widehat{\beta}_{\text{Coffee}}, \widehat{\beta}_{\text{Stress}})$ through the point where the ellipse has a horizontal tangent onto the horizontal axis at $\beta_{\text{Stress}} = 0$. The thick blue line in this figure shows the confidence interval for the slope for Coffee in the simple regression model. The confidence interval doesn't cover the origin, so we reject $H_0 : \beta_{\text{Coffee}} = 0$ in the simple regression model. The oblique shadow of the red 95% confidence-interval ellipse onto the horizontal axis is slightly smaller. How much smaller is a function of the size of the coefficient for Stress.

We can go further. As we noted earlier, all linear combinations of variables or parameters in data or models correspond graphically to projections (shadows) onto certain sub-spaces. Let's assume that Coffee and Stress were measured on the same scales so it makes sense to ask if they have equal impacts on Heart disease in the joint model that includes them both. Figure 13 also shows an auxiliary axis through the origin with slope $= -1$ corresponding to values of $\beta_{\text{Stress}} - \beta_{\text{Coffee}}$. The orthogonal projection of the coefficient vector on this axis is the point estimate of $\widehat{\beta}_{\text{Stress}} - \widehat{\beta}_{\text{Coffee}}$ and the shadow of the red ellipse along this axis is the 95% confidence interval for the difference in slopes. This interval excludes 0, so we would reject the

hypothesis that Coffee and Stress have equal coefficients.

4.7 Measurement error

In classical linear models, the predictors are often considered to be fixed variables, or, if random, to be measured without error and independent of the regression errors; either condition, along with the assumption of linearity, guarantees unbiasedness of the standard OLS estimators. In practice, of course, predictor variables are often also observed indicators, subject to error, a fact that is recognized in errors-in-variables regression models and in more general structural equation models but often ignored otherwise. Ellipsoids in data space and β space are well suited to showing the effect of measurement error in predictors on OLS estimates.

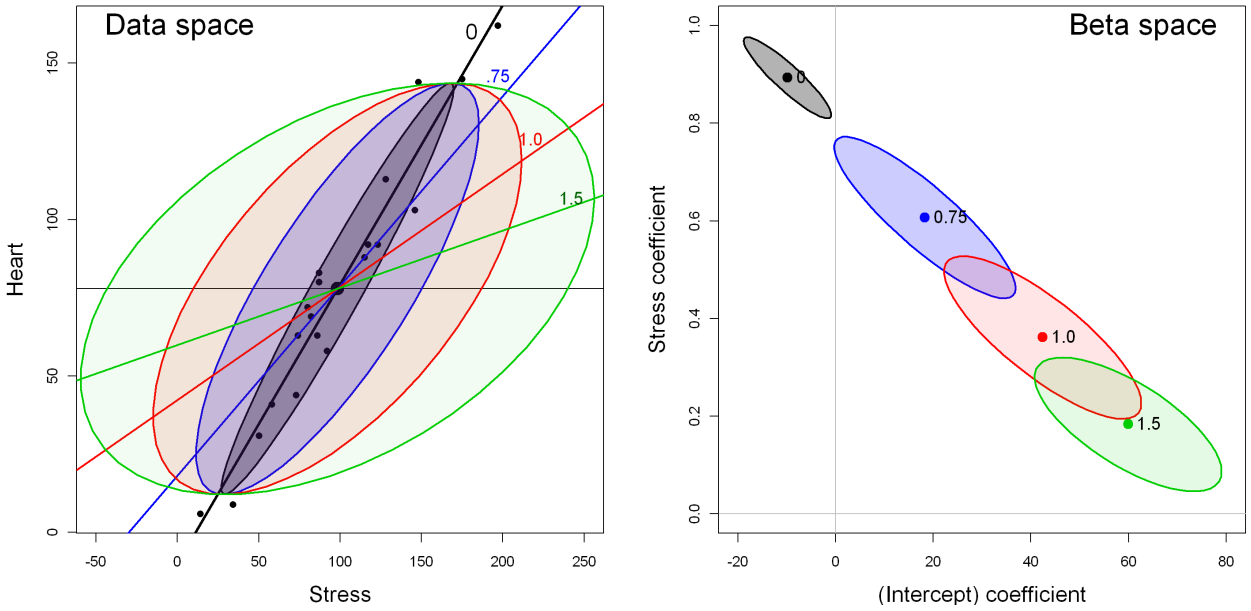

Figure 14: Effects of measurement error in Stress on the marginal relationship between Heart disease and Stress. Each panel starts with the observed data ($\delta = 0$), then adds random normal error, $\mathcal{N}(0, \delta \times \text{SD}_{Stress})$, with $\delta = \{0.75, 1.0, 1.5\}$, to the value of Stress. Increasing measurement error biases the slope for Stress toward 0. Left: 50% data ellipses; right: 50% confidence ellipses for $(\beta_0, \beta_{Stress})$.

The statistical facts are well known, though perhaps counter-intuitive in certain details: measurement error in a predictor biases regression coefficients, while error in the measurement in y increases the standard errors of the regression coefficients but does not introduce bias.

In the top row of Figure 11, adding measurement error to the Heart disease variable would expand the data ellipses vertically, but (apart from random variation) leave the slopes of the regression lines unchanged. Measurement error in a predictor variable, however, biases the corresponding estimated coefficient toward zero (sometimes called *regression attenuation*) as well as increasing standard errors.

Figure 14 demonstrates this effect for the marginal relation between Heart disease and stress, with data ellipses in data space and the corresponding confidence ellipses in β space. Each panel starts with the observed data (the darkest ellipse, marked 0), then adds random normal error, $\mathcal{N}(0, \delta \times \text{SD}_{Stress})$, with $\delta = \{0.75, 1.0, 1.5\}$, to the value of Stess, while keeping the mean of Stress the same. All of the data ellipses have the same vertical shadows (SD_{Heart}), while the horizontal shadows increase with δ, driving the slope for Stress toward 0. In β space, it can be seen that the estimated coefficients, $(\beta_0, \beta_{\text{Stress}})$ vary along a line and would reach $\beta_{\text{Stress}} = 0$ for δ sufficiently large. The vertical shadows of ellipses for $(\beta_0, \beta_{\text{Stress}})$ along the β_{Stress} axis also demonstrate the effects of measurement error on the standard error of β_{Stress}.

Perhaps less well-known, but both more surprising and interesting, is the effect that measurement error in one variable, x_1, has on the estimate of the coefficient for an *other* variable, x_2, in a multiple regression model. Figure 15 shows the confidence ellipses for $(\beta_{\text{Coffee}}, \beta_{\text{Stress}})$ in the multiple regression predicting Heart disease, adding random normal error $\mathcal{N}(0, \delta \times \text{SD}_{Stress})$, with $\delta = \{0, 0.2, 0.4, 0.8\}$, to the value of Stress alone. As can be plainly seen, while this measurement error in Stress attenuates its coefficient, it also has the effect of biasing the coefficient for Coffee toward that in the *marginal* regression of Heart disease on Coffee alone.

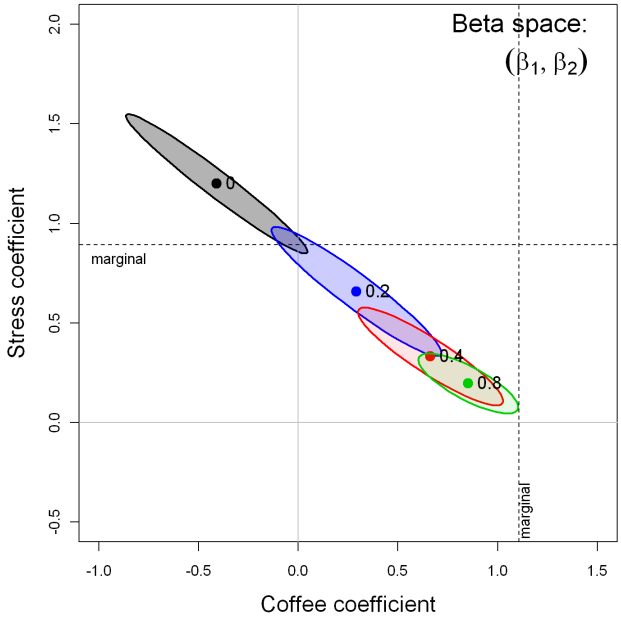

Figure 15: Biasing effect of measurement error in one variable (Stress) on on the coefficient of another variable (Coffee) in a multiple regression. The coefficient for Coffee is driven towards its value in the marginal model using Coffee alone, as measurement error in Stress makes it less informative in the joint model.

4.8 Ellipsoids in added-variable plots

In contrast to the marginal, bivariate views of the relationships of a response to several predictors (e.g., such as shown in the top row of the scatterplot matrix in Figure 11), *added-variable plots* (aka *partial regression plots*) show the partial relationship between the response and each predictor, where the effects of all other predictors have been controlled or adjusted for. Again we find that such plots have remarkable geometric properties, particularly when supplemented by ellipsoids.

Formally, we express the standard linear model in vector form as $\widehat{\boldsymbol{y}} \equiv \widehat{\boldsymbol{y}} \,|\, \boldsymbol{X} = \beta_0 \mathbf{1} + \beta_1 \boldsymbol{x}_1 + \beta_2 \boldsymbol{x}_2 + \cdots + \beta_p \boldsymbol{x}_p$, with model matrix $\boldsymbol{X} = [\mathbf{1}, \boldsymbol{x}_1, \ldots \boldsymbol{x}_p]$. Let $\boldsymbol{X}_{[-k]}$ be the model matrix omitting the column for variable k. Then algebraically, the added variable plot for variable k is the scatterplot of the residuals $(\boldsymbol{x}_k^\star, \boldsymbol{y}^\star)$ from two auxillary regressions,[10] fitting $\boldsymbol{y}$ and $\boldsymbol{x}_k$ from $\boldsymbol{X}_{[-k]}$,

$$\begin{aligned} \boldsymbol{y}^\star &\equiv \boldsymbol{y} \,|\, \text{others} = \boldsymbol{y} - \widehat{\boldsymbol{y}} \,|\, \boldsymbol{X}_{[-k]} \\ \boldsymbol{x}_k^\star &\equiv \boldsymbol{x}_k \,|\, \text{others} = \boldsymbol{x}_k - \widehat{\boldsymbol{x}}_k \,|\, \boldsymbol{X}_{[-k]} \quad . \end{aligned}$$

[10] These quantities can all be computed (Velleman and Welsh, 1981) from the results of a single regression for the full model.

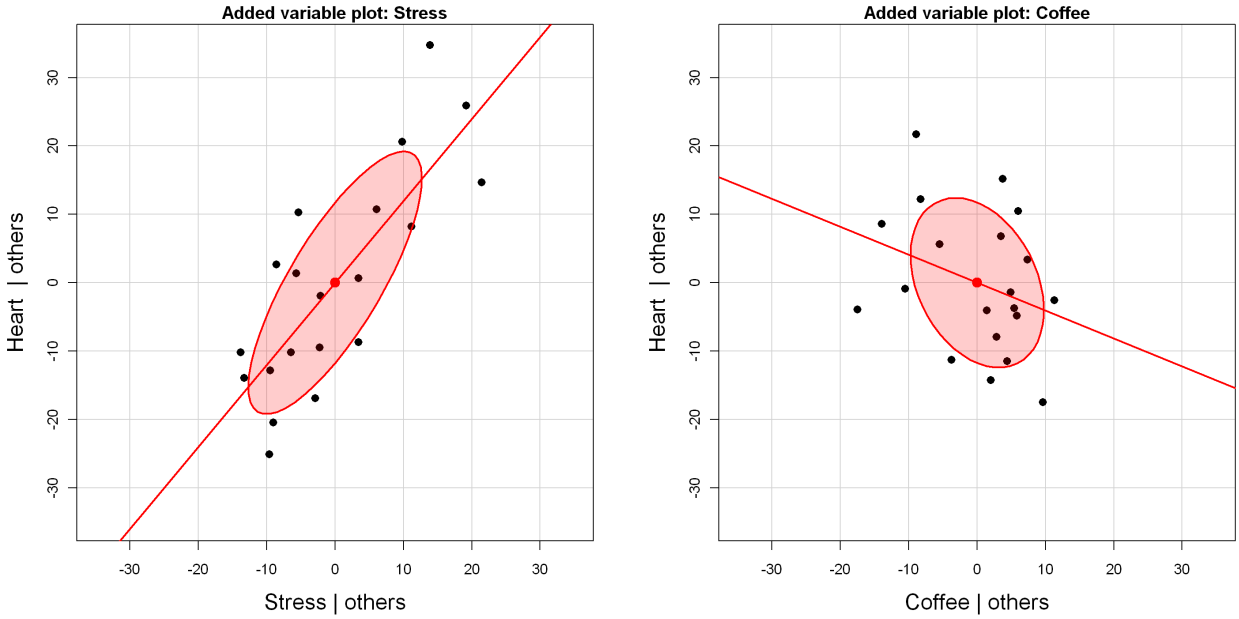

Figure 16: Added variable plots for Stress and Coffee in the multiple regression predicting Heart disease. Each panel also shows the 50% conditional data ellipse for residuals $(\boldsymbol{x}_k^\star, \boldsymbol{y}^\star)$, shaded red.

Geometrically, in the space of the observations,[11] the fitted vector $\widehat{\boldsymbol{y}}$ is the orthogonal projection of $\boldsymbol{y}$ onto the subspace spanned by $\boldsymbol{X}$. Then $\boldsymbol{y}^\star$ and $\boldsymbol{x}_k^\star$ are the projections onto the orthogonal complement of the subspace spanned by $\boldsymbol{X}_{[-k]}$, so the simple regression of $\boldsymbol{y}^\star$ on $\boldsymbol{x}_k^\star$ has slope $\hat{\beta}_k$ in the full model, and the residuals from the line $\widehat{\boldsymbol{y}}^\star = \hat{\beta}_k \boldsymbol{x}_k^\star$ in this plot are identically the residuals from the overall regression of $\boldsymbol{y}$ on $\boldsymbol{X}$.

Another way to describe the added-variable plot (AVP) for x_k is as a 2D projection of the space of $(\boldsymbol{y}, \boldsymbol{X})$, viewed in the plane defined by the intersection of two hyperplanes: the plane of the regression of $\boldsymbol{y}$ on all of $\boldsymbol{X}$, and the plane of regression of $\boldsymbol{y}$ on $\boldsymbol{X}_{[-k]}$. A third plane, that of the regression of x_k on $\boldsymbol{X}_{[-k]}$ also intersects in this space, and defines the horizontal axis in the AVP. This is illustrated in Figure 17, showing one view defined by the intersection of the three planes in the right panel.[12]

Figure 16 shows added-variable plots for Stress and Coffee in the multiple regression predicting Heart disease, supplemented by data ellipses for the residuals $(\boldsymbol{x}_k^\star, \boldsymbol{y}^\star)$. With reference to the properties of data ellipses in marginal scatterplots (see Figure 4), the following visual properties (among others) are useful in this discussion. These results follow simply from translating "marginal" into "conditional" (or "partial") in the present context. The essential idea is that the data ellipse of the AVP for $(x_k\star, y\star)$ is to the estimate of a coefficient in a multiple regression as the data ellipse of (x, y) is to simple regression. Thus:

1. The simple regression least squares fit of $\boldsymbol{y}^\star$ on $\boldsymbol{x}_k^\star$ has slope $\hat{\beta}_k$, the partial slope for x_k in the full model (and intercept = 0).
2. The residuals, $(\boldsymbol{y}^\star - \widehat{\boldsymbol{y}}^\star)$, shown in this plot are the residuals for $\boldsymbol{y}$ in the full model.
3. The correlation between $\boldsymbol{x}_k^\star$ and $\boldsymbol{y}^\star$, seen in the shape of the data ellipse for these variables, is the partial correlation between y and x_k with the other predictors in $\boldsymbol{X}_{[-k]}$ partialled out.
4. The horizontal half-width of the AVP data ellipse is proportional to the conditional standard deviation of x_k remaining after all other predictors have been accounted for, providing a visual interpretation of

[11] The "space of the observations" is yet a third, n-dimensional, space, in which the observations are the axes and each variable is represented as a point (or vector). See, e.g., Fox (2008, Ch. 10).

[12] Animated 3D movies of this plot are included among the supplementary materials for this paper.

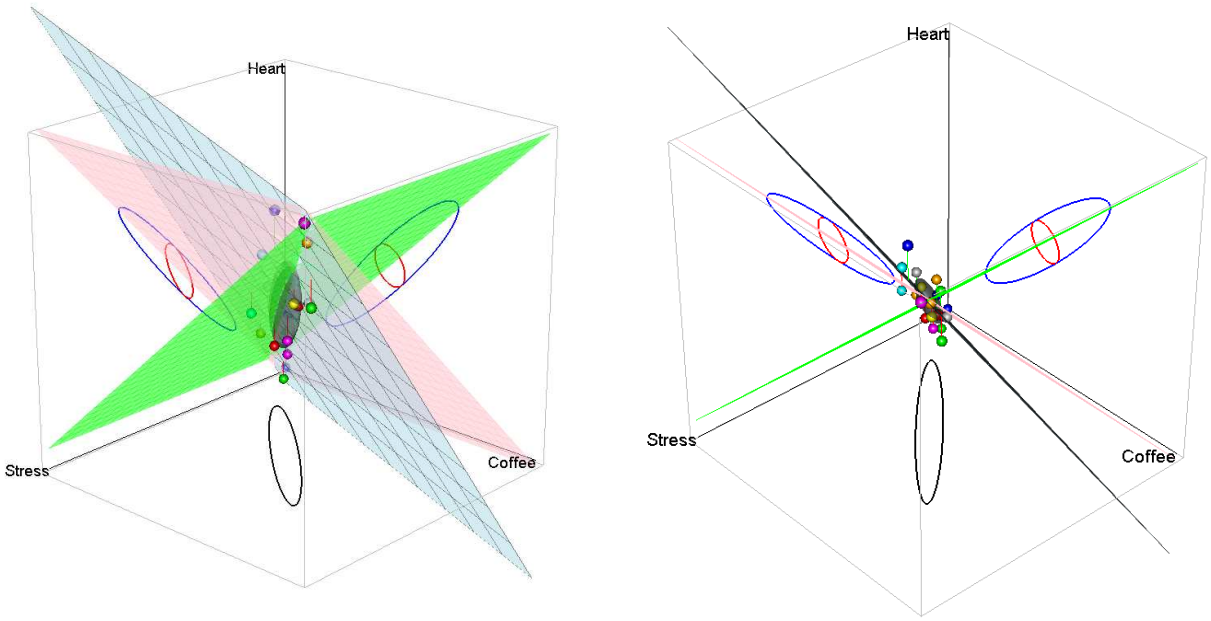

Figure 17: 3D views of the relationship between Heart, Coffee and Stress, showing the three regression planes for the marginal models, Heart $\sim$ Coffee (green), Heart $\sim$ Stress (pink), and the joint model, Heart $\sim$ Coffee + Stress (light blue). Left: a standard view; right: a view showing all three regression planes on edge. The ellipses in the side panels are 2D projections of the standard conditional (red) and marginal (blue) ellipsoids, as shown in Figure 18.

variance inflation due to collinear predictors, as we describe below.

5. The vertical half-width of the data ellipse is proportional to the residual standard deviation s_e in the multiple regression.
6. The squared horizontal positions, $(\boldsymbol{x}_k^\star)^2$, in the plot give the partial contributions to leverage on the coefficient $\hat{\beta}_k$ of x_k.
7. Items (3) and (7) imply that that the AVP for x_k shows the *partial* influence of individual observations on the coefficient $\hat{\beta}_k$, in the same way as in Figure 9 for marginal models. These influence statistics are often shown numerically as DFBETA statistics (Belsley *et al.*, 1980).
8. The last three items imply that the collection of added-variable plots for $\boldsymbol{y}$ and $\boldsymbol{X}$ provide an easy way to visualize the leverage and influence that individual observations—and indeed the joint influence of subsets of observations—have on the esimation of *each* coefficient in a given model.

