Elliptic Insights: Understanding Statistical Methods through Elliptic Geometry

Michael Friendly
York University

Georges Monette
York University

May 13, 2011

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 general linear models, multivariate linear models and mixed models. We emphasize the strong relations among statistical methods, matrix algebraic solutions and geometry that can often be easily understood in terms of ellipses.

Key words: Galton; Bayesian estimation; added variable plots; concentration ellipse; data ellipse; discriminant analysis; hypothesis-error plots; mixed models; regression paradoxes; statistical geometry; ridge regression;

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 relation 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, Ave. $(Y \mid X)$, and, for symmetry, 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 descend from these visual insights:[1] correlation and regression (Pearson, 1896), the bivariate normal distribution, principal components (Pearson, 1901, Hotelling, 1933) all descend from Galton's geometrical diagram.[2]

Basic geometry goes back 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.

Over time, the ellipse would be called to duty to provide simple explanations of phenomena once thought complex. Most notable is Keplers 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 observations by Tycho Brahe and others by a simple law: "The orbit of every planet is an ellipse with the sun at a focus." One century later, Isaac Newton was able to derive 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. We focus here on visual insights through ellipses in the areas of linear models, multivariate models and mixed models.

2 Notation and basic results

There are various representations of an ellipse (or ellipsoid in three or more dimensions), both geometric and statistical. To avoid repeating "or ellipsoid" we use "ellipsoid" as generic where context is clear. We make use of the following definitions and results in what follows for geometric and statistical properties.

2.1 Geometrical ellipsoids

A general unit ellipsoid in the p-dimensional space $\mathbb{R}^p$ centered at the origin, $\mathbf{0}$, may be defined by the quadratic form

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

where $\boldsymbol{x} = (x_1, x_2, \ldots, x_p)^\mathsf{T}$ is a vector referring to the coordinate axes and $\boldsymbol{C}$ is a symmetric non-negative definite $p \times p$ matrix. Some useful properties are:

- Translation: The unit ellipsoid centered at $\boldsymbol{x}_0$ has the equation $\mathcal{E} := \{\boldsymbol{x} : (\boldsymbol{x} - \boldsymbol{x}_0)^\mathsf{T} \boldsymbol{C} (\boldsymbol{x} - \boldsymbol{x}_0) = 1\}$.

[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."

[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' contribution and the correlation coefficient is often called the Bravais-Pearson coefficient in France (Denis, 2001).

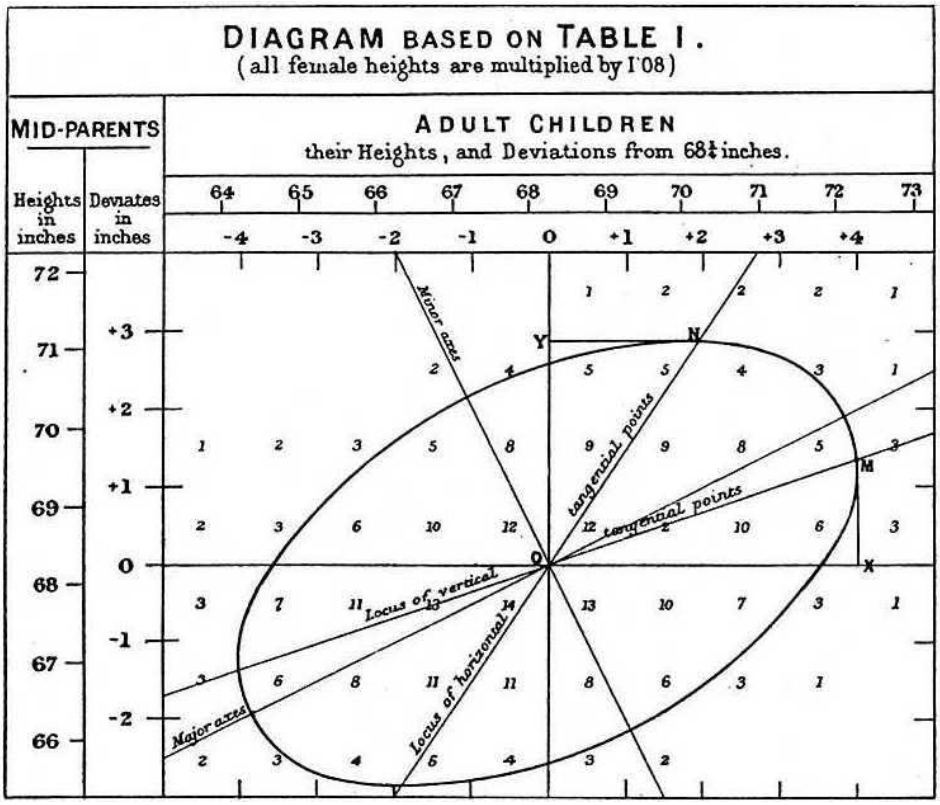

Figure 1: Galton's 1886 diagram, showing the relation of height of children to the average of their parents' height. The diagram is essentially an overlay of a geometrical interpretation on a bivariate grouped frequency distribution, shown as numbers.

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

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

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

- 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 $|\boldsymbol{C}|^{-1/2}$ and is given by $\pi^{p/2}/\det(\boldsymbol{C})^{1/2}\Gamma(\frac{p}{2}+1)$.

- 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.

- Inverse: Since 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.

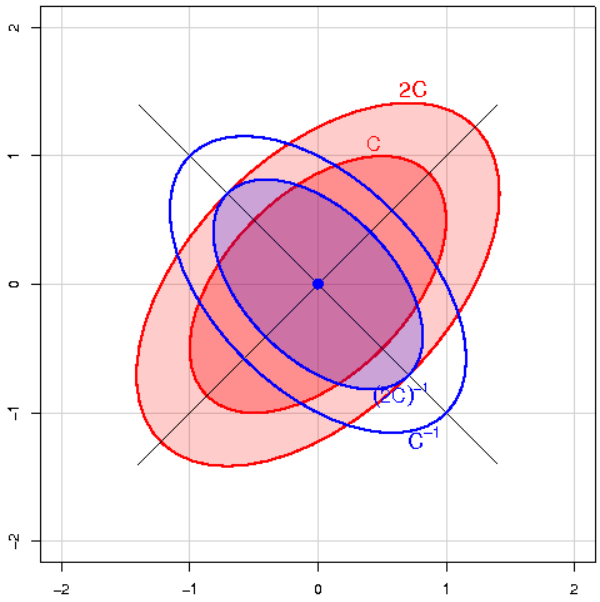

Figure 2: Some properties of geometric ellipsoids. Principal axes of an ellipsoid are given by the eigenvectors of $\boldsymbol{C}$, with radii $\sqrt{\lambda_i}$. For a standard unit ellipsoid defined by Eqn. (1), the comparable ellipsoid for $2\boldsymbol{C}$ has radii multiplied by $\sqrt{2}$. The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

In $\mathbb{R}^2$ (with appropriate scaling of the axes), the ellipsoid for $\boldsymbol{C}^{-1}$ is thus a $90°$ rotation of the ellipsoid for $\boldsymbol{C}$, as illustrated in Figure 2.

- Dimensionality: The ellipsoid is only p-dimensional if $\boldsymbol{C}$ is positive definite (all $\lambda_i > 0$). Each $\lambda_i = 0$ reduces dimensionality by one. For example, with $p = 3$, $\lambda_3 = 0$ gives a 2D ellipse in 3-space, and $\lambda_2 = \lambda_3 = 0$ gives a degenerate line.

- Projections: The projection of a p dimensional ellisoid into any subspace is $\boldsymbol{x}^\mathsf{T}(\boldsymbol{P}\boldsymbol{C}\boldsymbol{P}^\mathsf{T})\boldsymbol{x} = 1$, where $\boldsymbol{P}$ is an idempotent $p \times p$ 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.

- 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 ellopsoidal 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 $\boldsymbol{x}_\alpha^\mathsf{T}\boldsymbol{C}\boldsymbol{x} = 1$.

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 notations:

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

$$\mathcal{E} = \{\boldsymbol{x} : (\boldsymbol{x} - \boldsymbol{\mu})^\mathsf{T} \boldsymbol{\Sigma}^{-1} (x - \boldsymbol{\mu}) = 1\} \ . \tag{3}$$

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:	$\det(\boldsymbol{\Sigma}) = \prod^p \lambda_i$	area, volume	geometric mean
(b) Average variance:	$\mathrm{tr}(\boldsymbol{\Sigma}) = \sum^p \lambda_i$	linear sum	arithmetic mean
(c) Average variance:	$1/\mathrm{tr}(\boldsymbol{\Sigma^{-1}}) = 1/\sum^p (1/\lambda_i)$		harmonic mean
(d) Maximal variance:	λ_1	maximum dimension	supremum

In multivariate tests, these correspond (with suitable transformations) to (a) Wilks' Λ, (b) Pillai trace criteria, (c) Hotelling-Lawley and (d) Roy's maximum root test, as we describe below.

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

$$E(\boldsymbol{\mu}, \boldsymbol{W}) = \boldsymbol{\mu} + \boldsymbol{A}\mathcal{S} \ , \tag{4}$$

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 Section 2.3. Thus, for some results described below, a convenient notation in terms of $\boldsymbol{W}$ is

$$E(\boldsymbol{\mu}, \boldsymbol{W}) = \boldsymbol{\mu} \oplus \sqrt{\boldsymbol{W}} = \boldsymbol{\mu} \oplus \boldsymbol{W}^{1/2} \ , \tag{5}$$

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, $\boldsymbol{\mu} \oplus \boldsymbol{B} = \boldsymbol{\nu} \oplus \boldsymbol{C}$ (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, it is readily seen that under a linear transformation given by a matrix $\boldsymbol{L}$ the image of the ellipse is:

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

2.3 Conjugate axes and inner product spaces

For any $\boldsymbol{A}$ 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 space at that point is parallel to the span of all the other column vectors of $\boldsymbol{A}$.

For $p = 2$ this is illustrated in Figure 3(a) in which

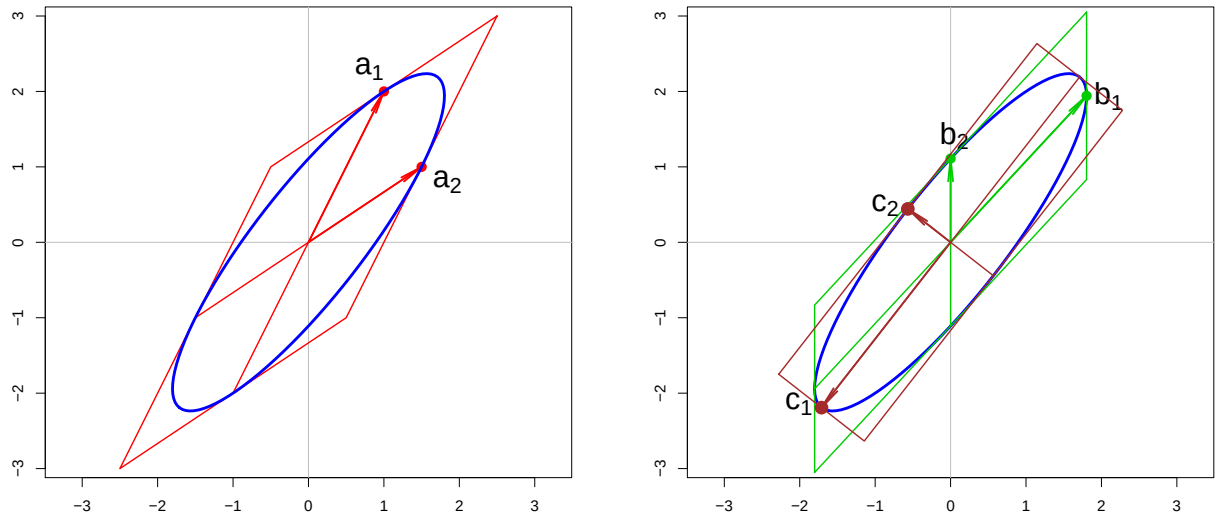

Figure 3: Conjugate axes of an ellipsoid with various factorizations of $\boldsymbol{W}$ and corresponding basis vectors. The conjugate vectors lie on the ellipsoid, and their tangents can be extended to form a parallelogram framing it. (a) Left: for an arbitrary factorization, given in Eqn. (6). (b) Right: for the Choleski factorization (green) and the principal component factorization (brown).