Elliptical insight also permits us to go further, to depict the relationship between conditional and marginal views directly. Figure 18 shows the same added-variable plots for Heart disease on Stress and Coffee as in Figure 16 (with a zoomed-out scaling), but here we also overlay the marginal data ellipses for (x_k, y), and and marginal regression lines for Stress and Coffee separately. In 3D data space, these are the shadows (projections) of the data ellipsoid onto the planes defined by the partial variables. In 2D AVP space, they are just the marginal data ellipses translated to the origin.

The most obvious feature of Figure 18 is that the AVP for Coffee has a negative slope in the conditional plot (suggesting that controlling for Stress, coffee consumption is good for your heart), while in the marginal plot increasing coffee seems to be bad for your heart. This serves as a regression example of Simpson's paradox, which we considered earlier.

Less obvious is the fact that the marginal and AVP ellipses are easily visualized as a shadow versus a slice of the full data ellipsoid. Thus, the AVP ellipse must be contained in the marginal ellipse, as we can see

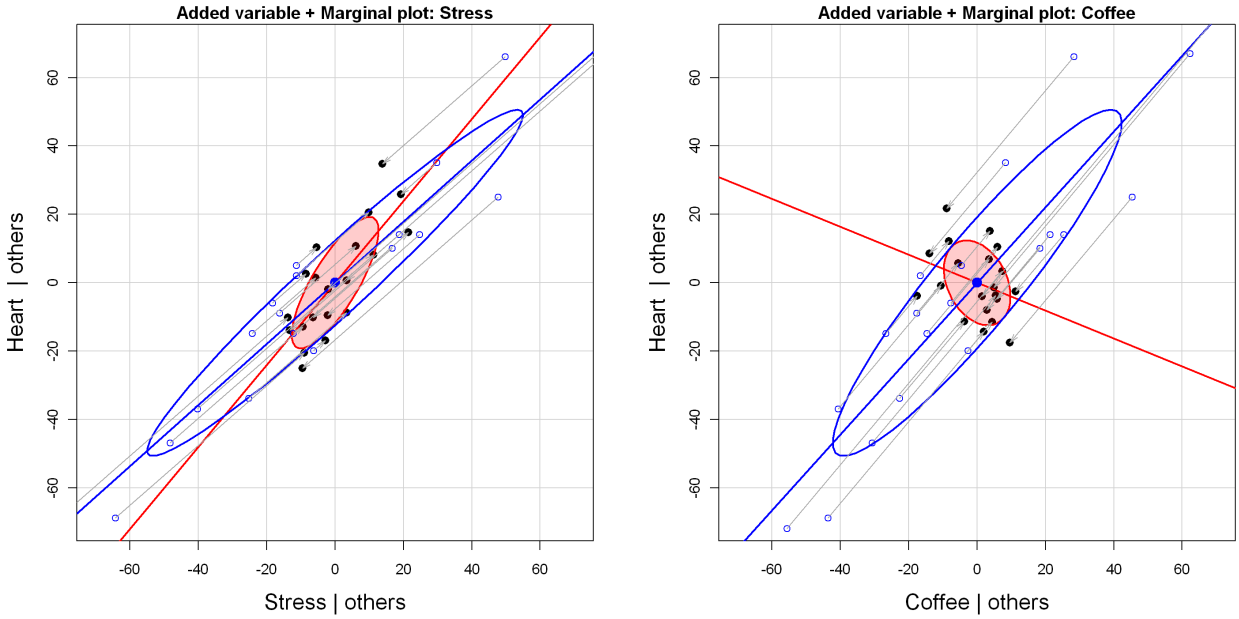

Figure 18: Added-variable + marginal plots for Stress and Coffee in the multiple regression predicting Heart disease. Each panel shows the 50% conditional data ellipse for $x_k^\star, y^\star$ residuals (shaded, red) as well as the marginal 50% data ellipse for the (x_k, y) variables, shifted to the origin. Arrows connect the mean-centered marginal points (open circles) to the residual points (filled circles).

in Figure 18. If there are only two xs, then the AVP ellipse must touch the marginal ellipse at two points. The shrinkage of the intersection of the AVP ellipse with the y axis represents improvement in fit due to other xs.

More importantly, the shrinkage of the width (projected onto a horizontal axis) represents the square root of the variance inflation factor (VIF), which can be shown to be the ratio of the horizontal width of the marginal ellipse of (x_k, y), with standard deviation $s(x_k)$ to the width of the conditional ellipse of $(x_k^\star, y^\star)$, with standard deviation $s(x_k \,|\, \text{others})$. This geometry implies interesting constraints among the three quantities: improvement in fit, VIF, and change from the marginal to conditional slope.

Finally, Figure 18 also shows how conditioning on other predictors works for individual observations, where each point of $(\boldsymbol{x}_k^\star, \boldsymbol{y}^\star)$ is the image of $(\boldsymbol{x}_k, \boldsymbol{y})$ along the path of the marginal regression. This reminds us that the AVP is a 2D projection of the full space, where the regression plane of $\boldsymbol{y}$ on $\boldsymbol{X}_{[-k]}$ becomes the vertical axis and the regression plane of $\boldsymbol{x}_k$ on $\boldsymbol{X}_{[-k]}$ becomes the horizontal axis.

5 Multivariate linear models: HE plots

Multivariate linear models (MvLMs) have a special affinity with ellipsoids and elliptical geometry, as described in this section. To set the stage and establish notation, we consider the MvLM (e.g., Timm (1975)) given by the equation $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{B} + \boldsymbol{U}$, where $\boldsymbol{Y}$ is an $n \times p$ matrix of responses in which each column represents a distinct response variable; $\boldsymbol{X}$ is the $n \times q$ model matrix of full column rank for the regressors; $\boldsymbol{B}$ is the $q \times p$ matrix of regression coefficients or model parameters; and $\boldsymbol{U}$ is the $n \times p$ matrix of errors, with $\text{vec}(\boldsymbol{U}) \sim \mathcal{N}_p(\boldsymbol{0}, \boldsymbol{I}_n \otimes \boldsymbol{\Sigma})$, where $\otimes$ is the Kronecker product.

A convenient feature of the MvLM for general multivariate responses is that *all* tests of linear hypotheses (for null effects) can be represented in the form of a general linear test,

$$H_0 : \underset{(h\times q)}{\boldsymbol{L}} \underset{(q\times p)}{\boldsymbol{B}} = \underset{(h\times p)}{\boldsymbol{0}} \;, \tag{11}$$

where $\boldsymbol{L}$ is a rank $h \leq q$ matrix of constants whose rows specify h linear combinations or contrasts of the parameters to be tested simultaneously by a multivariate test.

For *any* such hypothesis of the form given in Eqn. (11), the analogs of the univariate sums of squares for hypothesis (SS_H) and error (SS_E) are the $p \times p$ sum of squares and cross-products (SSP) matrices given by:

$$\boldsymbol{H} \equiv \boldsymbol{SSP}_H = (\boldsymbol{L}\widehat{\boldsymbol{B}})^{\mathsf{T}} \left[\boldsymbol{L}(\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X})^{-}\boldsymbol{L}^{\mathsf{T}}\right]^{-1} (\boldsymbol{L}\widehat{\boldsymbol{B}}) \ , \tag{12}$$

and

$$\boldsymbol{E} \equiv \boldsymbol{SSP}_E = \boldsymbol{Y}^{\mathsf{T}}\boldsymbol{Y} - \widehat{\boldsymbol{B}}^{\mathsf{T}}(\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X})\widehat{\boldsymbol{B}} = \widehat{\boldsymbol{U}}^{\mathsf{T}}\widehat{\boldsymbol{U}} \ , \tag{13}$$

where $\widehat{\boldsymbol{U}} = \boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{B}}$ is the matrix of residuals. Multivariate test statistics (Wilks's Λ, Pillai trace, Hotelling-Lawley trace, Roy's maximum root) for testing Eqn. (11) are based on the $s = \min(p, h)$ non-zero latent roots $\lambda_1 > \lambda_2 > \cdots > \lambda_s$ of the matrix $\boldsymbol{H}$ relative to the matrix $\boldsymbol{E}$, that is, the values of λ for which $\det(\boldsymbol{H} - \lambda\boldsymbol{E}) = 0$, or equivalently the latent roots ρ_i for which $\det[\boldsymbol{H} - \rho(\boldsymbol{H} + \boldsymbol{E})] = 0$. The details are shown in Table 1. These measures attempt to capture how "large" $\boldsymbol{H}$ is, relative to $\boldsymbol{E}$ in s dimensions, and correspond to various "means" as we described earlier. All of these statistics have transformations to F statistics giving either exact or approximate null-hypothesis F distributions. The corresponding latent vectors provide a set of s orthogonal linear combinations of the responses that produce maximal univariate F statistics for the hypothesis in Eqn. (11); we refer to these as the *canonical discriminant* dimensions.

Table 1: Multivariate test statistics as functions of the eigenvalues λ_i solving $\det(\boldsymbol{H} - \lambda\boldsymbol{E}) = 0$ or eigenvalues ρ_i solving $\det[\boldsymbol{H} - \rho(\boldsymbol{H} + \boldsymbol{E})] = 0$.

Criterion	Formula	"mean" of ρ	Partial η^2
Wilks's Λ	$\Lambda = \prod_i^s \frac{1}{1+\lambda_i} = \prod_i^s (1-\rho_i)$	geometric	$\eta^2 = 1 - \Lambda^{1/s}$
Pillai trace	$V = \sum_i^s \frac{\lambda_i}{1+\lambda_i} = \sum_i^s \rho_i$	arithmetic	$\eta^2 = \frac{V}{s}$
Hotelling-Lawley trace	$H = \sum_i^s \lambda_i = \sum_i^s \frac{\rho_i}{1-\rho_i}$	harmonic	$\eta^2 = \frac{H}{H+s}$
Roy maximum root	$R = \lambda_1 = \frac{\rho_1}{1-\rho_1}$	supremum	$\eta^2 = \frac{\lambda_1}{1+\lambda_1} = \rho_1$

Beyond the informal characterization of the four classical tests of hypotheses for multivariate linear models given in Table 1, there is an interesting geometrical representation that helps one to appreciate their relative power for various alternatives. This can be illustrated most simply in terms of the canonical representation, $(\boldsymbol{H} + \boldsymbol{E})^{\star}$, of the ellipsoid generated by $(\boldsymbol{H} + \boldsymbol{E})$ relative to $\boldsymbol{E}$, as shown in Figure 19 for $p = 2$.

With λ_i as described above, the eigenvalues and squared radii of $(\boldsymbol{H} + \boldsymbol{E})^{\star}$ are $\lambda_i + 1$, so the lengths of the major and minor axes are $a = \sqrt{\lambda_1 + 1}$ and $b = \sqrt{\lambda_2 + 1}$ respectively. The diagonal of the triangle comprising the segments a, b (labelled c) has length $c = \sqrt{a^2 + b^2}$. Finally, a line segment from the origin dropped perpendicularly to the diagonal joining the two ellipsoid axes is labelled d.

In these terms, Wilks's test, based on $\prod (1 + \lambda_i)^{-1}$ is equivalent to a test based on $a \times b$ which is proportional to the area of the framing rectangle, shown shaded in Figure 19. The Hotelling-Lawley trace test, based on $\sum \lambda_i$ is equivalent to a test based on $c = \sqrt{\sum \lambda_i + p}$. Finally, the Pillai Trace test, based on $\sum \lambda_i (1 + \lambda_i)^{-1}$ can be shown to be equal to $2 - d^{-2}$ for $p = 2$. Thus it is strictly monotone in d and equivalent to a test based directly on d.

The geometry makes it easy to see that if there is a large discrepancy between λ_1 and λ_2, Roy's test depends only on λ_1 while the Pillai test depends more on λ_2. Wilks's Λ and the Hotelling-Lawley trace criterion are also functional averages of λ_1 and λ_2, with the former being penalized when λ_2 is small. In practice, when $s \leq 2$, all four test criteria are equivalent, in that their standard transformations to F statistics are exact and give rise to identical p-values.

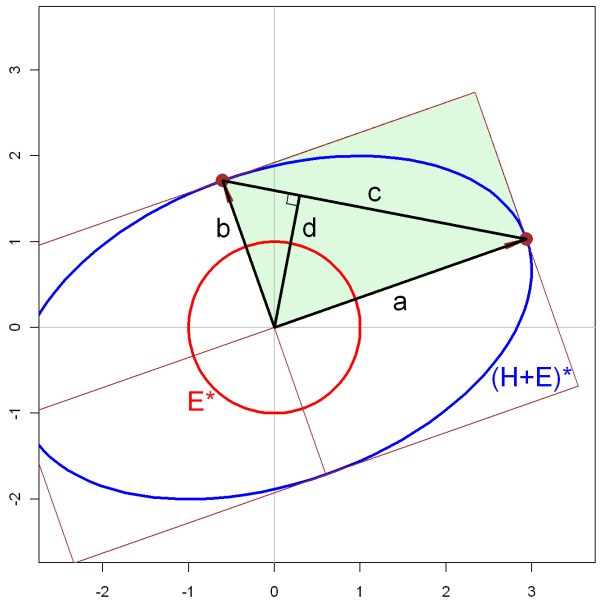

Figure 19: Geometry of the classical test statistics used in tests of hypotheses in multivariate linear models. The figure shows the representation of the ellipsoid generated by $(\boldsymbol{H} + \boldsymbol{E})$ relative to $\boldsymbol{E}$ in canonical space where $\boldsymbol{E}^\star = \boldsymbol{I}$ and $(\boldsymbol{H} + \boldsymbol{E})^\star$ is the corresponding transformation of $(\boldsymbol{H} + \boldsymbol{E})$.

5.1 Hypothesis-Error (HE) plots

The essential idea behind HE plots is that any multivariate hypothesis test, Eqn. (11), can be represented visually by ellipses (or ellipsoids beyond 2D) that express the size of covariation against a multivariate null hypothesis ($\boldsymbol{H}$) relative to error covariation ($\boldsymbol{E}$). The multivariate tests, based on the latent roots of $\boldsymbol{H}\boldsymbol{E}^{-1}$, are thus translated directly to the sizes of the $\boldsymbol{H}$ ellipses for various hypotheses, relative to the size of the $\boldsymbol{E}$ ellipse. Moreover, the shape and orientation of these ellipses show something more—the directions (linear combinations of the responses) that lead to various effect sizes and significance.

Figure 20 illustrates this idea for two variables from the iris dataset. Panel (a) shows the data ellipses for sepal length and petal length, equivalent to the corresponding plot in Figure 3. Panel (b) shows the HE plot for these variables from the one-way MANOVA model $\boldsymbol{y}_{ij} = \boldsymbol{\mu}_i + \boldsymbol{u}_{ij}$ testing equal mean vectors across species, $H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \boldsymbol{\mu}_3$. Let $\widehat{\boldsymbol{Y}}$ be the $n \times p$ matrix of fitted values for this model, i.e., $\widehat{\boldsymbol{Y}} = \{\bar{\boldsymbol{y}}_{i\cdot}\}$. Then $\boldsymbol{H} = \widehat{\boldsymbol{Y}}^\mathsf{T}\widehat{\boldsymbol{Y}} - n\bar{\boldsymbol{y}}\bar{\boldsymbol{y}}^\mathsf{T}$ (where $\bar{\boldsymbol{y}}$ is the grand-mean vector), and the $\boldsymbol{H}$ ellipse in the figure is then just the 2D projection of the data ellipsoid of the fitted values, scaled as described below. Similarly, $\widehat{\boldsymbol{U}} = \boldsymbol{Y} - \widehat{\boldsymbol{Y}}$, and $\boldsymbol{E} = \widehat{\boldsymbol{U}}^\mathsf{T}\widehat{\boldsymbol{U}} = (N-g)\boldsymbol{S}_{\text{pooled}}$, so the $\boldsymbol{E}$ ellipse is the 2D projection of the data ellipsoid of the residuals. Visually, the $\boldsymbol{E}$ ellipsoid corresponds to shifting the separate within-group data ellipsoids to the centroid, as illustrated above in Figure 6(c).

In HE plots, the $\boldsymbol{E}$ matrix is first scaled to a covariance matrix $\boldsymbol{E}/df_e$, dividing by the error degrees of freedom, df_e. The ellipsoid drawn is translated to the centroid $\overline{\boldsymbol{y}}$ of the variables, giving $\overline{\boldsymbol{y}} \oplus c\boldsymbol{E}^{1/2}/df_e$. This scaling and translation also allows the means for levels of the factors to be displayed in the same space, facilitating interpretation. In what follows, we show these as "standard" bivariate ellipses of 68% coverage, using $c = \sqrt{2F^{.68}_{2,df_e}}$, except where noted otherwise.

The ellipse for $\boldsymbol{H}$ reflects the size and orientation of covariation against the null hypothesis. In relation to the $\boldsymbol{E}$ ellipse, the $\boldsymbol{H}$ ellipse can be scaled to show either the *effect size* or strength of *evidence* against H_0 (significance).

For effect-size scaling, each $\boldsymbol{H}$ is divided by df_e to conform to $\boldsymbol{E}$. The resulting ellipse is then exactly

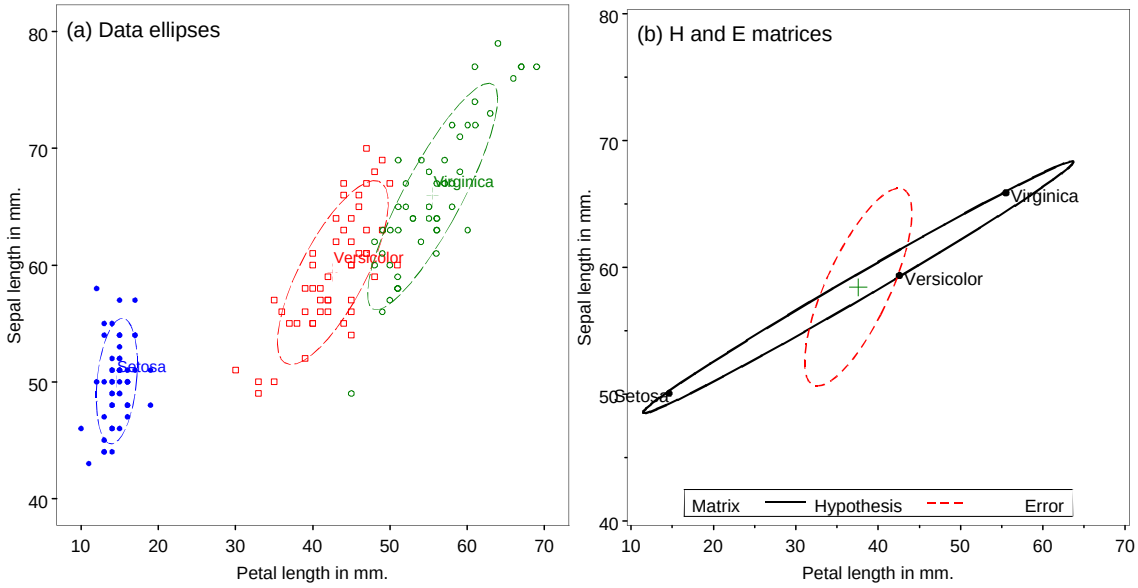

Figure 20: (a) Data ellipses and (b) corresponding HE plot for sepal length and petal length in the iris dataset. The $\boldsymbol{H}$ ellipse is the data ellipse of the fitted values defined by the group means, $\bar{\boldsymbol{y}}_{i\cdot}$. The $\boldsymbol{E}$ ellipse is the data ellipse of the residuals, $(\boldsymbol{y}_{ij} - \bar{\boldsymbol{y}}_{i\cdot})$. Using evidence ("significance") scaling of the $\boldsymbol{H}$ ellipse, the plot has the property that the multivariate test for a given hypothesis is significant by Roy's largest-root test *iff* the $\boldsymbol{H}$ ellipse protrudes anywhere outside the $\boldsymbol{E}$ ellipse.

the data ellipse of the fitted values, and corresponds visually to a multivariate analog of univariate effect-size measures (e.g., $(\bar{y}_1 - \bar{y}_2)/s_e$ where s_e is the within-group standard deviation).