$$\boldsymbol{A} = \begin{bmatrix} \boldsymbol{a}_1 & \boldsymbol{a}_2 \end{bmatrix} = \begin{bmatrix} 1 & 1.5 \\ 2 & 1 \end{bmatrix} \;\Rightarrow\; \boldsymbol{W} = \boldsymbol{A}\boldsymbol{A}^{\mathsf{T}} = \begin{bmatrix} 3.25 & 3.5 \\ 3.5 & 5 \end{bmatrix} \;. \tag{6}$$

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

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

Since $\boldsymbol{A}^{\mathsf{T}}\boldsymbol{W}^{-1}\boldsymbol{A} = \boldsymbol{A}^{\mathsf{T}}\left(\boldsymbol{A}\boldsymbol{A}^{\mathsf{T}}\right)^{-1}\boldsymbol{A} = \boldsymbol{A}^{\mathsf{T}}(\boldsymbol{A}^{\mathsf{T}})^{-1}\boldsymbol{A}^{-1}\boldsymbol{A} = \boldsymbol{I}$, 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 3(b) shows two such cases with special properties: In the Choleski factorization (shown 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 $\boldsymbol{W}$.

In the principal component (PC) factorization (shown 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.

As illustrated in Figure 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.

2.4 Ellipsoids in a generalized metric space

ToDo:Geo

ToDo: Smooth out and simplify this description. Add to the plot: unit vectors in data space, and their transformations in canonical space.

In the discussion above, 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 $\det(\boldsymbol{H} - \lambda \boldsymbol{E}) = 0$.

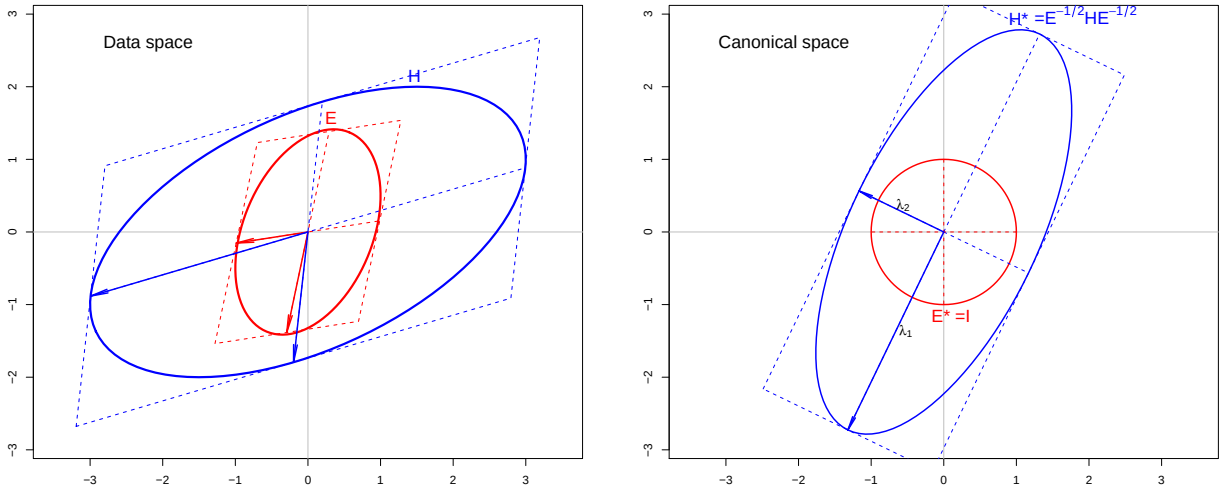

Figure 4: Left: Ellipses for $\boldsymbol{H}$ and $\boldsymbol{E}$ in Euclidean "data space". Right: Ellipses for $\boldsymbol{H}^\star$ and $\boldsymbol{E}^\star$ in the transformed "canonical space", with the eivenvectors of$\boldsymbol{H}$ relative to $\boldsymbol{E}$ shown as blue arrows, whose radii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

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 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 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 equivalent forms with the same λ_i, $\boldsymbol{v}_i$,

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

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 $\boldsymbol{E}^{1/2} = \boldsymbol{\Gamma}\boldsymbol{\Lambda}^{1/2}$ gives the ellipsoid of $\boldsymbol{H}^\star$ orthogonal axes corresponding to the $\boldsymbol{v}_i$, whose 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 sphere. In this transformed space ("canonical space"), functions of the the 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 which are successively largest in the metric of $\boldsymbol{E}$.

To illustrate, Figure 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 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 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 λ_1 and λ_2.

3 The data ellipse and ellipsoids

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. It 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) actual data, from which he drew Figure 1 as a conceptual diagram,[3] shown in Figure 5, where the frequency at each point is shown by a sunflower symbol. We also overlay the 40%, 68% and 95% data ellipses, as described below.

In Figure 5 the ellipses have the mean vector $(\bar{x}, \bar{y})$ as their center; the lengths of arms of the central cross show the standard deviation of each variable, which may be seen to correspond to the shadows of the 40% ellipse. In addition, the correlation coefficient may be visually estimated as the fraction of a vertical tangent line from $\bar{y}$ to the top of the ellipse that is below the regression line $\hat{y}|x$, shown by the arrow labeled 'r.' Finally, as Galton noted, the regression line for $\hat{y} \,|\, x$ (or $\hat{x} \,|\, y$) can be visually estimated as the locus of the points of vertical (or horizontal) tangents with the family of concentric ellipses. See Monette (1990, Fig.