For significance scaling, it turns out to be most visually convenient to use Roy's largest-root statistic as the test criterion. In this case, the $\boldsymbol{H}$ ellipse is scaled to $\boldsymbol{H}/(\lambda_\alpha df_e)$ where λ_α is the critical value of Roy's statistic.[13] Using this scaling gives a simple visual test of H_0: Roy's test rejects H_0 at a given α level *iff* the corresponding α-level $\boldsymbol{H}$ ellipse protrudes *anywhere* outside the $\boldsymbol{E}$ ellipse.[14] Moreover, the directions in which the hypothesis ellipse exceed the error ellipse are informative about the responses and their linear combinations that depart significantly from H_0. Thus, in Figure 20(b), the variation of the means of the iris species shown for these two variables appears to be largely one-dimensional, corresponding to a weighted sum (or average) of petal length and sepal length, perhaps a measure of overall size.

5.2 Linear hypotheses: geometries of contrasts and sums of effects

Just as in univariate ANOVA designs, important overall effects ($\text{df}_h > 1$) in MANOVA may be usefully explored and interpreted by the use of contrasts among the levels of the factors involved. In the general linear hypothesis test of Eqn. (11), contrasts are easily specified as one or more $(h_i \times q)$ $\boldsymbol{L}$ matrices, $\boldsymbol{L}_1, \boldsymbol{L}_2, \ldots$, each of whose rows sums to zero.

As an important special case, for an overall effect with df_h degrees of freedom (and balanced sample sizes), a set of df_h pairwise orthogonal $(1 \times q)$ $\boldsymbol{L}$ matrices ($\boldsymbol{L}_i^\mathsf{T}\boldsymbol{L}_j = 0$ for $i \neq j$) gives rise to a set of df_h rank-one $\boldsymbol{H}_i$ matrices that additively decompose the overall hypothesis SSCP matrix (by a multivariate

[13]The F test based on Roy's largest root uses the approximation $F = (df_2/df_1)\lambda_1$ with degrees of freedom df_1, df_2, where $df_1 = \max(df_h, df_e)$ and $df_2 = df_e - df_1 + df_h$. Inverting the F statistic gives the critical value for an α-level test: $\lambda_\alpha = (df_1/df_2)F^{1-\alpha}_{df_1,df_2}$

[14]Other multivariate tests (Wilks's Λ, Hotelling-Lawley trace, Pillai trace) also have geometric interpretations in HE plots (e.g., Wilks's Λ is the ratio of areas (volumes) of the $\boldsymbol{H}$ and $\boldsymbol{E}$ ellipses (ellipsoids); Hotelling-Lawley trace is based on the sum of the λ_i), but these statistics do not provide such simple visual comparisons. All HE plots shown in this paper use significance scaling, based on Roy's test.

analog of Pythagoras' Theorem),

$$\boldsymbol{H} = \boldsymbol{H}_1 + \boldsymbol{H}_2 + \cdots + \boldsymbol{H}_{\mathrm{df}_h} \ ,$$

exactly as the univariate SS_H may be decomposed in an ANOVA. Each of these rank-one $\boldsymbol{H}_i$ matrices will plot as a vector in an HE plot, and their collection provides a visual summary of the overall test, as partitioned by these orthogonal contrasts. Even more generally, where the subhypothesis matrices may be of rank > 1, the subhypotheses will have hypothesis ellipses of dimension rank($\boldsymbol{H}_i$) that are conjugate with respect to the hypothesis ellipse for the joint hypothesis, provided that the estimators for the subhypotheses are statistically independent.

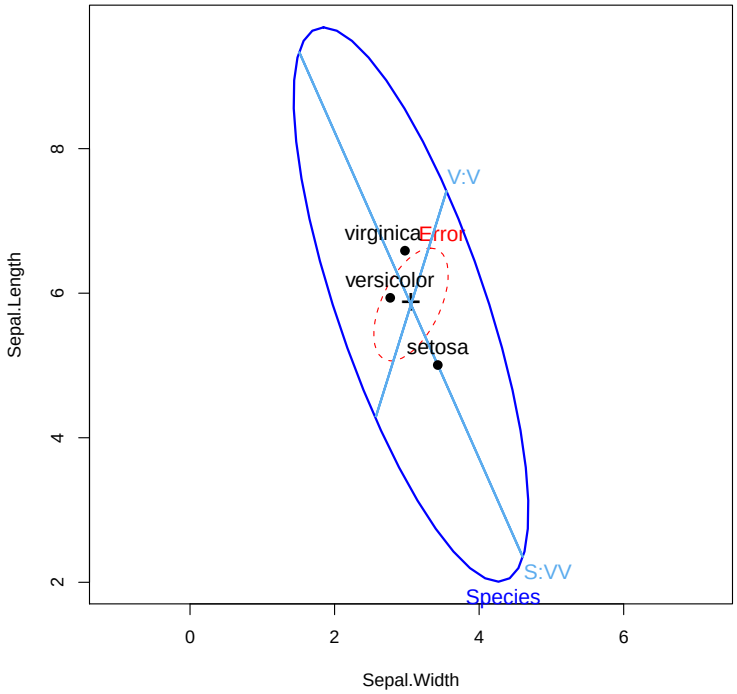

Figure 21: $\boldsymbol{H}$ and $\boldsymbol{E}$ matrices for sepal width and sepal length in the iris data, together with $\boldsymbol{H}$ matrices for testing two orthogonal contrasts in the species effect.

To illustrate, we show in Figure 21 an HE plot for the sepal width and sepal length variables in the iris data, corresponding to panel (1:2) in Figure 3. Overlayed on this plot are the one-df $\boldsymbol{H}$ matrices obtained from testing two orthogonal contrasts among the iris species: *setosa* vs. the average of *versicolor* and *virginica* (labeled "S:VV"), and *versicolor* vs. *virginica* ("V:V"), for which the contrast matrices are

$$\begin{aligned} \boldsymbol{L}_1 &= \begin{pmatrix} -2 & 1 & 1 \end{pmatrix} \\ \boldsymbol{L}_2 &= \begin{pmatrix} 0 & 1 & -1 \end{pmatrix} , \end{aligned}$$

where the species (columns) are taken in alphabetical order. In this view, the joint hypothesis testing equality of the species means has its major axis in data space largely in the direction of sepal length. The 1D degenerate "ellipse" for $\boldsymbol{H}_1$, representing the contrast of setosa with the average of the other two species, is closely aligned with this axis. The "ellipse" for $\boldsymbol{H}_2$ has a relatively larger component aligned with sepal width.

5.3 Canonical projections: ellipses in data space and canonical space

HE plots show the covariation leading toward rejection of a hypothesis relative to error covariation for two variables in data space. To visualize these relationships for more than two response variables, we can use the obvious generalization of a scatterplot matrix showing the 2D projections of the $\boldsymbol{H}$ and $\boldsymbol{E}$ ellipsoids for all pairs of variables. Alternatively, a transformation to canonical space permits visualization of all response variables in the reduced-rank 2D (or 3D) space in which $\boldsymbol{H}$ covariation is maximal.

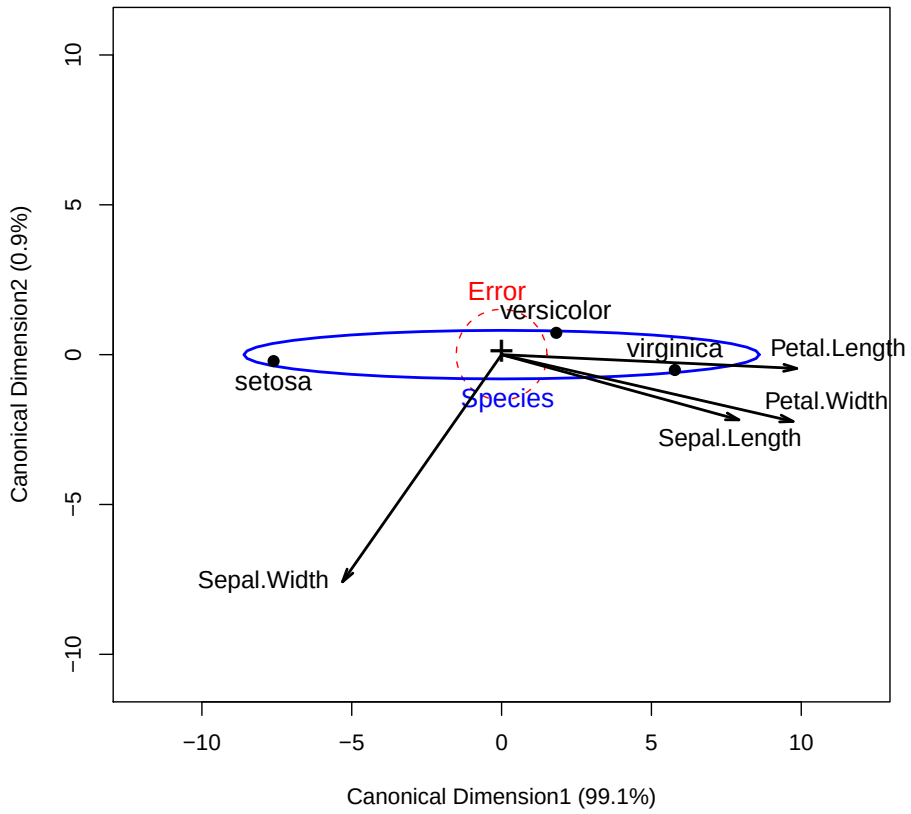

Figure 22: Canonical HE plot for the Iris data. In this plot, the $\boldsymbol{H}$ ellipse is shown using effect-size scaling to preserve resolution, and the variable vectors have been multiplied by a constant to approximately fill the plot space. The projections of the variable vectors on the coordinate axes show the correlations of the variables with the canonical dimensions.

In the MANOVA context, the analysis is called canonical discriminant analysis (CDA), where the emphasis is on dimension-reduction rather than hypothesis testing. For a one-way design with g groups and p-variate observations i in group j, $\boldsymbol{y}_{ij}$, CDA finds a set of $s = \min(p, g-1)$ linear combinations, $z_1 = \boldsymbol{c}_1^\mathsf{T}\boldsymbol{y}$, $z_2 = \boldsymbol{c}_2^\mathsf{T}\boldsymbol{y}$, $\ldots$, $z_s = \boldsymbol{c}_s^\mathsf{T}\boldsymbol{y}$, so that: (a) all z_k are mutually uncorrelated; (b) the vector of weights $\boldsymbol{c}_1$ maximizes the univariate F statistic for the linear combination z_1; (c) each successive vector of weights, $\boldsymbol{c}_k, k = 2, \ldots, s$, maximizes the univariate F-statistic for z_k, subject to being uncorrelated with all other linear combinations.

The canonical projection of $\boldsymbol{Y}$ to canonical scores $\boldsymbol{Z}$ is given by

$$\boldsymbol{Y}_{n\times p} \mapsto \boldsymbol{Z}_{n\times s} = \boldsymbol{Y}\boldsymbol{E}^{-1}\boldsymbol{V}/df_e \ , \tag{14}$$

where $\boldsymbol{V}$ is the matrix whose columns are the eigenvectors of $\boldsymbol{H}\boldsymbol{E}^{-1}$ associated with the ordered non-zero eigenvalues, $\lambda_i, i = 1, \ldots, s$. A MANOVA of all s linear combinations is statistically equivalent to that of

the raw data. The λ_i are proportional to the fractions of between-group variation expressed by these linear combinations. Hence, to the extent that the first one or two eigenvalues are relatively large, a two-dimensional display will capture the bulk of between-group differences. The 2D canonical discriminant HE plot is then simply an HE plot of the scores $\boldsymbol{z}_1$ and $\boldsymbol{z}_2$ on the first two canonical dimensions. (If $s \geq 3$, an analogous 3D version may also be obtained.)

Because the $\boldsymbol{z}$ scores are all mutually uncorrelated, the $\boldsymbol{H}$ and $\boldsymbol{E}$ matrices will always have their axes aligned with the canonical dimensions. When, as here, the $\boldsymbol{z}$ scores are standardized, the $\boldsymbol{E}$ ellipse will be circular, assuming that the axes in the plot are equated so that a unit data length has the same physical length on both axes.

Moreover, we can show the contributions of the original variables to discrimination as follows: Let $\boldsymbol{P}$ be the $p \times s$ matrix of the correlations of each column of $\boldsymbol{Y}$ with each column of $\boldsymbol{Z}$, often called *canonical structure* coefficients. Then, for variable j, a vector from the origin to the point whose coordinates $\boldsymbol{p}_{\cdot j}$ are given in row j of $\boldsymbol{P}$ has projections on the canonical axes equal to these structure coefficients and squared length equal to the sum squares of these correlations.

Figure 22 shows the canonical HE plot for the iris data, the view in canonical space corresponding to Figure 21 in data space for two of the variables (omitting the contrast vectors). Note that for $g = 3$ groups, $df_h = 2$, so $s = 2$ and the representation in 2D is exact. This provides a very simple interpretation: Nearly all (99.1%) of the variation in species means can be accounted for by the first canonical dimension, which is seen to be aligned with three of the four variables, most strongly with petal length. The second canonical dimension is mostly related to variation in the means on sepal width, and this variable is negatively correlated with the other three.

Finally, imagine a 4D version of the HE plot of Figure 21 in data space, showing the four-dimensional ellipsoids for $\boldsymbol{H}$ and $\boldsymbol{E}$. Add to this plot unit vectors corresponding to the coordinate axes, scaled to some convenient constant length. Some rotation would show that the $\boldsymbol{H}$ ellipsoid is really only two-dimensional, while $\boldsymbol{E}$ is 4D. Applying the transformation given by $\boldsymbol{E}^{-1}$ as in Figure A.4 and projecting into the 2D subspace of the non-zero dimensions of $\boldsymbol{H}$ would give a view equivalent to the canonical HE plot in Figure 22. The variable vectors in this plot are just the shadows of the original coordinate axes.

6 Kissing ellipsoids

In this section, we consider some circumstances in which there is a data stratification factor or there are two (or more) principles or procedures for deriving estimates of a parameter vector $\boldsymbol{\beta}$ of a linear model, each with its associated estimated covariance matrix, e.g., $\widehat{\boldsymbol{\beta}}^A$ with covariance matrix $\widehat{\mathrm{Var}}(\hat{\boldsymbol{\beta}}^A)$, and $\widehat{\boldsymbol{\beta}}^B$ with covariance matrix $\widehat{\mathrm{Var}}(\hat{\boldsymbol{\beta}}^B)$. The simplest motivating example is two-group discriminant analysis (Section 6.2). In data space, solutions to this statistical problem can be described geometrically in terms of the property that the data ellipsoids around the group centroids will just "kiss" (or *osculate*) along a path between the two centroids. We call this path the *locus of osculation*, whose properties are described in Section 6.1.

Perhaps more interesting and more productive is that the same geometric ideas apply equally well in parameter (β) space. Consider, for example, method A to be OLS estimation and several alternatives for method B, such as ridge regression (Section 6.3) or Bayesian estimation (Section 6.4). The remarkable fact is that the geometry of such kissing ellipsoids provides a clear visual interpretation of these cases and others, whenever we consider a convex combination of information from two sources. In all cases, the locus of osculation is interpretable in terms of the statistical goal to be achived.

6.1 Locus of osculation

The problems mentioned above all have a similar and simple physical interpretation: Imagine two stones dropped into a pond at locations with coordinates $\boldsymbol{m}_1$ and $\boldsymbol{m}_2$. The waves emanating from the centers

form concentric circles which osculate along the line from $\boldsymbol{m}_1$ to $\boldsymbol{m}_2$. Now imagine a world with ellipse-generating stones, where instead of circles, the waves form concentric ellipses determined by the shape matrices $\boldsymbol{A}_1$ and $\boldsymbol{A}_2$. The *locus of osculation* of these ellipses will be the set of points where the tangents to the two ellipses are parallel (or equivalently, that their normals are parallel). An example is shown in Figure 23, using $\boldsymbol{m}_1 = (-2, 2)$, $\boldsymbol{m}_2 = (2, 6)$, and

$$\boldsymbol{A}_1 = \begin{pmatrix} 1.0 & 0.5 \\ 0.5 & 1.5 \end{pmatrix} , \qquad \boldsymbol{A}_2 = \begin{pmatrix} 1.5 & -0.3 \\ -0.3 & 1.0 \end{pmatrix} , \tag{15}$$

where we have found points of osculation by trial and error.

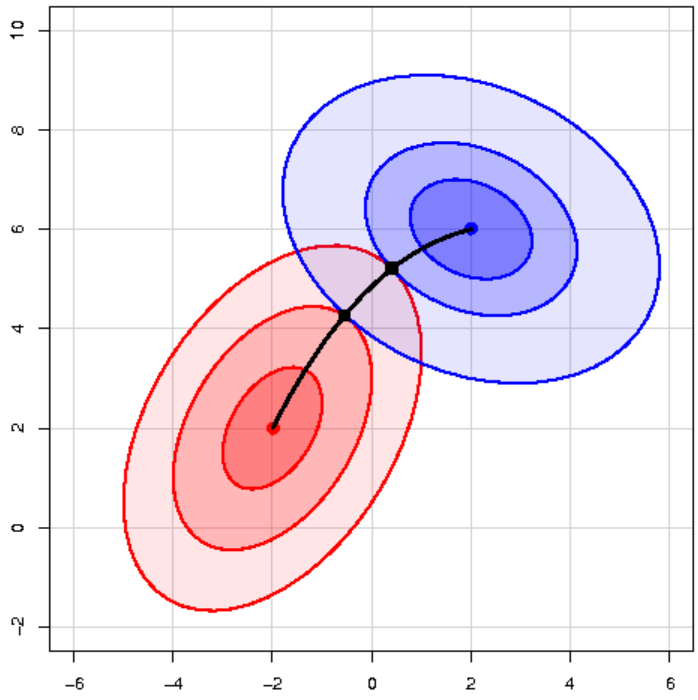

Figure 23: Locus of osculation for two families of ellipsoidal level curves, with centers at $\boldsymbol{m}_1 = (-2, 2)$ and $\boldsymbol{m}_2 = (2, 6)$, and shape matrices $\boldsymbol{A}_1$ and $\boldsymbol{A}_2$ given in Eqn. (15). The left ellipsoids (red) have radii=1, 2, 3. The right ellipsoids have radii=1, 1.74, 3.1, where the last two values were chosen to make them kiss at the points marked with squares. The black curve is an approximation to the path of osculation, using a spline function connecting $\boldsymbol{m}_1$ to $\boldsymbol{m}_2$ via the marked points of osculation.

An exact general solution can be described as follows: Let the ellipses be given by

$$\begin{aligned} f_1(\boldsymbol{x}) &= \boldsymbol{m}_1 \oplus \sqrt{\boldsymbol{A}_1} = (\boldsymbol{x} - \boldsymbol{m}_1)^{\mathsf{T}} \boldsymbol{A}_1 (\boldsymbol{x} - \boldsymbol{m}_1) \\ f_2(\boldsymbol{x}) &= \boldsymbol{m}_2 \oplus \sqrt{\boldsymbol{A}_2} = (\boldsymbol{x} - \boldsymbol{m}_2)^{\mathsf{T}} \boldsymbol{A}_2 (\boldsymbol{x} - \boldsymbol{m}_2) \ , \end{aligned}$$

and denote their gradient-vector functions as

$$\nabla f(x_1, x_2) = \left(\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2} \right) \tag{16}$$

so that

$$\begin{aligned} \nabla f_1(\boldsymbol{x}) &= 2\boldsymbol{A}_1(\boldsymbol{x}-\boldsymbol{m}_1) \\ \nabla f_2(\boldsymbol{x}) &= 2\boldsymbol{A}_2(\boldsymbol{x}-\boldsymbol{m}_2) \ . \end{aligned}$$

Then, the points where ∇f_1 and ∇f_2 are parallel can be expressed in terms of the condition that their vector cross product, $\boldsymbol{u} \circledast \boldsymbol{v} = u_1 v_2 - u_2 v_1 = \boldsymbol{v}^\mathsf{T} \boldsymbol{C} \boldsymbol{u} = 0$, where $\boldsymbol{C}$ is the skew-symmetric matrix

$$\boldsymbol{C} = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}$$

satisfying $\boldsymbol{C} = -\boldsymbol{C}^\mathsf{T}$. Thus, the locus of osculation is the set $\mathcal{O}$, given by $\mathcal{O} = \{\boldsymbol{x} \in \mathbb{R}^2 \,|\, \nabla f_1(\boldsymbol{x}) \circledast \nabla f_2(\boldsymbol{x}) = 0\}$, which implies

$$(\boldsymbol{x}-\boldsymbol{m}_2)^\mathsf{T} \, \boldsymbol{A}_2^\mathsf{T} \, \boldsymbol{C} \, \boldsymbol{A}_1 \, (\boldsymbol{x}-\boldsymbol{m}_1) = 0 \ . \tag{17}$$

Eqn. (17) is a biquadratic form in $\boldsymbol{x}$, with central matrix $\boldsymbol{A}_2^\mathsf{T} \, \boldsymbol{C} \, \boldsymbol{A}_1$, implying that $\mathcal{O}$ is a conic section in the general case. Note that when $\boldsymbol{x} = \boldsymbol{m}_1$ or $\boldsymbol{x} = \boldsymbol{m}_2$, Eqn. (17) is necessarily zero, so the locus of osculation always passes through $\boldsymbol{m}_1$ and $\boldsymbol{m}_2$.

A visual demonstration of the theory above is shown in Figure 24 (left), which overlays the ellipses in Figure 23 with contour lines (hyperbolae, here) of the vector cross-product function contained in Eqn. (17). When the contours of f_1 and f_2 have the same shape ($\boldsymbol{A}_1 = c\boldsymbol{A}_2$), as in the right panel of Figure 24, Eqn. (17) reduces to a line, in accord with the stones-in-pond interpretation. The above can be readily extended to ellipsoids in higher dimension, where the development is more easily understood in terms of normals to the surfaces.

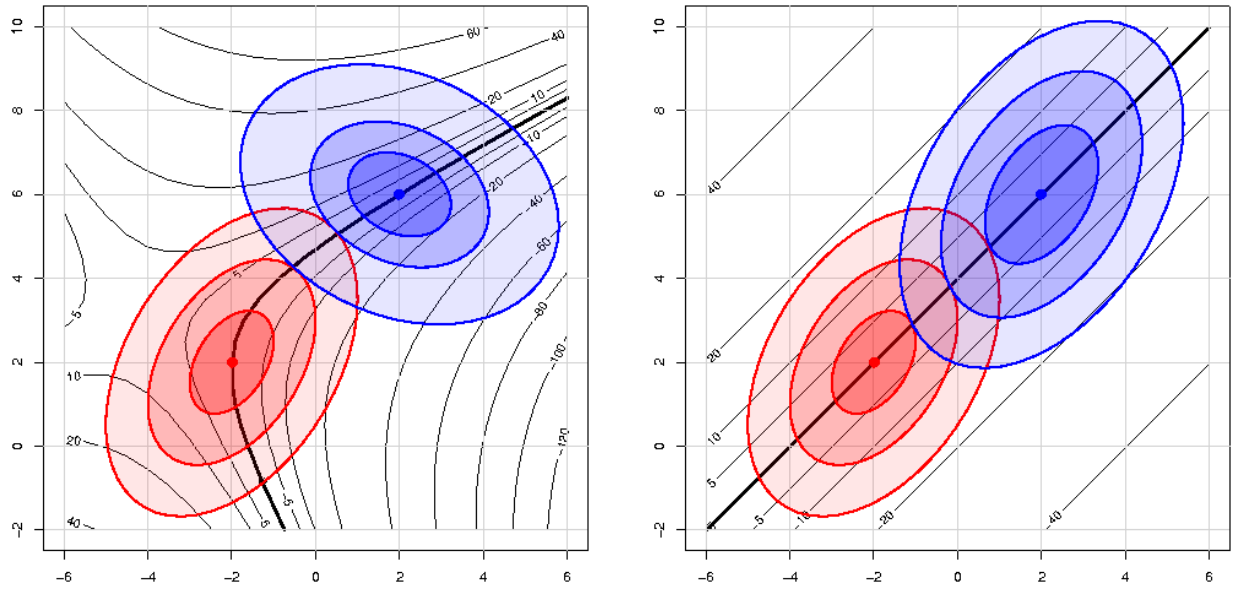

Figure 24: Locus of osculation for two families of ellipsoidal level curves, showing contour lines of the vector cross-product function Eqn. (17). The thick black curve shows the complete locus of osculation for these two families of ellipses, where the cross-product function equals 0. Left: with parameters as in Figure 23 and Eqn. (15). Right: with the same shape matrix $\boldsymbol{A}_1$ for both ellipsoids.

6.2 Discriminant analysis

The right panel of Figure 24, considered in data space, provides a visual interpretation of the classical, normal theory two-group discriminant analysis problem under the assumption of equal population covariance matrices, $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$. Here, we imagine that the plot shows the contours of data ellipsoids for two groups, with mean vectors $\boldsymbol{m}_1$ and $\boldsymbol{m}_2$, and common covariance matrix $\boldsymbol{A} = \boldsymbol{S}_{\text{pooled}} = [(n_1 - 1)\boldsymbol{S}_1 + (n_2 - 1)\boldsymbol{S}_2]/(n_1 + n_2 - 2)$.

The discriminant axis is the locus of osculation between the two families of ellipsoids. The goal in discriminant analysis, however, is to determine a classification rule based on a linear function, $\mathcal{D}(\boldsymbol{x}) = \boldsymbol{b}^\mathsf{T}\boldsymbol{x}$, such that an observation $\boldsymbol{x}$ will be classified as belonging to Group 1 if $\mathcal{D}(\boldsymbol{x}) \leq d$, and to Group 2 otherwise. In linear discriminant analysis, the discriminant function coefficients are given by

$$\boldsymbol{b} = \boldsymbol{S}_{\text{pooled}}^{-1}(\boldsymbol{m}_1 - \boldsymbol{m}_2) \ .$$

All boundaries of the classification regions determined by d will then be the tangent lines (planes) to the ellipsoids at points of osculation. The location of the classification region along the line from $\boldsymbol{m}_1$ to $\boldsymbol{m}_2$ typically takes into account both the prior probabilities of membership in Groups 1 and 2, and the costs of misclassification. Similarly, the left panel of Figure 24 is a visual representation of the same problem when $\boldsymbol{\Sigma}_1 \neq \boldsymbol{\Sigma}_2$, giving rise to quadratic classification boundaries.

6.3 Ridge regression

In the univariate linear model, $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, high multiple correlations among the predictors in $\boldsymbol{X}$ lead to problems of *collinearity*—unstable OLS estimates of the parameters in $\boldsymbol{\beta}$ with inflated standard errors and coefficients that tend to be too large in absolute value. Although collinearity is essentially a data problem (Fox, 2008), one popular (if questionable) approach is ridge regression, which shrinks the estimates toward $\mathbf{0}$ (introducing bias) in an effort to reduce sampling variance.

Suppose the predictors and response have been centered at their means and the unit vector is omitted from $\boldsymbol{X}$. Further, rescale the columns of $\boldsymbol{X}$ to unit length, so that $\boldsymbol{X}^\mathsf{T}\boldsymbol{X}$ is a correlation matrix. Then, the OLS estimates are given by

$$\widehat{\boldsymbol{\beta}}^{\text{OLS}} = (\boldsymbol{X}^\mathsf{T}\boldsymbol{X})^{-1}\boldsymbol{X}^\mathsf{T}\boldsymbol{y} \ . \tag{18}$$

Ridge regression replaces the standard residual sum of squares criterion with a penalized form,

$$\text{RSS}(k) = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^\mathsf{T}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) + k\boldsymbol{\beta}^\mathsf{T}\boldsymbol{\beta} \qquad (k \geq 0) \ , \tag{19}$$

whose solution is easily seen to be

$$\begin{aligned} \widehat{\boldsymbol{\beta}}_k^{\text{RR}} &= (\boldsymbol{X}^\mathsf{T}\boldsymbol{X} + k\boldsymbol{I})^{-1}\boldsymbol{X}^\mathsf{T}\boldsymbol{y} \\ &= \boldsymbol{G}\,\widehat{\boldsymbol{\beta}}^{\text{OLS}} \ , \end{aligned} \tag{20}$$

where $\boldsymbol{G} = \left[\boldsymbol{I} + k(\boldsymbol{X}^\mathsf{T}\boldsymbol{X})^{-1}\right]^{-1}$. Thus, as the "ridge constant" k increases, $\boldsymbol{G}$ decreases, driving $\widehat{\boldsymbol{\beta}}_k^{\text{RR}}$ toward $\mathbf{0}$ (Hoerl and Kennard, 1970a,b). The addition of a positive constant k to the diagonal of $\boldsymbol{X}^\mathsf{T}\boldsymbol{X}$ drives $\det(\boldsymbol{X}^\mathsf{T}\boldsymbol{X} + k\boldsymbol{I})$ away from zero even if $\det(\boldsymbol{X}^\mathsf{T}\boldsymbol{X}) \approx 0$.

The penalized Lagrangian formulation in Eqn. (19) has an equivalent form as a constrained minimization problem,

$$\widehat{\boldsymbol{\beta}}^{\text{RR}} = \underset{\boldsymbol{\beta}}{\text{argmin}}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^\mathsf{T}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) \qquad \text{subject to} \qquad \boldsymbol{\beta}^\mathsf{T}\boldsymbol{\beta} \leq t(k) \ , \tag{21}$$

which makes the size constraint on the parameters explicit, with $t(k)$ an inverse function of k. This form provides a visual interpretation of ridge regression, as shown in Figure 25. Depicted in the figure are the

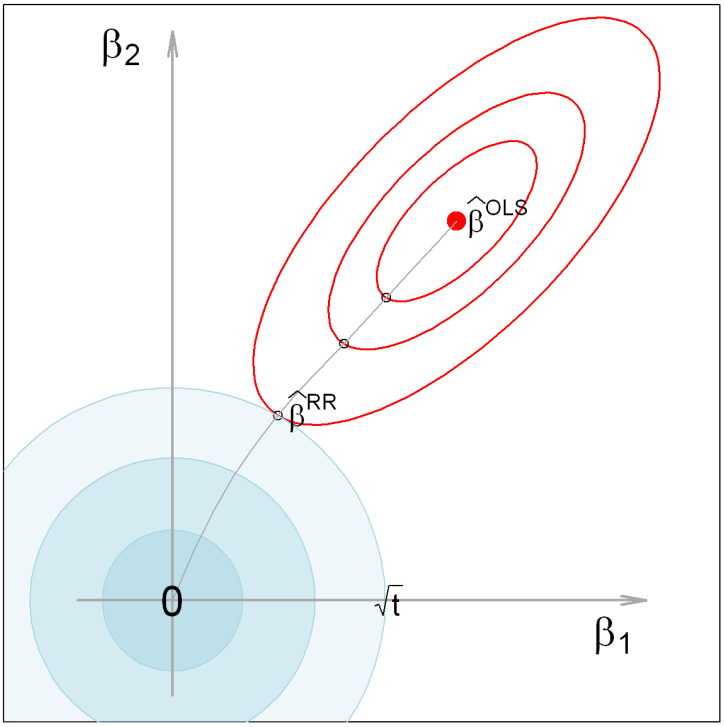

Figure 25: Elliptical contours of the OLS residual sum of squares for two parameters in a regression, together with circular contours for the constraint function, $\beta_1^2 + \beta_2^2 \leq t$. Ridge regression finds the point $\boldsymbol{\beta}^{RR}$ where the OLS contours just kiss the contstraint region.

elliptical contours of the OLS regression sum of squares, $\mathrm{RSS}(0)$ around $\widehat{\boldsymbol{\beta}}^{\mathrm{OLS}}$. Each ellipsoid marks the point closest to the origin, i.e., with $\min \boldsymbol{\beta}^{\mathsf{T}}\boldsymbol{\beta}$. It is easily seen that the ridge regression solution is the point where the elliptical contours just kiss the constraint contour.

Another insightful interpretation of ridge regression (Marquardt, 1970) sees the ridge estimator as equivalent to an OLS estimator, when the actual data in $\boldsymbol{X}$ are supplemented by some number of fictitious observations, $n(k)$, with uncorrelated predictors, giving rise to an orthogonal $\boldsymbol{X}_k^0$ matrix, and where $y = 0$ for all supplementary observations. The linear model then becomes,

$$\begin{pmatrix} \boldsymbol{y} \\ \mathbf{0} \end{pmatrix} = \begin{pmatrix} \boldsymbol{X} \\ \boldsymbol{X}_k^0 \end{pmatrix} \boldsymbol{\beta}^{\mathrm{RR}} + \begin{pmatrix} \boldsymbol{e} \\ \boldsymbol{e}_k^0 \end{pmatrix} \ , \tag{22}$$

which gives rise to the solution,

$$\widehat{\boldsymbol{\beta}}^{\mathrm{RR}} = [\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X} + (\boldsymbol{X}_k^0)^{\mathsf{T}}\boldsymbol{X}_k^0]^{-1}\boldsymbol{X}^{\mathsf{T}}\boldsymbol{y} \ . \tag{23}$$

But because $\boldsymbol{X}_k^0$ is orthogonal, $(\boldsymbol{X}_k^0)^{\mathsf{T}}\boldsymbol{X}_k^0$ is a scalar multiple of $\boldsymbol{I}$, so there exists some value of k making Eqn. (23) equivalent to Eqn. (20). As promised, the ridge regression estimator then reflects a weighted average of the data $[\boldsymbol{X}, \boldsymbol{y}]$ with $n(k)$ observations $[\boldsymbol{X}_k^0, \mathbf{0}]$ biased toward $\boldsymbol{\beta} = \mathbf{0}$. In Figure 25, it is easy to imagine that there is a direct translation between the size of the constraint region, $t(k)$, and an equivalent supplementary sample size, $n(k)$, in this interpretation.

This classic version of the ridge regression problem can be generalized in a variety of ways, giving other geometric insights. Rather than a constant multiplier k of $\boldsymbol{\beta}^{\mathsf{T}}\boldsymbol{\beta}$ as the penalty term in Eqn. (19), consider a penalty of the form $\boldsymbol{\beta}^{\mathsf{T}}\boldsymbol{K}\boldsymbol{\beta}$ with a positive definite matrix $\boldsymbol{K}$. The choice $\boldsymbol{K} = \mathrm{diag}\,(k_1, k_2, \dots)$ gives rise to a version of Figure 25 in which the constraint contours are ellipses aligned with the coordinate axes, with axis

lengths inversely proportional to k_i. These constants allow for differential shrinkage of the OLS coefficients. The visual solution to the obvious modification of Eqn. (21) is again the point where the elliptical contours of $\mathrm{RSS}(0)$ kiss the contours of the (now elliptical) constraint region.

6.3.1 Bivariate ridge trace plots

Ridge regression is touted (optimistically we think) as a method to counter the effects of collinearity by trading off a small amount of bias for an advantageous decrease in variance. The results are often visualized in a a *ridge trace plot* (Hoerl and Kennard, 1970b), showing the changes in individual coefficient estimates as a function of k. A bivariate version of this plot, with confidence ellipses for the parameters is introduced here. This plot provides greater insight into the effects of k on coefficient variance.[15]

Confidence ellipsoids for the OLS estimator are generated from the estimated covariance matrix of the coefficients,

$$\widehat{\mathrm{Var}}(\boldsymbol{\beta}^{\mathrm{OLS}}) = \hat{\sigma}_e^2 (\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X})^{-1} \ .$$

For the ridge estimator, this becomes (Marquardt, 1970)

$$\widehat{\mathrm{Var}}(\boldsymbol{\beta}^{\mathrm{RR}}) = \hat{\sigma}_e^2 [\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X} + k\boldsymbol{I}]^{-1} (\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X}) [\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X} + k\boldsymbol{I}]^{-1} \ , \tag{24}$$

which coincides with the OLS result when $k = 0$.

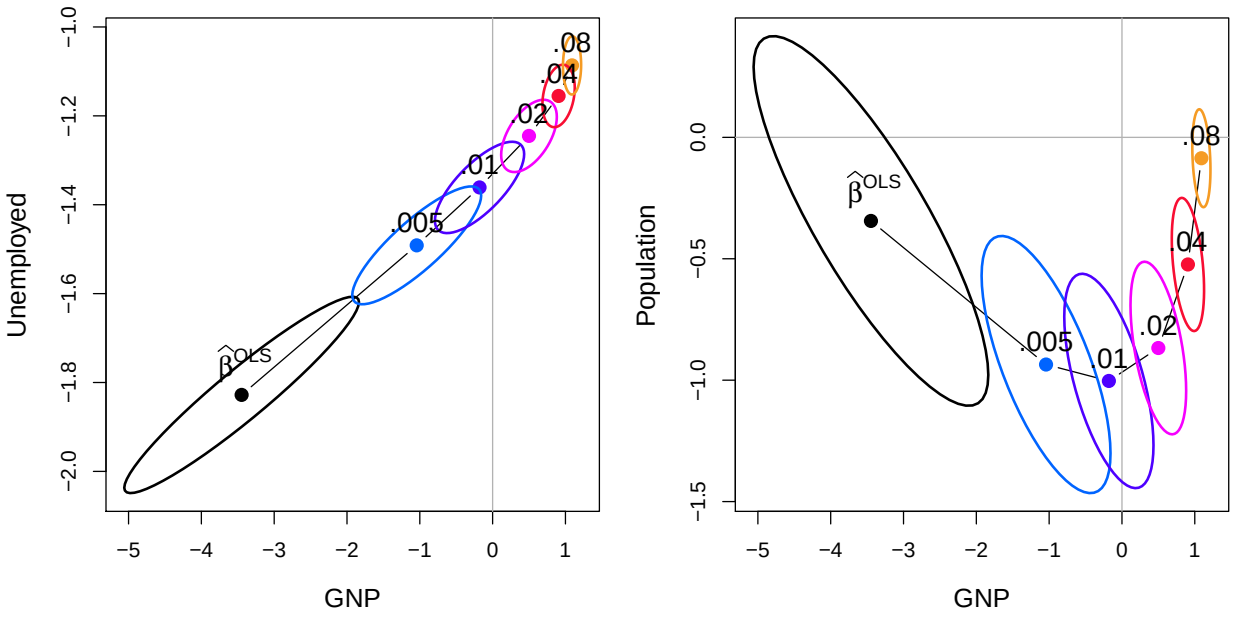

Figure 26: Bivariate ridge trace plots for the coefficients of Unemployed and Population against the coefficient for GNP in Longley's data, with $k = 0, 0.005, 0.01, 0.02, 0.04, 0.08$. In both cases the coefficients are driven on average toward zero, but the bivariate plot also makes clear the reduction in variance. To reduce overlap, all variance ellipses are shown with 1/2 the standard radius.