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

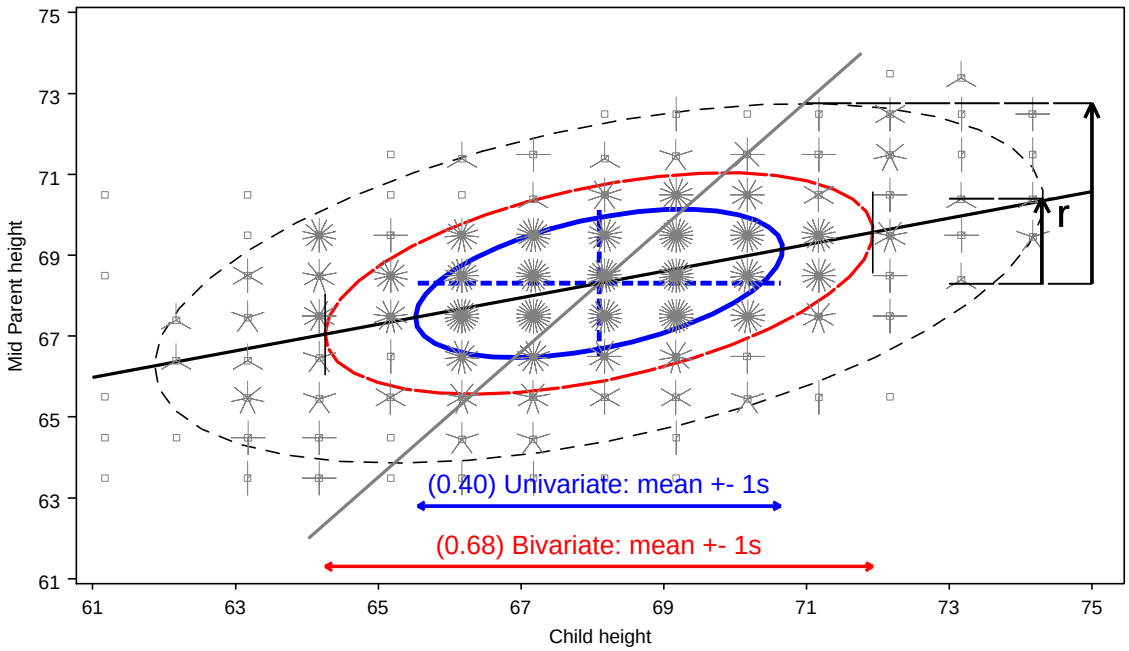

Figure 5: Sunflower plot of Galton's data on heights of parents and their children (in.), with 40%, 68% and 95% data ellipses and the regression lines of y on x (black) and x on y (grey). The ratio of the vertical to the regression line (labeled 'r') to the vertical to the top of the ellipse gives a visual estimate of the correlation (r=0.46, here). Shadows (projections) on the coordinate axes give standard intervals, $\bar{x}_i \pm s_i$, with various coverage properties. Plotting children's height on the abscissa follows Galton.

5.1–5.2) and Friendly (1991, p. 183) for illustrations and further discussion of the properties of the data ellipse.

More formally (Dempster, 1969, Monette, 1990), for a p-dimensional sample, $\mathbf{Y}_{n\times p}$, we recognize the quadratic form in Eqn. (3) 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 define the 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,

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

where $\boldsymbol{S}$ is the sample variance-covariance matrix, $\boldsymbol{S} = (n-1)^{-1} \sum_{i=1}^n (\boldsymbol{y}_i - \bar{\boldsymbol{y}})^\mathsf{T} (\boldsymbol{y}_i - \bar{\boldsymbol{y}})$. In the computational notation of Eqn. (5), the boundary of the data ellipsoid of radius c is thus

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

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 represents a contour of constant density in their joint distribution. In this case $D^2_M(\boldsymbol{y})$ 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 5, 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, $\bar{y} \pm 1s$. The same property of univariate coverage pertains to any linear combination of y_1 and y_2.
- 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 $\bar{y} \pm 1s$ interval. The univariate shadows, or those of any linear combination, then correspond to standard intervals taking fishing in a $p = 2$-dimensional space into account.

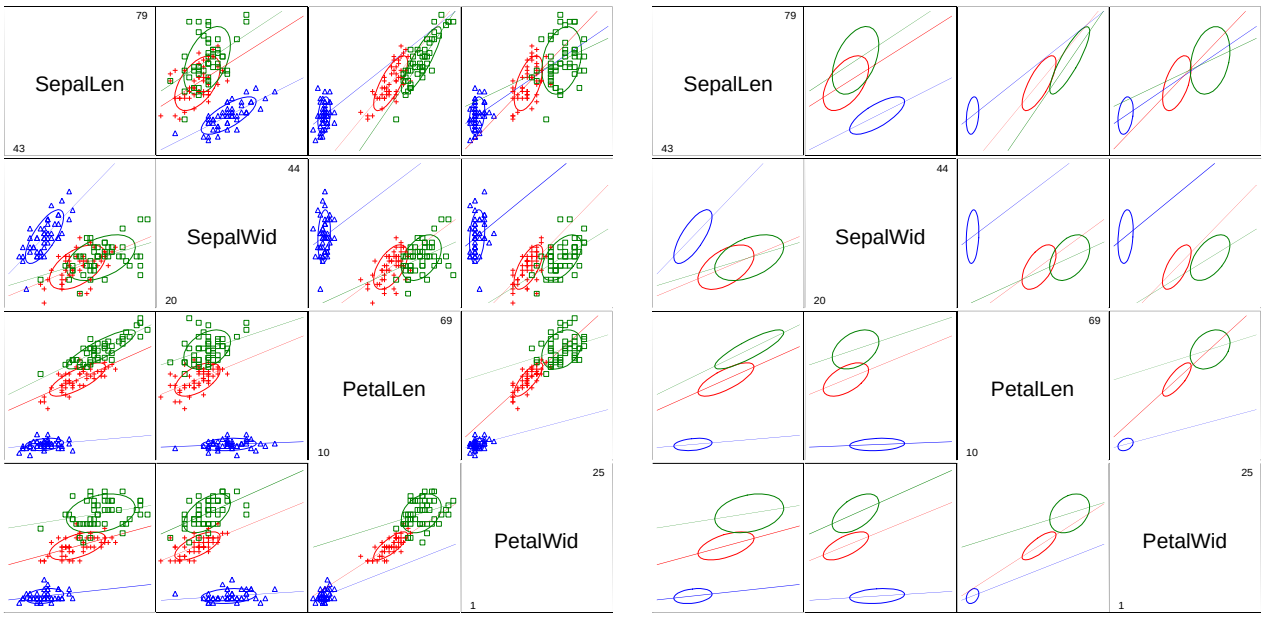

Figure 6: Scatterplot matrices of Anderson's iris data: (a) showing data, separate 68% data ellipses and regression lines for each species; (b) showing only ellipses and regression lines. Key– *Iris setosa*: blue, △s; *Iris versicolor*: red, +; *Iris virginca*: green, □.