Figure 26 uses the classic Longley (1967) data to illustrate bivariate ridge trace plots. The data consist of an economic time series ($n = 16$) observed yearly from 1947 to 1962, with the number of people Employed as the response and the following predictors: GNP, Unemployed, Armed.Forces, Population, Year,

[15] Bias and mean-squared error are a different matter: Although Hoerl and Kennard (1970a) demonstrate that there is a range of values for the ridge constant k for which the MSE of the ridge estimator is smaller than that of the OLS estimator, to know where this range is located requires knowledge of $\boldsymbol{\beta}$. As we explain in the following subsection, the constraint on $\hat{\boldsymbol{\beta}}$ incorporated in the ridge estimator can be construed as a Bayesian prior; the fly in the ointment of ridge regression, however, is that there is no reason to suppose that the ridge-regression prior is in general reasonable.

and GNP.deflator (using 1954 as 100).[16] For each value of k, the plot shows the estimate $\widehat{\boldsymbol{\beta}}$, together with the variance ellipse. For the sake of this example, we assume that GNP is a primary predictor of Employment, and we wish to know how other predictors modify the regression estimates and their variance when ridge regression is used.

For these data, it can be seen that even small values of k have substantial impact on the estimates $\widehat{\boldsymbol{\beta}}$. What is perhaps more dramatic (and unseen in univariate trace plots) is the impact on the size of the variance ellipse. Moreover, shrinkage in variance is generally in a similar direction to the shrinkage in the coefficients. This new graphical method is developed more fully in Friendly (2012), including 2D and 3D plots, as well as more informative representations of shrinkage by ellipsoids in the transformed space of the SVD of the predictors.

6.4 Bayesian linear models

In a Bayesian alternative to standard least squares estimation, consider the case where our prior information about $\boldsymbol{\beta}$ can be encapsulated in a distribution with a prior mean $\boldsymbol{\beta}^{\text{prior}}$ and covariance matrix $\boldsymbol{A}$. We show that under reasonable conditions the Bayesian posterior estimate, $\widehat{\boldsymbol{\beta}}^{\text{posterior}}$, turns out to be a weighted average of the prior coefficients $\boldsymbol{\beta}^{\text{prior}}$ and the OLS solution $\widehat{\boldsymbol{\beta}}^{\text{OLS}}$, with weights proportional to the conditional prior precision, $\boldsymbol{A}^{-1}$, and the data precision given by $\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X}$. Once again, this can be understood geometrically as the locus of osculation of ellipsoids that characterize the prior and the data.

Under Gaussian assumptions, the conditional likelihood can be written as

$$\mathcal{L}(\boldsymbol{y} \,|\, \boldsymbol{X}, \boldsymbol{\beta}, \sigma^2) \propto (\sigma^2)^{-n/2} \, \exp\left[-\frac{1}{2\sigma^2}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})^{\mathsf{T}}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})\right] \ .$$

To focus on alternative estimators, we can complete the square around $\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}^{\text{OLS}}$ to give

$$(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})^{\mathsf{T}}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}) = (\boldsymbol{y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}})^{\mathsf{T}}(\boldsymbol{y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}) + (\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}})^{\mathsf{T}}(\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X})(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}) \ . \tag{25}$$

With a little manipulation, a conjugate prior, of the form $\Pr(\boldsymbol{\beta}, \sigma^2) = \Pr(\boldsymbol{\beta} \,|\, \sigma^2) \times \Pr(\sigma^2)$ can be expressed with $\Pr(\sigma^2)$ an inverse gamma distribution depending on the first term on the right hand side of Eqn. (25) and $\Pr(\boldsymbol{\beta} \,|\, \sigma^2)$ a normal distribution,

$$\Pr(\boldsymbol{\beta} \,|\, \sigma^2) \propto (\sigma^2)^{-p} \times \exp\left[-\frac{1}{2\sigma^2}(\boldsymbol{\beta}-\boldsymbol{\beta}^{\text{prior}})^{\mathsf{T}} \, \boldsymbol{A}(\boldsymbol{\beta}-\boldsymbol{\beta}^{\text{prior}})\right] \ . \tag{26}$$

The posterior distribution is then $\Pr(\boldsymbol{\beta}, \sigma^2 \,|\, \boldsymbol{y}, \boldsymbol{X}) \propto \Pr(\boldsymbol{y} \,|\, \boldsymbol{X}, \boldsymbol{\beta}, \sigma^2) \times \Pr(\boldsymbol{\beta} \,|\, \sigma^2) \times \Pr(\sigma^2)$, whence, after some simplification, the posterior mean can be expressed as

$$\widehat{\boldsymbol{\beta}}^{\text{posterior}} = (\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X}\widehat{\boldsymbol{\beta}}^{\text{OLS}} + \boldsymbol{A}\boldsymbol{\beta}^{\text{prior}})(\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X} + \boldsymbol{A})^{-1} \tag{27}$$

with covariance matrix $(\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X} + \boldsymbol{A})^{-1}$. The posterior coefficients are therefore a weighted average of the prior coefficients and the OLS estimates, with weights given by the conditional prior precision, $\boldsymbol{A}^{-1}$, and the data precision, $\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X}$. Thus, as we increase the strength of our prior precision (decreasing prior variance), we place greater weight on our prior beliefs relative to the data.

In this context, ridge regression can be seen as the special case where $\widehat{\boldsymbol{\beta}}^{\text{prior}} = \mathbf{0}$ and $\boldsymbol{A} = k\boldsymbol{I}$, and where Figure 25 provides an elliptical visualization. In Eqn. (23), the number of observations, $n(k)$ corresponding to $\boldsymbol{X}_k^0$ can be seen as another way of expressing the weight of the prior in relation to the data.

[16] Longley (1967) used these data to demonstrate the effects of numerical instability and round-off error in least squares computations based on direct computation of the crossproducts matrix, $\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X}$. Longley's paper sparked the development of a wide variety of numerically stable least squares algorithms (QR, modified Gram-Schmidt, etc.) now used is almost all statistical software. Even ignoring numerical problems (not to mention problems due to lack of independence), these data would be anticipated to exhibit high collinearity because a number of the predictors would be expected to have strong associations with year and/or population, yet both of these are also included among the predictors.

6.5 Mixed models: BLUEs and BLUPs

In this section we make implicit use of the duality between data space and β space, where lines in one map into points in the other and ellipsoids help to visualize the precision of estimates in the context of the linear mixed model for hierarchical data. We also show visually how the best linear unbiased predictors (BLUPs) from the mixed model can be seen as a weighted average of the best linear unbiased estimates (BLUEs) derived from OLS regressions performed *within* clusters of related data and pooled OLS estimates computed *ignoring* clusters.

The mixed model for hierarchical data provides a general framework for dealing with dependence among observations in linear models, such as occurs when students are sampled within schools, schools within counties, and so forth (e.g., Raudenbush and Bryk, 2002). In these situations, the assumption of OLS that the errors are conditionally independent is probably violated, because, for example, students nested within the same school are likely to have more similar outcomes than those from different schools. Essentially the same model, with provision for serially correlated errors, can be applied to longitudinal data (e.g., Laird and Ware, 1982), although we will not pursue this application here.

The mixed model for the $n_i \times 1$ response vector $\boldsymbol{y}_i$ in cluster i can be given as

$$\begin{aligned} \boldsymbol{y}_i &= \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{u}_i + \boldsymbol{\epsilon}_i \\ \boldsymbol{u}_i &\sim \mathcal{N}_q(\boldsymbol{0}, \boldsymbol{G}) \\ \boldsymbol{\epsilon}_i &\sim \mathcal{N}_{n_i}(\boldsymbol{0}, \boldsymbol{R}_i) \end{aligned} \tag{28}$$

where $\boldsymbol{\beta}$ is a $p \times 1$ vector of parameters corresponding to the fixed effects in the $n_i \times p$ model matrix $\boldsymbol{X}_i$; $\boldsymbol{u}_i$ is a $q \times 1$ vector of coefficients corresponding to the random effects in the $n_i \times q$ model matrix $\boldsymbol{Z}_i$; $\boldsymbol{G}$ is the $q \times q$ covariance matrix of the random effects in $\boldsymbol{u}_i$; and $\boldsymbol{R}_i$ is the $n_i \times n_i$ covariance matrix of the errors in $\boldsymbol{\epsilon}_i$.

Stacking the $\boldsymbol{y}_i$, $\boldsymbol{X}_i$, $\boldsymbol{Z}_i$, and so forth in the obvious way then gives

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u} + \boldsymbol{\epsilon} \tag{29}$$

where $\boldsymbol{u}$ and $\boldsymbol{\epsilon}$ are assumed to have normal distributions with mean $\boldsymbol{0}$ and

$$\mathrm{Var}\begin{pmatrix}\boldsymbol{u} \\ \boldsymbol{\epsilon}\end{pmatrix} = \begin{bmatrix}\boldsymbol{G} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{R}\end{bmatrix} \quad . \tag{30}$$

The variance of $\boldsymbol{y}$ is therefore $\boldsymbol{V} = \boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}^\mathsf{T} + \boldsymbol{R}$, and when $\boldsymbol{Z} = \boldsymbol{0}$ and $\boldsymbol{R} = \sigma^2\boldsymbol{I}$, the mixed model in Eqn. (29) reduces to the standard linear model.

Assume that our interest lies primarily in estimating the fixed-effect parameters in $\boldsymbol{\beta}$. At one extreme (complete pooling) we could simply ignore clusters and calculate the pooled OLS estimate,

$$\widehat{\boldsymbol{\beta}}^{\text{pooled}} = (\boldsymbol{X}^\mathsf{T}\boldsymbol{X})^{-1}\boldsymbol{X}^\mathsf{T}\boldsymbol{y} \qquad \text{with} \qquad \widehat{\boldsymbol{S}} \equiv \widehat{\mathrm{Var}}(\widehat{\boldsymbol{\beta}}^{\text{pooled}}) = \hat{\sigma}^2(\boldsymbol{X}^\mathsf{T}\boldsymbol{X})^{-1} \quad . \tag{31}$$

At the other extreme (no pooling), we can ignore the variation across clusters and calculate the separate BLUE estimate within each cluster, giving the results of a fixed-effects analysis,

$$\widehat{\boldsymbol{\beta}}_i^{\text{unpooled}} = (\boldsymbol{X}_i^\mathsf{T}\boldsymbol{X}_i)^{-1}\boldsymbol{X}_i^\mathsf{T}\boldsymbol{y}_i \qquad \text{with} \qquad \widehat{\boldsymbol{S}}_i \equiv \widehat{\mathrm{Var}}(\widehat{\boldsymbol{\beta}}_i^{\text{unpooled}}) = \hat{\sigma}^2(\boldsymbol{X}_i^\mathsf{T}\boldsymbol{X}_i)^{-1} \quad . \tag{32}$$

Both extremes have drawbacks: whereas the pooled analysis ignores variation among clusters, the unpooled analysis ignores the cluster-averaged result and overstates variation within each cluster, making the clusters appear to differ more than they actually do.

This dilemma led to the development of BLUPs in models with random effects (Henderson, 1975, Robinson, 1991, Speed, 1991). In the case considered here, $\boldsymbol{Z}_i = \boldsymbol{X}_i$, and so $\widehat{\boldsymbol{u}}_i$ gives "estimates" of the random

effects with $\widehat{\mathrm{Var}}(\widehat{\boldsymbol{u}}_i) = \widehat{\boldsymbol{G}}$. The BLUPs are a weighted average of the $\widehat{\boldsymbol{\beta}}_i$ and $\widehat{\boldsymbol{u}}_i$ using the precision (Var^{-1}) as weights,

$$\widetilde{\boldsymbol{\beta}}_i^{\text{blup}} = \left(\widehat{\boldsymbol{\beta}}_i^{\text{unpooled}} \widehat{\boldsymbol{S}}_{\boldsymbol{i}}^{-1} + \widehat{\boldsymbol{u}}_{\boldsymbol{i}} \widehat{\boldsymbol{G}}^{-1}\right) \left(\widehat{\boldsymbol{S}}_{\boldsymbol{i}}^{-\mathbf{1}} + \widehat{\boldsymbol{G}}^{-\mathbf{1}}\right)^{-1} \quad . \tag{33}$$

This "partial pooling"optimally combines the information from cluster i with the information from all clusters, shrinking the $\widehat{\boldsymbol{\beta}}_i$ toward $\widehat{\boldsymbol{\beta}}^{\text{pooled}}$. Shrinkage for a given parameter β_{ij} is greater when the sample size n_i is small or when the estimated variance of the corresponding random effect, g_{jj}, is small.

Eqn. (33) is of the same form as Eqn. (27) and other convex combinations of estimates considered earlier in this section. So once again, we can understand these results geometrically as the locus of osculation of ellipsoids.

6.5.1 Example: Math achievement and SES

To illustrate, we use a classic data set from Bryk and Raudenbush (1992) and Raudenbush and Bryk (2002) dealing with math achievement scores for a subsample of 7,185 students from 160 schools in the 1982 High School & Beyond survey of U.S. public and Catholic high schools conducted by the National Center for Education Statistics (NCES). The data set contains 90 public schools and 70 Catholic schools, with sample sizes ranging from 14 to 67.

The response is a standardized measure of math achievement, while student-level predictor variables include sex and student socioeconomic status (SES), and school-level predictors include sector (public or Catholic) and mean SES for the school (among other variables). Following Raudenbush and Bryk (2002), student SES is considered the main predictor, and is typically analyzed centered within schools, $\mathrm{CSES}_{ij} = \mathrm{SES}_{ij} - (\mathrm{meanSES})_i$, for ease of interpretation (making the within-school intercept for school i equal to the mean SES in that school).

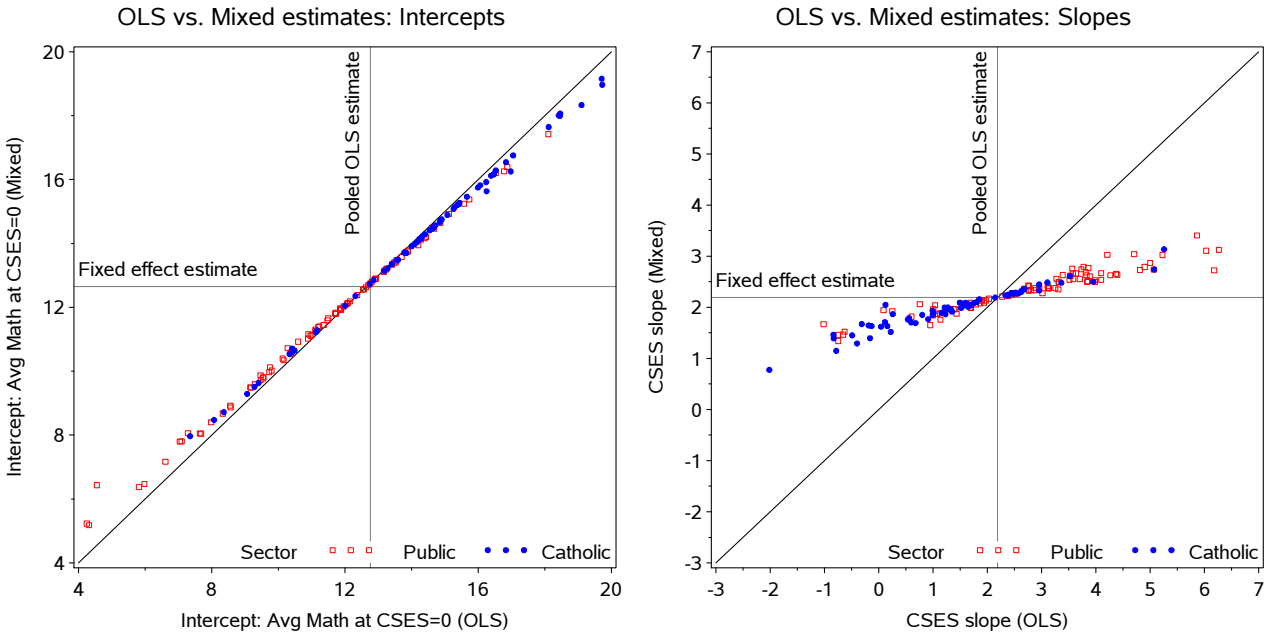

Figure 27: Comparing BLUEs and BLUPs. Each panel plots the OLS estimates from separate regressions for each school (BLUEs) versus the mixed model estimates from the random intercepts and slopes model (BLUPs). Left: intercepts; Right: slopes for CSES. The shrinkage of the BLUPs toward the OLS estimate is much greater for slopes than intercepts.

For simplicity, we consider the case of CSES as a single quantitative predictor in $\boldsymbol{X}$ in the example below. We fit and compare the following models:

$$\begin{aligned}
\boldsymbol{y}_i &\sim \mathcal{N}(\beta_0 + x_i\beta_1 + \epsilon_i, \sigma^2) \qquad \text{pooled OLS} && (34)\\
\boldsymbol{y}_i &\sim \mathcal{N}(\beta_{0i} + x_i\beta_{1i} + \epsilon_i, \sigma_i^2) \qquad \text{unpooled BLUEs} && (35)\\
\boldsymbol{y}_i &\sim \mathcal{N}(\beta_{0i} + x_i\beta_{1i} + u_{0i} + x_i u_{1i} + \epsilon_i, \sigma_i^2) \qquad \text{random intercepts and slopes} && (36)
\end{aligned}$$

and also include a fixed effect of sector, common to all models; for compactness, the sector effect is elided in the notation above.

In expositions of mixed-effects models, such models are often compared visually by plotting predicted values in data space, where each school appears as a fitted line under one of the models above (sometimes called "spaghetti plots"). Our geometric approach leads us to consider the equivalent but simpler plots in the dual β space, where each school appears as a point.

Figure 27 plots the unpooled BLUE estimates against those from the random effects model, with separate panels for intercepts and slopes to illustrate the shrinkage of different parameters. In these data, the variance in intercepts (average math achievement for students at $\text{CSES} = 0$), g_{00}, among schools in each sector is large, so the mixed effects estimates have small weight and there is little shrinkage. On the other hand, the variance component for slopes, g_{11}, is relatively small, so there is greater shrinkage toward the BLUE, fixed effect estimate.

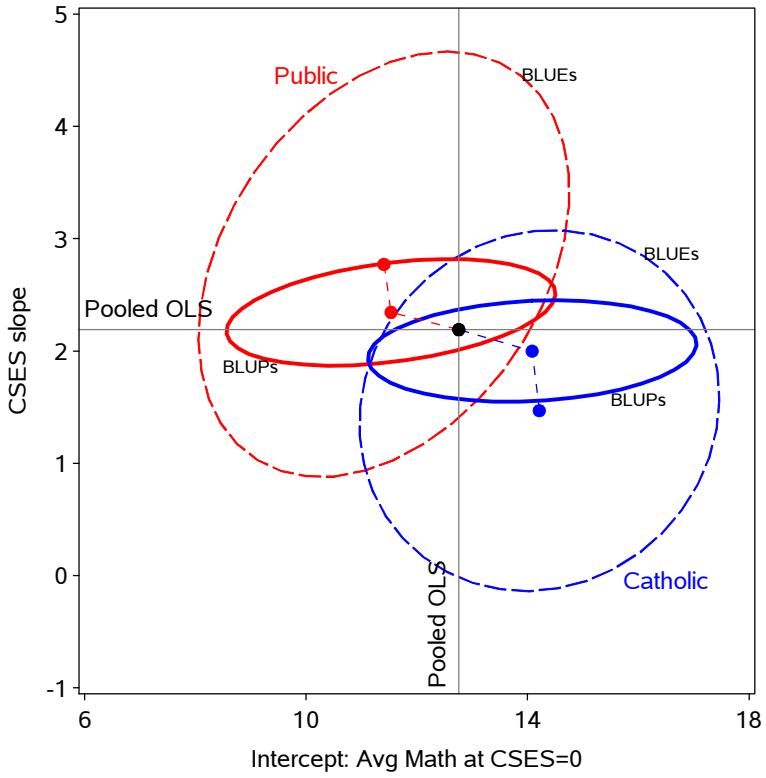

Figure 28: Comparing BLUEs and BLUPs. The plot shows ellipses of 50% coverage for the estimates of intercepts and slopes from OLS regressions (BLUEs) and the mixed model (BLUPs), separately for each sector. The centers of the ellipses illustrate how the BLUPS can be considered a weighted average of the BLUEs and the pooled OLS estimate, ignoring sector. The relative sizes of the ellipses reflect the smaller variance for the BLUPs compared to the BLUEs, particularly for slope estimates.

For the present purposes, a more useful visual representation of these model comparisons can be shown together in the space of (β_0, β_1), as in Figure 28. Estimates for individual schools are not shown, but rather

these are summarized by the ellipses of 50% coverage for the BLUEs and BLUPs within each sector. The centers of the ellipsoids indicate the relatively greater shrinkage of slopes compared to intercepts. The sizes of ellipsoids show directly the greater precision of the BLUPs, particularly for slopes.

6.6 Multivariate meta-analysis

A related situation arises in random effects multivariate meta-analysis (Berkey *et al.*, 1998, Nam *et al.*, 2003), where several outcome measures are observed in a series of similar research studies and it is desired to synthesize those studies to provide an overall (pooled) summary of the outcomes, together with meta-analytic inferences and measures of heterogeneity across studies.

The application of mixed model ideas in this context differs from the standard situation in that individual data are usually unavailable and use is made instead of summary data (estimated treatment effects and their covariances) from the published literature. The multivariate extension of standard univariate methods of meta-analysis allows the correlations among outcome effects to be taken into account and estimated, and regression versions can incorporate study-specific covariates to account for some inter-study heterogeneity. More importantly, we illustrate a graphical method based (of course) on ellipsoids that serves to illustrate bias, heterogeneity, and shrinkage in BLUPs, and the optimism of fixed-effect estimates when study heterogeneity is ignored.

The general mixed-effects multivariate meta-analysis model can be written as

$$\boldsymbol{y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{\delta}_i + \boldsymbol{e}_i \ , \tag{37}$$

where $\boldsymbol{y}_i$ is a vector of p outcomes (means or treatment effects) for study i; $\boldsymbol{X}_i$ is the matrix of study-level predictors for study i or a unit vector when no covariates are available; $\boldsymbol{\beta}$ is the population-averaged vector of regression parameters or effects (intercepts, means) when there are no covariates; $\boldsymbol{\delta}_i$ is the p-vector of random effects associated with study i, whose $p \times p$ covariance matrix $\boldsymbol{\Delta}$ represents the between-study heterogeneity unaccounted for by $\boldsymbol{X}_i\boldsymbol{\beta}$; and, finally, $\boldsymbol{e}_i$ is the p-vector of random sampling errors (independent of $\boldsymbol{\delta}_i$) within study i, having the $p \times p$ covariance matrix $\boldsymbol{S}_i$.

With suitable distributional assumptions, the mixed=effects model in Eqn. (37) implies that

$$\boldsymbol{y}_i \sim \mathcal{N}_p(\boldsymbol{X}_i\boldsymbol{\beta}, \boldsymbol{\Delta} + \boldsymbol{S}_i) \tag{38}$$

with $\mathrm{Var}(\boldsymbol{y}_i) = \boldsymbol{\Delta} + \boldsymbol{S}_i$. When all the $\boldsymbol{\delta}_i = \boldsymbol{0}$, and thus $\boldsymbol{\Delta} = \boldsymbol{0}$, Eqn. (37) reduces to a fixed-effects model, $\boldsymbol{y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{e}_i$, which can be estimated by GLS to give,

$$\widehat{\boldsymbol{\beta}}^{\mathrm{GLS}} = (\boldsymbol{X}^\mathsf{T}\boldsymbol{S}\boldsymbol{X})^{-1}\boldsymbol{X}^\mathsf{T}\boldsymbol{S}^{-1}\boldsymbol{y} \tag{39}$$

$$\widehat{\mathrm{Var}}(\widehat{\boldsymbol{\beta}}^{\mathrm{GLS}}) = (\boldsymbol{X}^\mathsf{T}\boldsymbol{S}\boldsymbol{X})^{-1} \ , \tag{40}$$

where $\boldsymbol{y}$ and $\boldsymbol{X}$ are the stacked $\boldsymbol{y}_i$ and $\boldsymbol{X}_i$, and $\boldsymbol{S}$ is the block-diagonal matrix containing the $\boldsymbol{S}_i$. The fixed-effects model ignores unmodelled heterogeneity among the studies, however, and consequently the estimated effects in Eqn. (39) may be biased and the estimated uncertainty of these effects in Eqn. (40) may be too small.

The example we use here concerns the comparison of surgical (S) and non-surgical (NS) procedures for the treatment of moderate periodontal disease in five randomized split-mouth design clinical trials (Antczak-Bouckoms *et al.*, 1993, Berkey *et al.*, 1998). The two outcome measures for each patient were pre- to post-treatment changes after one year in probing depth (PD) and attachment level (AL), in mm, where successful treatment should decrease probing depth and increase attachment level. Each study was summarized by the mean difference, $\boldsymbol{y}_i = (\boldsymbol{y}_i^S - \boldsymbol{y}_i^{NS})$, between S and NS treated teeth, together with the covariance matrix $\boldsymbol{S}_i$ for each study. Sample sizes ranged from 14 to 89 across studies.

The left panel of Figure 29 shows the individual study estimates of PD and AL together with their covariances ellipses in a generic form that we propose as a more useful visualization of multivariate meta-analysis

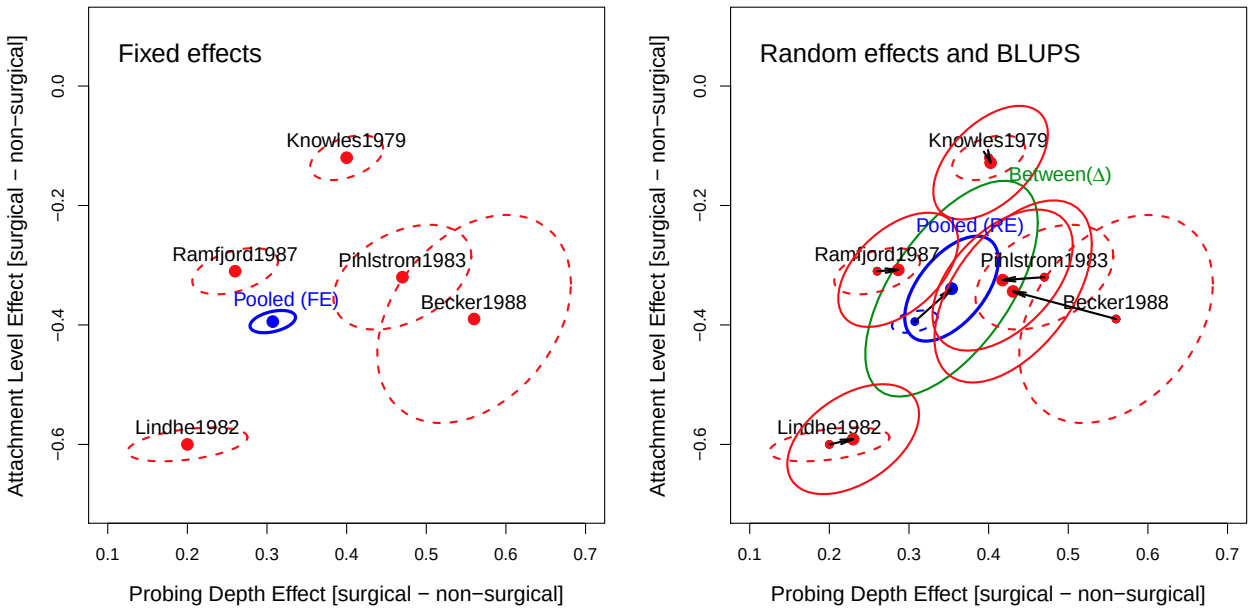

Figure 29: Multivariate meta-analysis visualizations for five periodontal treatment studies with outcome measures PD and AL. Left: Individual study estimates $\boldsymbol{y}_i$ and 40% standard ellipses for $\boldsymbol{S}_i$ (dashed, red) together with the pooled, fixed effects estimate and its associated covariance ellipse (blue). Right: BLUPs from the random effects multivariate meta-analysis model and their associated covariance ellipses (red, solid), together with the pooled, population averaged estimate and its covariance ellipse (blue), and the estimate of the between-study covariance matrix, Δ (green). Arrows show the differences between the FE and the RE models.

results than standard tabular displays: individual estimates plus model-based summary, all with associated covariance ellipsoids.[17]

It can be seen that all studies show that surgical treatment yields better probing depth (estimates are positive), while non-surgical treatment results in better attachment level (all estimates are negative). As well, within each study, there is a consistently positive correlation between the two outcome effects: patients with a greater surgical vs. non-surgical difference on one measure tend to have a greater such difference on the other, and greater within-study variation on PD than on AL.[18] The overall sizes of the ellipses largely reflect (inversely) the sample sizes in the various studies. As far as we know, these results were not noted in previous analyses of these data. Finally, the fixed-effect estimate, $\widehat{\boldsymbol{\beta}}^{\text{GLS}} = (0.307, -0.394)$, and its covariance ellipse suggest that these effects are precisely estimated.

The random-effects model is more complex because $\boldsymbol{\beta}$ and $\boldsymbol{\Delta}$ must be estimated jointly. A variety of methods have been proposed (full maximum likelihood, restricted maximum likelihood, method of moments, Bayesian methods, etc.), whose details (for which see Jackson *et al.*, 2011) are not relevant to the present discussion. Given an estimate $\widehat{\Delta}$, however, the pooled, population-averaged point estimate of effects under the random-effects model can be expressed as

$$\widehat{\boldsymbol{\beta}}^{\text{RE}} = \left(\sum_i \boldsymbol{X}_i^{\mathsf{T}} \boldsymbol{\Sigma}_i^{-1} \boldsymbol{X}_i \right)^{-1} \left(\sum_i \boldsymbol{X}_i^{\mathsf{T}} \boldsymbol{\Sigma}_i^{-1} \boldsymbol{y}_i \right) \quad , \tag{41}$$

where $\boldsymbol{\Sigma}_i = \boldsymbol{S}_i + \widehat{\boldsymbol{\Delta}}$. The first term in Eqn. (41) gives the estimated covariance matrix $\boldsymbol{V} \equiv \widehat{\text{Var}}(\widehat{\boldsymbol{\beta}}^{\text{RE}})$ of the

[17] The analyses described here were carried out using the **mvmeta** package for R (Gasparrini, 2012).

[18] Both PD and AL are measured on the same scale (mm), and the plots have been scaled to have unit aspect ratio, justifying this comparison.

random-effect pooled estimates. For the present example, this is shown in the right panel of Figure 29 as the blue ellipse. The green ellipse shows the estimate of the between-study covariance, $\widehat{\boldsymbol{\Delta}}$, whose shape indicates that studies with a larger estimate of PD also tend to have a larger estimate of AL (with correlation = 0.61). It is readily seen that, relative to the fixed-effects estimate $\widehat{\boldsymbol{\beta}}^{\text{GLS}}$, the unbiased estimate $\widehat{\boldsymbol{\beta}}^{\text{RE}} = (0.353, -0.339)$ under the random-effects model has been shifted toward the centroid of the individual study estimates and that its covariance ellipse is now considerably larger, reflecting between-study heterogeneity. In contrast to the fixed-effect estimates, inferences on $H_0 : \widehat{\boldsymbol{\beta}}^{\text{RE}} = \mathbf{0}$ pertain to the entire population of potential studies of these effects.

Figure 29 (right) also shows the best linear unbiased predictions of individual study estimates and their associated covariance ellipses, superposed (purely for didactic purposes) on the fixed-effects estimates to allow direct comparison. For random-effects models, the BLUPs have the form

$$\widehat{\boldsymbol{\beta}}_i^{\text{BLUP}} = \widehat{\boldsymbol{\beta}}^{\text{RE}} + \widehat{\boldsymbol{\Delta}}\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{y}_i - \widehat{\boldsymbol{\beta}}^{\text{RE}}) \ , \tag{42}$$

with covariance matrices

$$\widehat{\text{Var}}(\widehat{\boldsymbol{\beta}}_i^{\text{BLUP}}) = \boldsymbol{V} + (\widehat{\boldsymbol{\Delta}} - \widehat{\boldsymbol{\Delta}}\boldsymbol{\Sigma}_i^{-1}\widehat{\boldsymbol{\Delta}}) \ . \tag{43}$$

Algebraically, the BLUP outcome estimates in Eqn. (42) are thus a weighted average of the population-averaged estimates and the study-specific estimates, with weights depending on the relative sizes of the within- and between-study covariance matrices $\boldsymbol{S}_i$ and $\boldsymbol{\Delta}$. The point BLUPs borrow strength from the assumption of an underlying multivariate distribution of study parameters with covariance matrix $\boldsymbol{\Delta}$, shrinking toward the mean inversely proportional to the within-study covariance. Geometrically, these estimates may be described as occurring along the locus of osculation of the ellipses $\mathcal{E}(\boldsymbol{y}_i, \boldsymbol{S}_i)$ and $\mathcal{E}(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\Delta}})$.

Finally, the right panel of Figure 29 also shows the covariance ellipses of the BLUPs from Eqn. (42). It is clear that their orientation is a blending of the correlations in $\boldsymbol{V}$ and $\boldsymbol{S}_i$, and their size reflects the error in the average point estimates $\boldsymbol{V}$ and the error in the random deviation predicted for each study.

7 Discussion and Conclusions

> I know of scarcely anything so apt to impress the imagination as the wonderful form of cosmic order expressed by the "[Elliptical] Law of Frequency of Error." The law would have been personified by the Greeks and deified, if they had known of it. ... It is the supreme law of Unreason. Whenever a large sample of chaotic elements are taken in hand and marshaled in the order of their magnitude, an unsuspected and most beautiful form of regularity proves to have been latent all along.
>
> Sir Francis Galton, *Natural Inheritance*, London: Macmillan, 1889 ("[Elliptical]" added).

We have taken the liberty to add the word "Elliptical" to this famous quotation from Galton (1889). His "supreme law of Unreason" referred to univariate distributions of observations tending to the Normal distribution in large samples. We believe he would not take us remiss, and might perhaps welcome us for extending this view to two and more dimensions, where ellipsoids often provide an "unsuspected and most beautiful form of regularity."

In statistical data, theory, and graphical methods, one fundamental organizing distinction can be made depending on the dimensionality of the problem. A coarse but useful scale considers the essential defining distinctions to be among:

- ONE (univariate),
- TWO (bivariate),
- MANY (multivariate).

This scale[19] at least implicitly organizes much of current statistical teaching, practice, and software. But within this classification, the data, theory, and graphical methods are often treated separately (1D, 2D, nD), without regard to geometric ideas and visualizations that help to unify them.

This paper starts from the premise that one geometric form—the ellipsoid—provides a unifying framework for many statistical phenomena, with simple representations in 1D (a line) and 2D (an ellipse) that extend naturally to higher dimensions (an ellipsoid). The intellectual leap in statistical thinking from ONE to TWO in Galton (1886) was enormous. Galton's visual insights derived from the ellipse quickly led to an understanding of the ellipse as a contour of a bivariate normal surface. From here, the step from TWO to MANY would take another 20–30 years, but it is hard to escape the conclusion that geometric insight from the ellipse to the general ellipsoid in nD played an important role in the development of multivariate statistical methods.

In this paper, we have tried to show how ellipsoids can be useful tools for visual statistical thinking, data analysis, and pedagogy in a variety of contexts often treated separately and from a univariate perspective. Even in bivariate and multivariate problems, first-moment summaries (a 1D regression line or 2+D regression surface) show only part of the story—that of the expectation of a response $\boldsymbol{y}$ given predictors $\boldsymbol{X}$. In many cases, the more interesting part of the story concerns the *precision* of various methods of estimation, which we've shown to be easily revealed through data ellipsoids and elliptical confidence regions for parameters.

The general relationships among statistical methods, matrix algebra, and geometry are not new here. To our knowledge, Dempster (1969) was the first to exploit these relationships in a systematic fashion, establishing the connections among abstract vector spaces, algebraic coordinate systems, matrix operations and properties, the dualities between observation space and variable space, and the geometry of ellipses and projections.The roots of these connections go back much further—to Cramér (1946) (the idea of the concentration ellipsoid), to Pearson (1901) and Hotelling (1933) (principal components), and, we maintain, ultimately to Galton (1886). Throughout this development, elliptical geometry has played a fundamental role, leading to important visual insights.

The separate and joint roles of statistical computation and computational graphics should not be underestimated in appreciation of these developments. Dempster's analysis of the connections among geometry, algebra, and statistical methods was fueled by the development and software implementation of algorithms (Gram-Schmidt orthogonalization, Cholesky decomposition, sweep and multistandardize operators from Beaton, 1964) that allowed him to show precisely the translation of theoretical relations from abstract algebra to numbers and thence to graphs and diagrams.

Monette (1990) took these ideas several steps further, developing interactive 3D graphics focused on linear models, geometry, and ellipsoids, and demonstrating how many statistical properties and results could be understood through the geometry of ellipsoids. Yet, even at this later date, the graphical facilities of readily available statistical software were still rather primitive, and 3D graphics was available only on high-end workstations.

Several features of the current discussion may help to present these ideas in a new light. First, the examples we have presented rely heavily on software for statistical graphics developed separately and jointly by all three authors. These have allowed us to create what we hope are compelling illustrations, all statistically and geometrically exact, of the principles and ideas that form the body of the paper. Moreover, these are now general methods, implemented in a variety of R packages, e.g., (Fox and Weisberg, 2011, Friendly, 2007b) and a large collection of SAS macros (`http://datavis.ca/sasmac`), so we hope this paper will contribute to turning the theory we describe into practice.

Second, we have illustrated, in a wide variety of contexts, comprising all classical (Gaussian) linear models, multivariate linear models, and several extensions, how ellipsoids can contribute substantially to the understanding of statistical relationships, both in data analysis and in pedagogy. One graphical theme un-

[19]This idea, as a unifying classification principle for data analysis and graphics, was first suggested to the first author in seminars by John Hartigan at Princeton, c. 1968.

derlying a number of our examples is how the simple addition of ellipses to standard 2D graphical displays provides an efficient *visual* summary of important bivariate statistical quantities (means, variances, correlation, regression slopes, etc.) While first-moment visual summaries are now common adjuncts to graphical displays in standard software, often by default, we believe that the second-moment visual summaries of ellipses (and ellipsoids in 3D) now deserve a similar place in statistical practice.

Finally, we have illustrated several recent or entirely new visualizations of statistical methods and results, all based on elliptical geometry. HE plots for MANOVA designs (Friendly, 2007b) and their projections into canonical space (Section 5) provide one class of examples where ellipsoids provide simple visual summaries of otherwise complex statistical results. Our analysis of the geometry of added variable-plots suggested the idea of superposing marginal and conditional plots, as in Figure 18, leading to direct visualization of the difference between marginal and conditional relationships in linear models. The bivariate ridge trace plots described in Section 6.3.1 are a direct outgrowth of the geometric approach taken here, emphasizing the duality between views in data space and in parameter (β) space. We believe these all embody von Humboldt's (1811) dictum, quoted in the introduction.