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 6 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) 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 these serve as sufficient visual summaries of the important statistical properties (first and second moments) by removing the data points from the plots in the version at the right.

3.1 Robust data ellipsoids

We recognize that a normal-theory summary (first and second moments), shown visually or numerically, can be distorted by multivariate outliers, particularly in smaller samples. Such effects can be countered by using robust covariance estimates such as multivariate trimming (Gnanadesikan and Kettenring, 1972) or the

high-breakdown bound Minimum Volume Ellipsoid (MVE) and Minimum Covariance Determinant (MCD) methods developed by Rousseeuw and others (Rousseeuw and Leroy, 1987, Rousseeuw and Van Driessen, 1999). In what follows, it should be noted that robust covariance estimates could, in principle, be substituted for the classical, normal-theory estimates in all cases. To save space, we don't explore these possibilities further here.

4 Linear models: data ellipses and confidence ellipses

Here we consider how ellipses help to visualize relations 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 7. These properties are also useful in more complex contexts.

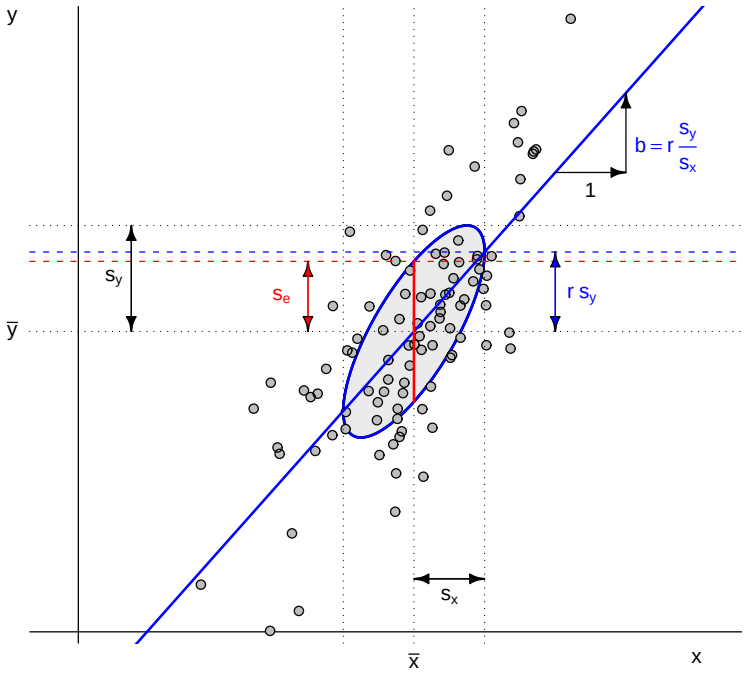

Figure 7: Annotated standard data ellipse showing standard deviations of x and y, residual standard deviation (s_e), slope (b), and correlation (r).

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

- Since 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 may be seen 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 $H_0 : \beta = 0$ may be seen in Figure 8. From the formula for a 95% confidence interval, $CI_{.95}(\beta) = b \pm t^{0.975}_{n-2} \times SE(b)$, we can take $t^{0.975}_{n-2} \approx 2$ and $SE(b) \approx \left(\frac{1}{\sqrt{n}}\right) \frac{s_e}{s_x}$, leading to

$$CI_{.95}(\beta) \approx b \pm \frac{2}{\sqrt{n}} \times \frac{s_e}{s_x} \tag{9}$$

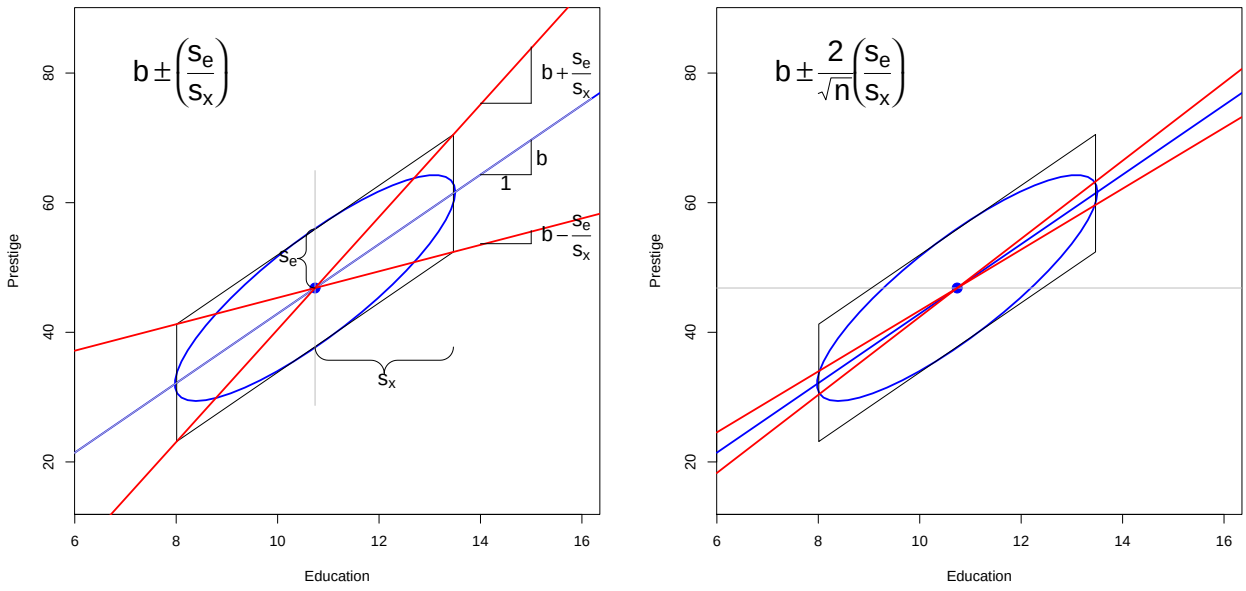

Figure 8: Visual 95% confidence interval for the slope in linear regression. Left: Standard data ellipse surrounded by the regression parallelogram. Right: Shrinking the diagonal lines by a factor of $2/\sqrt{n}$, giving the approximate 95% confidence interval for β.