8 Supplementary materials

All figures in this paper were constructed with either SAS or R software. The SAS examples use a collection of SAS macros from `http://datavis.ca/sasmac`; the R examples employ a variety of R packages available from the CRAN web site, `http://cran.us.r-project.org/` and the R-Forge development server at `https://r-forge.r-project.org/`. SAS and R scripts to generate many of the figures are included as supplementary materials for this article, along with several 3D movies.

9 Acknowledgments

This work was supported by Grant OGP0138748 from the National Sciences and Engineering Research Council of Canada to Michael Friendly, and by grants from the Social Sciences and Humanities Research Council of Canada to John Fox. We are grateful to Antonio Gasparini for helpful discussion regarding multivariate meta-analysis models, to Duncan Murdoch for advice on 3D graphics and comments on an earlier draft, and to the reviewers and Associate Editor for many helpful suggestions.

References

Alker, H. R. (1969). A typology of ecological fallacies. In M. Dogan and S. Rokkam, eds., *Social Ecology*, (pp. 69–86). Boston: The MIT Press.

Anderson, E. (1935). The irises of the Gaspé peninsula. *Bulletin of the American Iris Society*, 35, 2–5.

Antczak-Bouckoms, A., Joshipura, K., Burdick, E., and F., T. J. (1993). Meta-analysis of surgical versus non-surgical methods of treatment for periodontal disease. *Journal of Clinical Periodontology*, 20(4), 259–268.

Beaton, A. E. (1964). *The use of special matrix operators in statistical calculus*. Ed.d. thesis, Harvard University. Reprinted as Educational Testing Service Research Bulletin 64-51, Princeton, NJ.

Belsley, D. A., Kuh, E., and Welsch, R. E. (1980). *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. New York: John Wiley and Sons.

Berkey, C. S., Hoaglin, D. C., Antczak-Bouckoms, A., Mosteller, F., and Colditz, G. A. (1998). Meta-analysis of multiple outcomes by regression with random effects. *Statistics in Medicine*, 17(22), 2537–2550.

Boyer, C. B. (1991). Apollonius of Perga. In *A History of Mathematics, 2nd ed.*, (pp. 156–157). New York: John Wiley & Sons, Inc.

Bravais, A. (1846). Analyse mathématique sur les probabilités des erreurs de situation d'un point. *Mémoires présentés par divers savants à l'Académie royale des sciences de l'Institut de France*, 9, 255–332.

Bryant, P. (1984). Geometry, statistics, probability: Variations on a common theme. *The American Statistician*, 38(1), pp. 38–48.

Bryk, A. S. and Raudenbush, S. W. (1992). *Hierarchical linear models: Applications and data analysis methods*. Thousand Oaks, CA: Sage.

Campbell, N. A. and Atchley, W. R. (1981). The geometry of canonical variate analysis. *Systematic Zoology*, 30(3), 268–280.

Cramér, H. (1946). *Mathematical Models of Statistics*. Princeton, NJ: Princeton University Press.

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

Denis, D. (2001). The origins of correlation and regression: Francis Galton or Auguste Bravais and the error theorists. *History and Philosophy of Psychology Bulletin*, 13, 36–44.

Diez-Roux, A. V. (1998). Bringing context back into epidemiology: Variables and fallacies in multilevel analysis. *American Journal of Public Health*, 88, 216–222.

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

Fox, J. (2008). *Applied Regression Analysis and Generalized Linear Models*. Thousand Oaks, CA: Sage, 2nd edn.

Fox, J. and Suschnigg, C. (1989). A note on gender and the prestige of occupations. *Canadian Journal of Sociology*, 14, 353–360.

Fox, J. and Weisberg, S. (2011). *An R Companion to Applied Regression*. Thousand Oaks CA: Sage, 2nd edn.

Friendly, M. (1991). *SAS System for Statistical Graphics*. Cary, NC: SAS Institute, 1st edn.

Friendly, M. (2007a). A.-M. Guerry's *Moral Statistics of France*: Challenges for multivariable spatial analysis. *Statistical Science*, 22(3), 368–399.

Friendly, M. (2007b). HE plots for multivariate general linear models. *Journal of Computational and Graphical Statistics*, 16(2), 421–444.

Friendly, M. (2012). The generalized ridge trace plot: Visualizing bias and precision. *Journal of Computational and Graphical Statistics*, 21. In press; Accepted, 2/1/2012.

Galton, F. (1886). Regression towards mediocrity in hereditary stature. *Journal of the Anthropological Institute*, 15, 246–263.

Gasparrini, A. (2012). *mvmeta: multivariate meta-analysis and meta-regression.* R package version 0.2.4.

Guerry, A.-M. (1833). *Essai sur la statistique morale de la France*. Paris: Crochard. English translation: Hugh P. Whitt and Victor W. Reinking, Lewiston, N.Y. : Edwin Mellen Press, 2002.

Henderson, C. R. (1975). Best linear unbiased estimation and prediction under a selection model. *Biometrics*, 31, 423–448.

Hoerl, A. E. and Kennard, R. W. (1970a). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12, 55–67.

Hoerl, A. E. and Kennard, R. W. (1970b). Ridge regression: Applications to nonorthogonal problems (Corr: V12 p723). *Technometrics*, 12, 69–82.

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24, 417–441.

Jackson, D., Riley, R., and White, I. R. (2011). Multivariate meta-analysis: Potential and promise. *Statistics in Medicine*, 30, n/a–n/a.

Kramer, G. H. (1983). The ecological fallacy revisited: Aggregate- versus individual-level findings on economics and elections, and sociotropic voting. *The American Political Science Review*, 77(1), 92–111.

Laird, N. M. and Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics*, 38, 963–974.

Lichtman, A. J. (1974). Correlation, regression, and the ecological fallacy: A critique. *The Journal of Interdisciplinary History*, 4(3), 417–433.

Longley, J. W. (1967). An appraisal of least squares programs for the electronic computer from the point of view of the user. *Journal of the American Statistical Association*, 62, 819–841.

Marquardt, D. W. (1970). Generalized inverses, ridge regression, biased linear estimation, and nonlinear estimation. *Technometrics*, 12, 591–612.

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

Nam, I.-S., Mengersen, K., and Garthwaite, P. (2003). Multivariate meta-analysis. *Statistics in Medicine*, 22(14), 2309–2333.

Pearson, K. (1896). Contributions to the mathematical theory of evolution—III, regression, heredity and panmixia. *Philosophical Transactions of the Royal Society of London*, 187, 253–318.

Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, 6(2), 559–572.

Pearson, K. (1920). Notes on the history of correlation. *Biometrika*, 13(1), 25–45.

Raudenbush, S. W. and Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods*. Newbury Park, CA: Sage, 2nd edn.

Riley, M. W. (1963). Special problems of sociological analysis. In M. W. Riley, ed., *Sociological research I: a case approach*, (pp. 700–725). New York: Harcourt, Brace, and World.

Robinson, G. K. (1991). That blup is a good thing: The estimation of random effects. *Statistical Science*, 6(1), 15–32.

Robinson, W. S. (1950). Ecological correlations and the behavior of individuals. *American Sociological Review*, 15, 351–357.

Saville, D. and Wood, G. (1991). *Statistical methods: the geometric approach*. Springer texts in statistics. New York: Springer-Verlag.

Simpson, E. H. (1951). The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society, Series B*, 30, 238–241.

Speed, T. (1991). That BLUP is a good thing: The estimation of random effects: Comment. *Statistical Science*, 6(1), pp. 42–44.

Stigler, S. M. (1986). *The History of Statistics: The Measurement of Uncertainty before 1900*. Cambridge, MA: Harvard University Press.

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

Velleman, P. F. and Welsh, R. E. (1981). Efficient computing of regression diagnostics. *The American Statistician*, 35(4), 234–242.

von Humboldt, A. (1811). *Essai Politique sur le Royaume de la Nouvelle-Espagne. (Political Essay on the Kingdom of New Spain: Founded on Astronomical Observations, and Trigonometrical and Barometrical Measurements)*, vol. 1. New York: I. Riley. Eng. trans. by John Black.

Wickens, T. D. (1995). *The Geometry of Multivariate Statistics*. Hillsdale, NJ: Lawrence Erlbaum Associates.

A Geometrical and statistical ellipsoids

This appendix outlines useful results and properties concerning the representation of geometric and statistical ellipsoids. A number of these can be traced to or have more general descriptions within the abstract formulation of Dempster (1969), but casting them in terms of ellipsoids provides a simpler and more easily visualized framework.

A.1 Taxonomy and representation of ellipsoids

Section 2.1 defined a *proper* (origin-centered) ellipsoid in $\mathbb{R}^p$ by $\mathcal{E} := \{\boldsymbol{x} : \boldsymbol{x}^\mathsf{T}\boldsymbol{C}\boldsymbol{x} \leq 1\}$ that is bounded with non-empty interior (call these "fat" ellipsoids). For more general purposes, particularly for statistical applications, it is useful to give ellipsoids a wider definition. To give a complete taxonomy, this wider definition should also include ellipsoids that may be unbounded in some directions in $\mathbb{R}^p$ (an infinite cylinder of ellipsoidal cross-section) and degenerate (singular) ellipsoids that are "flat" in $\mathbb{R}^p$ with empty interior, such as when a 3D ellipsoid has no extent in one dimension (collapsing to an ellipse), or in two dimensions (collapsing to a line).

The motivation for this more general representation is to allow a notation for a set of general ellipsoids to be algebraicly closed under operations (a) image and preimage under a linear transformation and (b) inversion, where we can think about, visualize, and compute a linear transformation of an ellipsoid with central matrix $\boldsymbol{C}$ or its' inverse transformation via an analog of $\boldsymbol{C}^{-1}$, which apply equally to unbounded and/or degenerate ellipsoids. Applications concern the relationship between a predictor data ellipsoid and the corresponding $\boldsymbol{\beta}$ confidence ellipsoid (Section 4.6): The $\boldsymbol{\beta}$ will be unbounded (some linear combinations will have infinite confidence intervals) *iff* the corresponding data ellipsoid is flat, as when $p > n$.

Defining ellipsoids with $\{\boldsymbol{x} : \boldsymbol{x}^\mathsf{T}\boldsymbol{C}\boldsymbol{x} \leq 1\}$ produces proper ellipsoids for $\boldsymbol{C}$ positive definite and unbounded 'fat' ellipsoids for $\boldsymbol{C}$ positive semi-definite. But it does not produce degenerate (i.e. 'flat') ellipsoids. On the other hand, the representation in Eqn. (3), $\mathcal{E} := \boldsymbol{A}\mathcal{S}$ produces proper ellipsoids when $\boldsymbol{C} = \left(\boldsymbol{A}^\mathsf{T}\boldsymbol{A}\right)^{-1}$ where $\boldsymbol{A}$ is a non-singular $p \times p$ matrix and degenerate ellipsoids when $\boldsymbol{A}$ is a singular.

One representation that works for all of fat or flat *and* bounded or unbounded ellipsoids can be based on an SVD representation $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Delta}\boldsymbol{V}^{\mathsf{T}}$, with

$$\mathcal{E} := \boldsymbol{U}(\boldsymbol{\Delta}\mathcal{S}) \ , \tag{A.1}$$

where $\boldsymbol{U}$ is orthogonal and $\boldsymbol{\Delta}$ is diagonal with non-negative reals or infinity. The 'inverse' of an ellipsoid $\mathcal{E}$ is then simply $\boldsymbol{U}(\boldsymbol{\Delta}^{-1}\mathcal{S})$. The connection with traditional representations is that, if $\boldsymbol{\Delta}$ is finite, $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Delta}\boldsymbol{V}^{\mathsf{T}}$ where $\boldsymbol{V}$ can be any orthogonal matrix and if $\boldsymbol{\Delta}^{-1}$ is finite, $\boldsymbol{C} = \boldsymbol{U}\boldsymbol{\Delta}^{-2}\boldsymbol{U}^{\mathsf{T}}$.

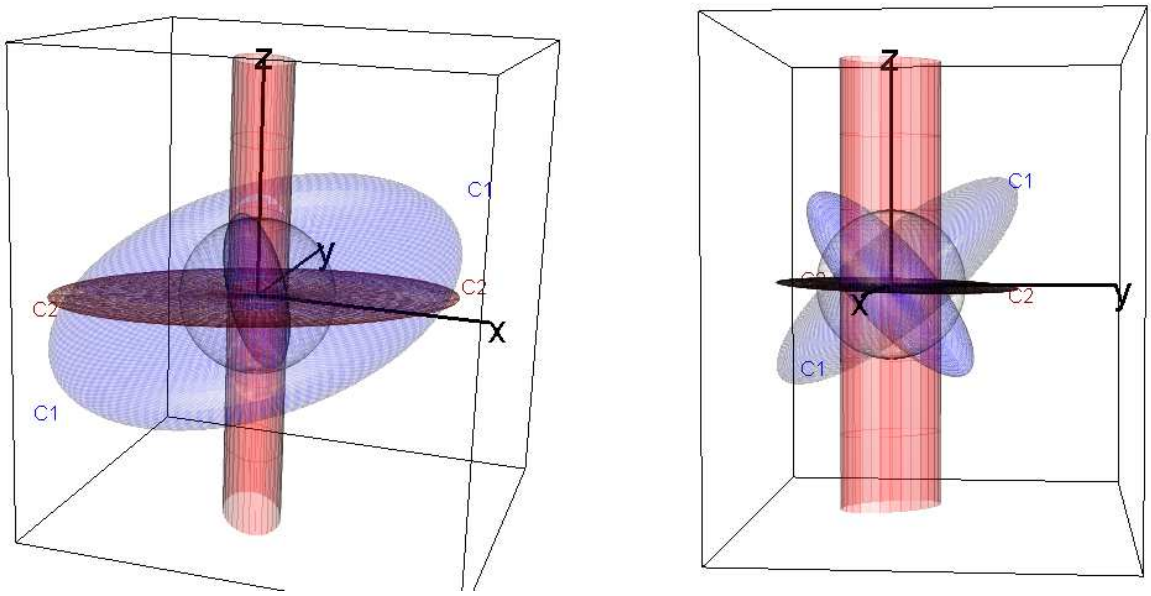

Figure A.1: Two views of an example of generalized ellipsoids. $\boldsymbol{C}_1$ (blue) determines a proper, fat ellipsoid; its inverse $\boldsymbol{C}_1^{-1}$ also generates a proper ellipsoid. $\boldsymbol{C}_2$ (red) determines an improper, flat ellipsoid, whose inverse $\boldsymbol{C}_2^{-1}$ is an unbounded cylinder of elliptical cross section. The scale of these images is defined by a unit sphere (gray). The right panel shows a view illustrating the orthogonality of each $\boldsymbol{C}$ and its dual, $\boldsymbol{C}^{-1}$.

Figure A.1 illustrates these ideas, using two generating matrices, $\boldsymbol{C}_1$ and $\boldsymbol{C}_2$ in this more general representation,

$$\boldsymbol{C}_1 = \begin{bmatrix} 6 & 2 & 1 \\ 2 & 3 & 2 \\ 1 & 2 & 2 \end{bmatrix} \ , \qquad \boldsymbol{C}_2 = \begin{bmatrix} 6 & 2 & 0 \\ 2 & 3 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

where $\boldsymbol{C}_1$ generates a proper ellipsoid and $\boldsymbol{C}_2$ generates an improper, flat ellipsoid. These varieties of ellipsoids are more easily seen in the 3D movies included in the online supplements.

A.2 Properties of geometric ellipsoids

- Translation: An ellipsoid centered at $\boldsymbol{x}_0$ has the definition $\mathcal{E} := \{\boldsymbol{x} : (\boldsymbol{x} - \boldsymbol{x}_0)^{\mathsf{T}}\boldsymbol{C}(\boldsymbol{x} - \boldsymbol{x}_0) = 1\}$ or $\mathcal{E} := \boldsymbol{x}_0 \oplus \boldsymbol{A}\mathcal{S}$ in the notation of Section 2.2.

- Orthogonality: If $\boldsymbol{C}$ is diagonal, the origin-centered ellipsoid has its axes aligned with the coordinate axes, and has the equation

$$\boldsymbol{x}^{\mathsf{T}}\boldsymbol{C}\boldsymbol{x} = c_{11}x_1^2 + c_{22}x_2^2 + \cdots + c_{pp}x_p^2 = 1 \ , \tag{A.2}$$

  where $1/\sqrt{c_{ii}} = c_{ii}^{-1/2}$ are the radii (semi-diameter lengths) along the coordinate axes.

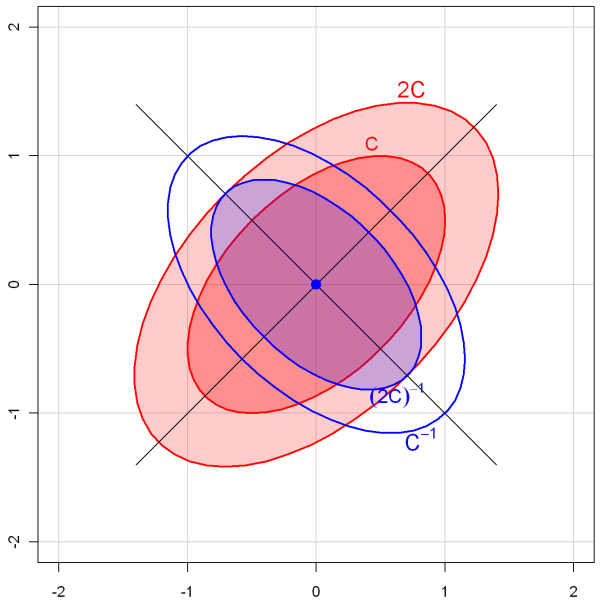

Figure A.2: Some properties of geometric ellipsoids. Principal axes of an ellipsoid are given by the eigenvectors of $\boldsymbol{C}$, with radii $1/\sqrt{\lambda_i}$. For an ellipsoid defined by Eqn. (1), the comparable ellipsoid for $2\boldsymbol{C}$ has radii multiplied by $1/\sqrt{2}$. The ellipsoid for $\boldsymbol{C}^{-1}$ has the same principal axes, but with radii $\sqrt{\lambda_i}$, making it small in the directions where $\boldsymbol{C}$ is large and vice-versa.

- Area and volume: In two dimensions, the area of the axis-aligned ellipse is $\pi(c_{11}c_{22})^{-1/2}$. For $p = 3$, the volume is $\frac{4}{3}\pi(c_{11}c_{22}c_{33})^{-1/2}$. In the general case, the hypervolume of the ellipsoid is proportional to $|\boldsymbol{C}|^{-1/2} = ||\boldsymbol{A}||$ and is given by $\pi^{p/2}\det(\boldsymbol{C})^{-1/2}/\left[\Gamma\left(\frac{p}{2}+1\right)\right]$, where the first two factors are familiar as the normalizing constant of the multivariate normal density function.

- Principal axes: In general, the eigenvectors, $\boldsymbol{v_i}, i = 1, \ldots, p$, of $\boldsymbol{C}$ define the principal axes of the ellipsoid and the inverse of the square roots of the ordered eigenvalues, $\lambda_1 > \lambda_2 \ldots, \lambda_p$, are the principal radii. Eigenvectors belonging to eigenvalues that are 0 are directions in which the ellipsoid is unbounded. With $\mathcal{E} = \boldsymbol{A}\mathcal{S}$, we consider the singular-value decomposition $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^{\mathsf{T}}$, with $\boldsymbol{U}$ and $\boldsymbol{V}$ orthogonal matrices and $\boldsymbol{D}$ a diagonal non-negative matrix with the same dimension as $\boldsymbol{A}$. The column vectors of $\boldsymbol{U}$, called the left singular vectors, correspond to the eigenvectors of $\boldsymbol{C}$ in the case of a proper ellipsoid. The positive diagonal elements of $\boldsymbol{D}$, $d_1 > d_2 > \ldots > d_p > 0$, are the principal radii of the ellipse with $d_i = 1/\sqrt{\lambda_i}$. In the singular case, the left singular vectors form a set of principal axes for the flattened ellipsoid.[20]

- Inverse: When $\boldsymbol{C}$ is positive definite, the eigenvectors of $\boldsymbol{C}$ and $\boldsymbol{C}^{-1}$ are identical, while the eigenvalues of $\boldsymbol{C}^{-1}$ are $1/\lambda_i$. It follows that the ellipsoid for $\boldsymbol{C}^{-1}$ has the same axes as that of $\boldsymbol{C}$, but with inversely proportional radii. In $\mathbb{R}^2$, the ellipsoid for $\boldsymbol{C}^{-1}$ is, with rescaling, a $90°$ rotation of the ellipsoid for $\boldsymbol{C}$, as illustrated in Figure A.2.

- Generalized inverse: A definition for an inverse ellipsoid that is equivalent in the case of proper ellipsoids,

$$\mathcal{E}^{-1} := \{\boldsymbol{y} : |\boldsymbol{x}^{\mathsf{T}}\boldsymbol{y}| \leq 1 \ , \quad \forall \boldsymbol{x} \in \mathcal{E}\} \ , \tag{A.3}$$

  generalizes to all ellipsoids. The inverse of a singular ellipsoid is an improper ellipsoid and vice versa.

[20] Corresponding left singular vectors and eigenvectors are not necessarily equal but sets that belong to the same eigenvalue/singular value span the same space.

- Dimensionality: The ellipsoid is bounded if $\boldsymbol{C}$ is positive definite (all $\lambda_i > 0$). Each $\lambda_i = 0$ increases the dimension of the space along which the ellipsoid is unbounded by one. For example, with $p = 3$, $\lambda_3 = 0$ gives a cylinder with an elliptical cross-section in 3-space, and $\lambda_2 = \lambda_3 = 0$ gives an infinite slab with thickness $2\sqrt{\lambda_1}$. With $\mathcal{E} = \boldsymbol{A}\mathcal{S}$, the dimension of the ellipsoid is equal to the number of positive singular values of $\boldsymbol{A}$.

- Projections: The projection of a p dimensional ellisoid into any subspace is $\boldsymbol{P}\mathcal{E}$, where $\boldsymbol{P}$ is an idempotent $p \times p$ (projection) matrix, i.e., $\boldsymbol{P}\boldsymbol{P} = \boldsymbol{P}^2 = \boldsymbol{P}$. For example, in $\mathbb{R}^2$ and $\mathbb{R}^3$, the matrices

$$\boldsymbol{P}_2 = \begin{bmatrix} 1 & 1 \\ 0 & 0 \end{bmatrix} \ , \qquad \boldsymbol{P}_3 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

  project, respectively, an ellipse onto the line $x_1 = x_2$, and an ellipsoid into the (x_1, x_2) plane. If $\boldsymbol{P}$ is symmetric, then $\boldsymbol{P}$ is the matrix of an orthogonal projection, and it is easy to visualize $\boldsymbol{P}\mathcal{E}$ as the shadow of $\mathcal{E}$ cast perpendicularly onto $\mathrm{span}(\boldsymbol{P})$. Generally, $\boldsymbol{P}\mathcal{E}$ is the shadow of $\mathcal{E}$ onto $\mathrm{span}(\boldsymbol{P})$ along the null space of $\boldsymbol{P}$.

- Linear transformations: A linear transformation of an ellipsoid is an ellipsoid, and the pre-image of an ellipsoid under a linear transformation is an ellipsoid. A non-singular linear transformation maps a proper ellipsoid into a proper ellipsoid.

- Slopes and tangents: The slopes of the ellipsoidal surface in the directions of the coordinate axes are given by $\partial/\partial\boldsymbol{x}\ (\boldsymbol{x}^\mathsf{T}\boldsymbol{C}\boldsymbol{x}) = 2\boldsymbol{C}\boldsymbol{x}$. From this, it follows that the tangent hyperplane to the unit ellipsoidal surface at the point $\boldsymbol{x}_\alpha$, where $\boldsymbol{x}_\alpha^\mathsf{T}\,\partial/\partial\boldsymbol{x}(\boldsymbol{x}^\mathsf{T}\boldsymbol{C}\boldsymbol{x}) = 0$, has the equation $\boldsymbol{x}_\alpha^\mathsf{T}\boldsymbol{C}\boldsymbol{x} = 1$.

A.3 Conjugate axes and inner-product spaces

For any non-singular $\boldsymbol{A}$ in Eqn. (5) that generates an ellipsoid, the columns of $\boldsymbol{A} = [\boldsymbol{a}_1, \boldsymbol{a}_2, \cdots, \boldsymbol{a}_p]$ form a set of "conjugate axes" of the ellipse. (Two diameters are conjugate *iff* the tangent line at the endpoint of one diameter is parallel to the other diameter.) Each vector $\boldsymbol{a}_i$ lies on the ellipse, and the tangent hyperplane at that point is parallel to the span of all the other column vectors of $\boldsymbol{A}$. For $p = 2$ this result is illustrated in Figure A.3 (left) in which

$$\boldsymbol{A} = \begin{bmatrix} \boldsymbol{a}_1 & \boldsymbol{a}_2 \end{bmatrix} = \begin{bmatrix} 1 & 1.5 \\ 2 & 1 \end{bmatrix} \ \Rightarrow\ \boldsymbol{W} = \boldsymbol{A}\boldsymbol{A}^\mathsf{T} = \begin{bmatrix} 3.25 & 3.5 \\ 3.5 & 5 \end{bmatrix} \ . \tag{A.4}$$

Consider the inner-product space with inner product matrix $\boldsymbol{W}^{-1} = \begin{bmatrix} 1.25 & -0.875 \\ -0.875 & 0.8125 \end{bmatrix}$ and inner product

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle = \boldsymbol{x}'\boldsymbol{W}^{-1}\boldsymbol{y} \ .$$

Because $\boldsymbol{A}^\mathsf{T}\boldsymbol{W}^{-1}\boldsymbol{A} = \boldsymbol{A}^\mathsf{T}\left(\boldsymbol{A}\boldsymbol{A}^\mathsf{T}\right)^{-1}\boldsymbol{A} = \boldsymbol{A}^\mathsf{T}(\boldsymbol{A}^\mathsf{T})^{-1}\boldsymbol{A}^{-1}\boldsymbol{A} = \boldsymbol{I}$, we see that $\boldsymbol{a}_1$ and $\boldsymbol{a}_2$ are orthogonal unit vectors (in fact, an orthonormal basis) in this inner product:

$$\begin{aligned} \langle \boldsymbol{a}_i, \boldsymbol{a}_i \rangle &= \boldsymbol{a}_i^\mathsf{T}\boldsymbol{W}^{-1}\boldsymbol{a}_i = 1 \\ \langle \boldsymbol{a}_1, \boldsymbol{a}_2 \rangle &= \boldsymbol{a}_1'\boldsymbol{W}^{-1}\boldsymbol{a}_2 = 0 \ . \end{aligned}$$

Now, if $\boldsymbol{W} = \boldsymbol{B}\boldsymbol{B}^\mathsf{T}$ is any other factorization of $\boldsymbol{W}$, then the columns of $\boldsymbol{B}$ have the same properties as the columns of $\boldsymbol{A}$. Particular factorizations yield interesting and statistically useful sets of conjugate axes. The illustration in Figure A.3 (right) shows two such cases with special properties: In the Choleski

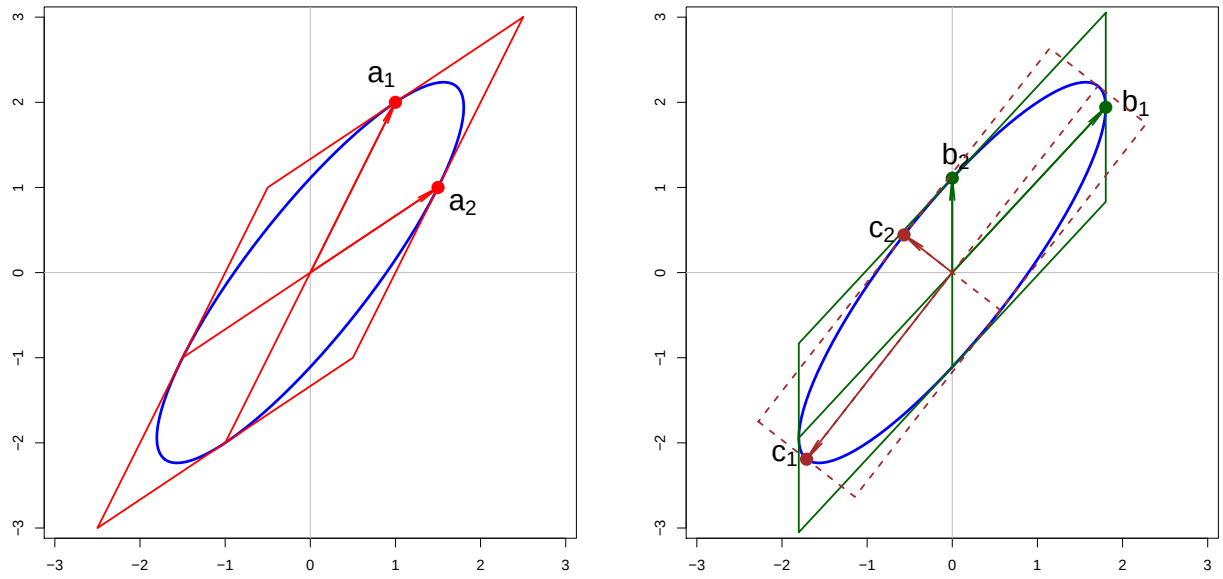

Figure A.3: Conjugate axes of an ellipsoid with various factorizations of $\boldsymbol{W}$ and corresponding basis vectors. The conjugate vectors lie on the ellipsoid, and their tangents can be extended to form a parallelogram framing it. Left: for an arbitrary factorization, given in Eqn. (A.4). Right: for the Choleski factorization (solid, green, b_1, b_2) and the principal component factorization (dashed, brown, c_1, c_2).

factorization (shown solid in green), where $\boldsymbol{B}$ is lower triangular, the last conjugate axis, $\boldsymbol{b}_2$, is aligned with the coordinate axis $\boldsymbol{x}_2$. Each previous axis ($\boldsymbol{b}_1$, here) is the orthogonal complement to all later axes in the inner-product space of $\boldsymbol{W}^{-1}$. The Choleski factorization is unique in this respect, subject to a permutation of the rows and columns of $\boldsymbol{W}$. The subspace $\{c_1\boldsymbol{b}_1 + ... + c_{p-1}\boldsymbol{b}_{p-1}, c_i \in \mathbb{R}\}$, is the plane of the regression of the last variable on the others, a fact that generalizes naturally to ellipsoids that are not necessarily centerered at the origin.

In the principal-component (PC) factorization (shown dashed in brown) $\boldsymbol{W} = \boldsymbol{C}\boldsymbol{C}^\mathsf{T}$, where $\boldsymbol{C} = \boldsymbol{\Gamma}\boldsymbol{\Lambda}^{1/2}$ and hence $\boldsymbol{W} = \boldsymbol{\Gamma}\boldsymbol{\Lambda}\boldsymbol{\Gamma}'$ is the spectral decomposition of $\boldsymbol{W}$. Here, the ellipse axes are orthogonal in the space of the ellipse (so the bounding tangent parallelogram is a rectangle) *as well as* in the inner-product space of $\boldsymbol{W}^{-1}$. The PC factorization is unique in this respect (up to reflections of the axis vectors).

As illustrated in Figure A.3, each pair of conjugate axes has a corresponding bounding tangent parallelogram. It can be shown that all such parallelograms have the same area and equal sums of squares of the lengths of their diameters.

A.4 Ellipsoids in a generalized metric space

In Appendix A.3, we considered the positive semi-definite matrix $\boldsymbol{W}$ and corresponding ellipsoid to be referred to a Euclidean space, perhaps with different basis vectors. We showed that various measures of the "size" of the ellipsoid could be defined in terms of functions of the eigenvalues λ_i of $\boldsymbol{W}$.

We now consider the generalized case of an analogous $p \times p$ positive semi-definite symmetric matrix $\boldsymbol{H}$, but where measures of length, distance, and angles are referred to a metric defined by a positive-definite symmetric matrix $\boldsymbol{E}$. As is well known, the generalized eigenvalue problem is to find the scalars λ_i and vectors $\boldsymbol{v}_i, i = 1, 2, \ldots p$, such that $\boldsymbol{H}\boldsymbol{v} = \lambda\boldsymbol{E}\boldsymbol{v}$, that is, the roots of $\det(\boldsymbol{H} - \lambda\boldsymbol{E}) = 0$.

For such $\boldsymbol{H}$ and $\boldsymbol{E}$, we can always find a factor $\boldsymbol{A}$ of $\boldsymbol{E}$, so that $\boldsymbol{E} = \boldsymbol{A}\boldsymbol{A}^\mathsf{T}$, whose colums will be conjugate directions for $\boldsymbol{E}$ and whose rows will also be conjugate directions for $\boldsymbol{H}$, in that $\boldsymbol{H} = \boldsymbol{A}^\mathsf{T}\boldsymbol{D}\boldsymbol{A}$, where $\boldsymbol{D}$ is diagonal. Geometrically, this means that there exists a unique pair of bounding parallelograms

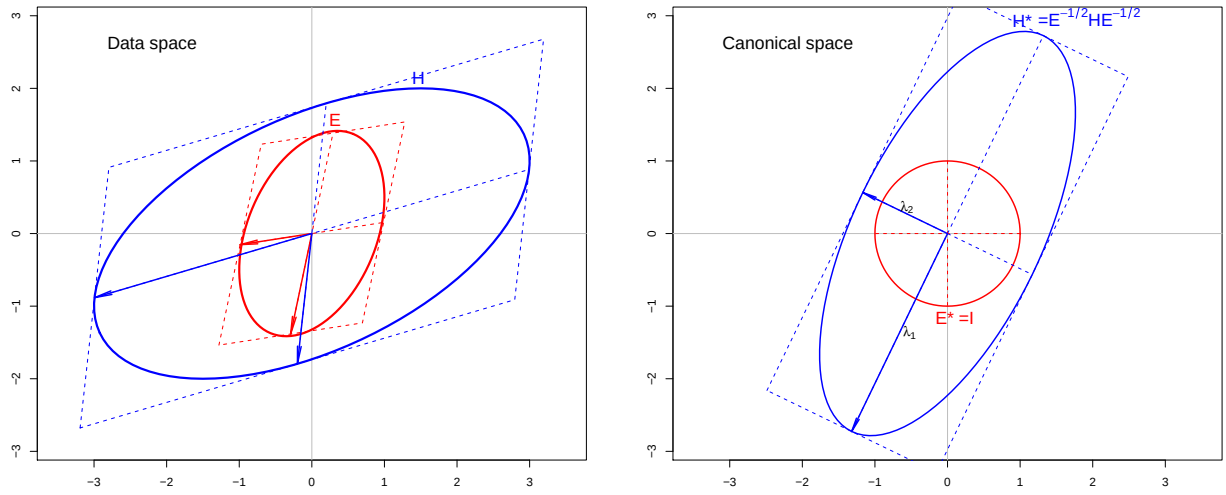

Figure A.4: Left: Ellipses for $\boldsymbol{H}$ and $\boldsymbol{E}$ in Euclidean "data space." Right: Ellipses for $\boldsymbol{H}^\star$ and $\boldsymbol{E}^\star$ in the transformed "canonical space," with the eigenvectors of $\boldsymbol{H}$ relative to $\boldsymbol{E}$ shown as blue arrows, whose radii are the corresponding eigenvalues, λ_1, λ_2.

for the $\boldsymbol{H}$ and $\boldsymbol{E}$ ellipsoids whose corresponding sides are parallel. A linear transformation of $\boldsymbol{E}$ and $\boldsymbol{H}$ that transforms the parallelogram for $\boldsymbol{E}$ to a square (or cuboid), and hence $\boldsymbol{E}$ to a sphere (or spheroid), generates an equivalent view in what we describe below as canonical space.

In statistical applications (e.g., MANOVA, canonical correlation), the generalized eigenvalue problem is transformed to an ordinary eigenvalue problem by considering the following equivalent forms with the same λ_i, $\boldsymbol{v}_i$,

$$\begin{aligned} (\boldsymbol{H} - \lambda \boldsymbol{E})\boldsymbol{v} &= \boldsymbol{0} \\ \Rightarrow (\boldsymbol{H}\,\boldsymbol{E}^{-1} - \lambda \boldsymbol{I})\boldsymbol{v} &= \boldsymbol{0} \\ \Rightarrow (\boldsymbol{E}^{-1/2}\,\boldsymbol{H}\,\boldsymbol{E}^{-1/2} - \lambda \boldsymbol{I})\boldsymbol{v} &= \boldsymbol{0}\ , \end{aligned}$$

where the last form gives a symmetric matrix, $\boldsymbol{H}^\star = \boldsymbol{E}^{-1/2}\,\boldsymbol{H}\,\boldsymbol{E}^{-1/2}$. Using the square root of $\boldsymbol{E}$ defined by the principal-component factorization $\boldsymbol{E}^{1/2} = \boldsymbol{\Gamma}\boldsymbol{\Lambda}^{1/2}$ produces the ellipsoid $\boldsymbol{H}^\star$, the orthogonal axes of which correspond to the $\boldsymbol{v}_i$, whose squared radii are the corresponding eigenvalues λ_i. This can be seen geometrically as a rotation of "data space" to an orientation defined by the principal axes of $\boldsymbol{E}$, followed by a re-scaling, so that the $\boldsymbol{E}$ ellipsoid becomes the unit spheroid. In this transformed space ("canonical space"), functions of the the squared radii λ_i of the axes of $\boldsymbol{H}^\star$ give direct measures of the "size" of $\boldsymbol{H}$ relative to $\boldsymbol{E}$. The orientation of the eigenvectors $\boldsymbol{v}_i$ can be related to the (orthogonal) linear combinations of the data variables that are successively largest in the metric of $\boldsymbol{E}$.

To illustrate, Figure A.4 (left) shows the ellipses generated by

$$\boldsymbol{H} = \begin{bmatrix} 9 & 3 \\ 3 & 4 \end{bmatrix} \quad \text{and} \quad \boldsymbol{E} = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 2 \end{bmatrix}$$

together with their conjugate axes. For $\boldsymbol{E}$, the conjugate axes are defined by the columns of the right factor, $\boldsymbol{A}^\mathsf{T}$, in $\boldsymbol{E} = \boldsymbol{A}\boldsymbol{A}^\mathsf{T}$; for $\boldsymbol{H}$, the conjugate axes are defined by the columns of $\boldsymbol{A}$. The transformation to $\boldsymbol{H}^\star = \boldsymbol{E}^{-1/2}\,\boldsymbol{H}\,\boldsymbol{E}^{-1/2}$ is shown in the right panel of Figure A.4. In this "canonical space," angles and lengths have the orginary interpretation of Euclidean space, so the size of $\boldsymbol{H}^\star$ can be interpreted directly in terms of functions of the radii $\sqrt{\lambda_1}$ and $\sqrt{\lambda_2}$.