To show this visually, the left panel of Figure 8 shows 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 3. Simple algebra shows 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, 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 8. In the data used for this example, $n = 102$, so the factor is approximately 0.2 here. Now consider the horizontal line through the center of the data ellipse. If this line is outside the envelope of the confidence lines, we can reject $H_0 : \beta = 0$.

4.3 Simpson's paradox, marginal and conditional relations

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 relations among variables, ignoring some important factor or covariate, and the conditional relations, adjusting (controlling) for that factor or covariate.

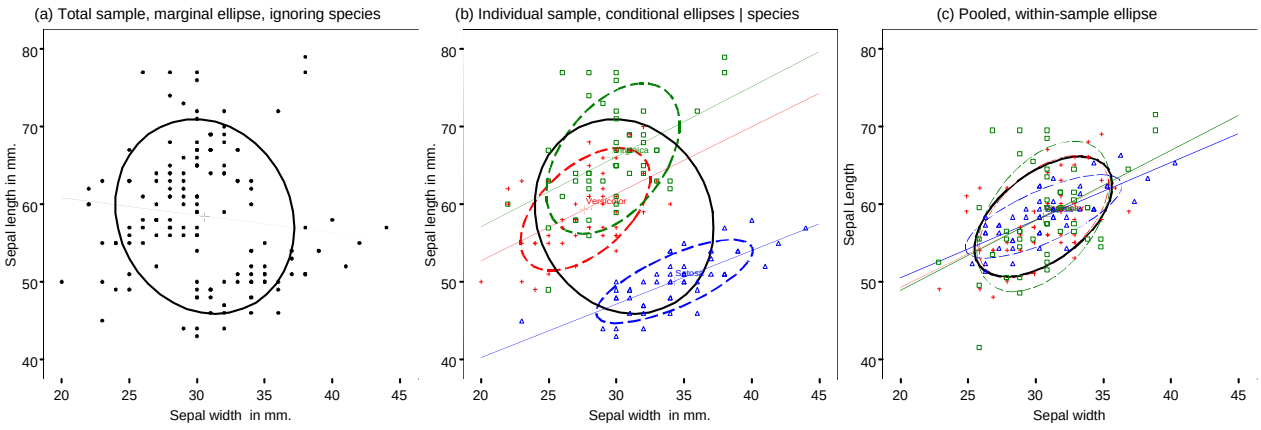

Figure 9: Marginal (a), conditional (b), and pooled within-sample (c) relations of Sepal length and Sepal width in the iris data. Total sample data ellipses are shown as black, solid curves; individual group data and ellipses are shown with colors and dashed lines

Simpson's paradox (Simpson, 1951) occurs when the marginal and conditional relations differ in direction. This may be seen in the plots of Sepal length and Sepal width shown in Figure 9. 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 relation, accounting for the negative marginal correlation.

A correct analysis of the (conditional) relation between these variables, controlling, or adjusting for mean differences among species would be 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 9. In this figure, the data for *each* species were first transformed to deviations from the 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 GLM, fitting, 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, 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 relation between variables y and x of interest, but 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 (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 pooled, marginal association between y and x, giving a slope β_{pooled} and correlation r_{pooled}. 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 is 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 10 illustrates these estimates in a simulation of five groups, with $n = 10$, means $\bar{x}_i = 2\,i + \mathcal{U}(-0.4, 0.4)$ and $\bar{y}_i = \bar{x}_i + \mathcal{N}(0, 0.5)$, so that $r_{\text{between}} \approx 0.95$. For simplicity, we have set the within-group covariance matrices to be identical in all groups, with $\mathsf{Var}(x) = 6$, $\mathsf{Var}(y) = 2$ and $\mathsf{Cov}(x, y) = \pm 3$ 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. However, in general, it can be shown that

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

which is also evident in the right panel, where the within group slope is negative. This follows from the fact that the data ellipse for the pooled total sample has a shape which is a convex combination (weighted average) of the average within group covariance of (x, y), shown by the green ellipse in Figure 10 and the covariance of the means $(\bar{x}_i, \bar{y}_i)$, shown by the red between-group ellipse.

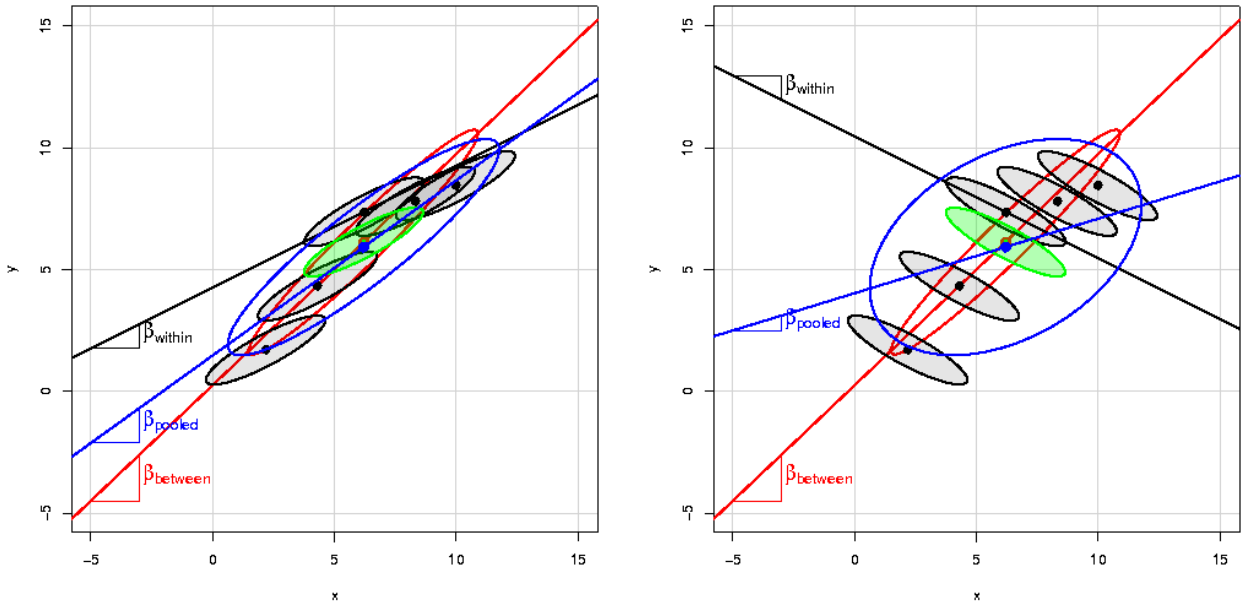

Figure 10: Paradoxes and fallacies: between (ecological), within (conditional) and pooled (marginal) associatons. In both panels, five groups have the same group means, $\mathsf{Var}(x) = 6$, and $\mathsf{Var}(y) = 2$ within each group. In the left panel, the within-group correlation is $r = +0.87$ in all groups and is $r = -0.87$ in the right panel. The green ellipse shows the average within group data ellipse.

The right panel of Figure 10 provides a prototypical illustration of Simpson's paradox, where β_{within} and β_{pooled} 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 relation (β_{pooled}) and believing that it reflects the conditional relation, 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 relations in the separate panels reflect the pairwise conditional relations with other variables controlled. In a regression context, the antidote to the marginal fallacy is the added-variable plot, which displays the conditional relation directly, controlling for all other predictors.

The right panel of Figure 10 also illustrates Robinson's paradox (Robinson, 1950), where β_{within} and β_{between} can have opposite signs.[4] The more general *ecological fallacy* (Freedman, 2001) is to draw conclusions from aggregated data, estimating β_{between} or r_{between}, believing that they reflect relations 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 relations of these moral variables as if they reflected individual behavior.[5] As

[4]William Robinson (1950) examined the relation 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.

[5]Guerry was certainly aware of the logical problem of ecological inference, at least in general terms (Friendly, 2007), and carried

can be seen in Figure 10, 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 10), or when the relevant data are available only at the individual level, so that only the conditional relation, β_{within} can be estimated. The *atomistic fallacy* (sometimes called the *fallacy of composition*) is the inverse to the ecololgical fallacy, and consists of believing that you can draw conclusions about the marginal relation, β_{between} from the conditional one.

The atomistic fallacy occurs most often in the context of multilevel models, where it is desired to draw inferences regarding variablility of higher-level units (states, countries) from data collected from lower-level units. For example, imagine that the right panel of Figure 10 depicts the decreasing relation of mortality from heart disease (y) with individual income (y) 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 portions), leading to increased heart disease.

4.5 Leverage, influence *and* precision

The topic of leverage and influence in regression is often introduced with graphs similar to to Figure 11, what we call the "leverage-influence quartet." In these graphs, a bivariate sample of $n = 20$ points was first generated with $x \sim \mathcal{N}(40, 10)$ and $y \sim 10 + 0.75x + \mathcal{N}(0, 2.5)$. Then, in each of panels (b)–(d) a single point is added at the locations shown, to represent, respectively, a low-leverage point with large residual, 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 ($y - \hat{y}$) combine to produce influential points— those that affect the estimates of $\hat{\boldsymbol{\beta}} = (\beta_0, \beta_1)^{\mathsf{T}}$.

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 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.

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 $\hat{\boldsymbol{\beta}}$. 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 11 in a single graph shown in Figure 12 to allow direct comparison. Since 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 11) i 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 $\hat{\boldsymbol{\beta}}$.

---

out several side-analyses to examine potential confounding variables.

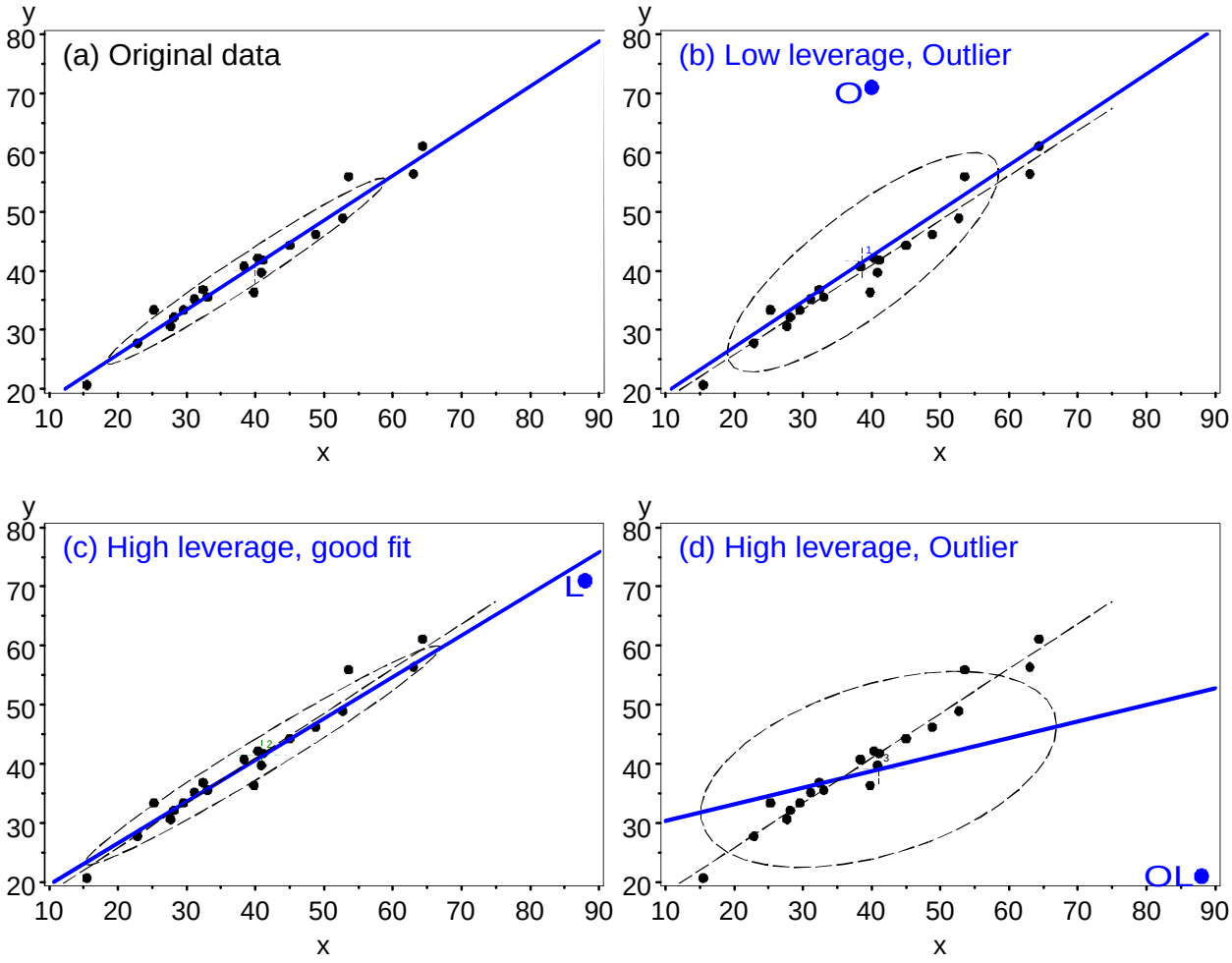

Figure 11: Leverage-Influence quartet with data ellipses. (a) Original data; (b) Adding one low leverage outlier; (c) Adding one "good" leverage point; (d) Adding one "bad" leverage point. In panels (b)–(d) the dashed black line is the fitted line for the original data, while the thick solid blue line reflects the regression including the additional point. The data ellipses show the effect of the additional point on precision.

4.6 Ellipsoids in data space and β 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 relations in data space. For linear models, particularly regression models with quantitative predictors, there is another space—β space, that provides deeper views of models and relations among them. In β 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 β space are *dual* to each other. In simple linear regression, for example, each line in data space corresponds to a point in β space, the set of points on any line in β space

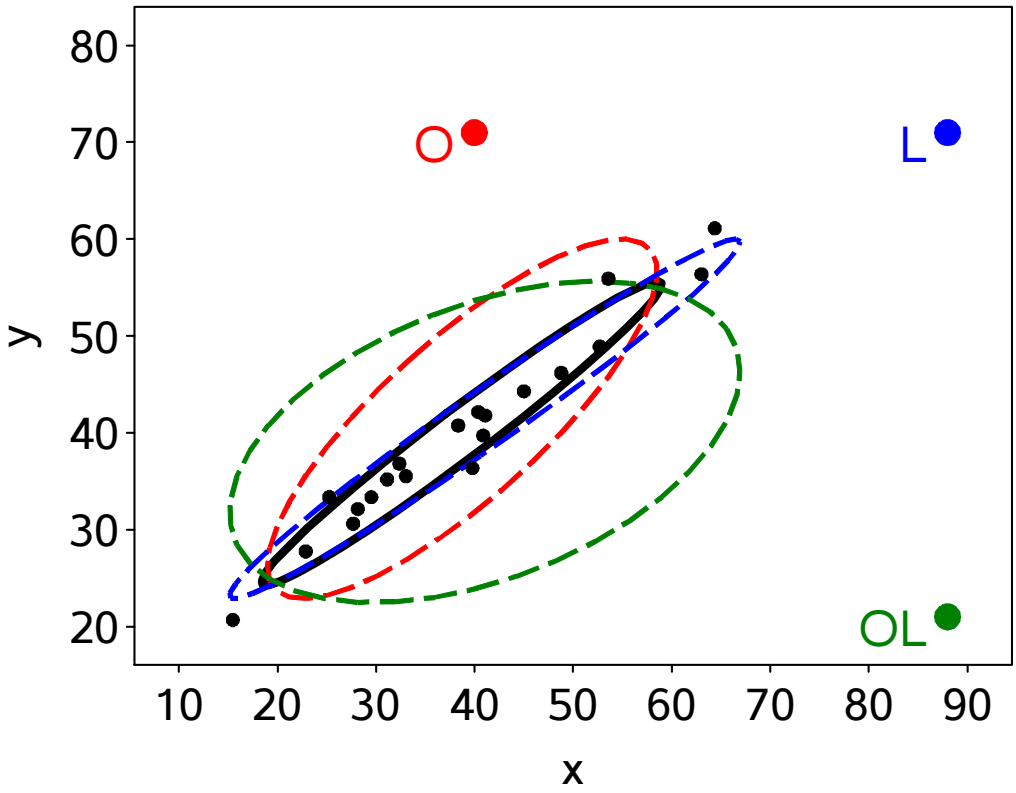

Figure 12: Data ellipses in the Leverage-Influence quartet. This graph overlays the data ellipses and additional points from the four panels of Figure 11. It can be seen that only the OL point affects the slope, while the O and L points affect precision of the estimates in opposite directions.

corresponds to a pencil of lines through a given point in data space, and the proposition that every pair of points define 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 $\bar{y} \oplus c\sqrt{\boldsymbol{S}}$. In the dual of $\boldsymbol{\beta}$ space, joint confidence regions for the parameters are given by ellipsoids of the form $\hat{\beta} \oplus c\sqrt{\boldsymbol{S}^{-1}}$. We illustrate these relations in the example below.

Figure 13 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 13 shows only the marginal relations 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?

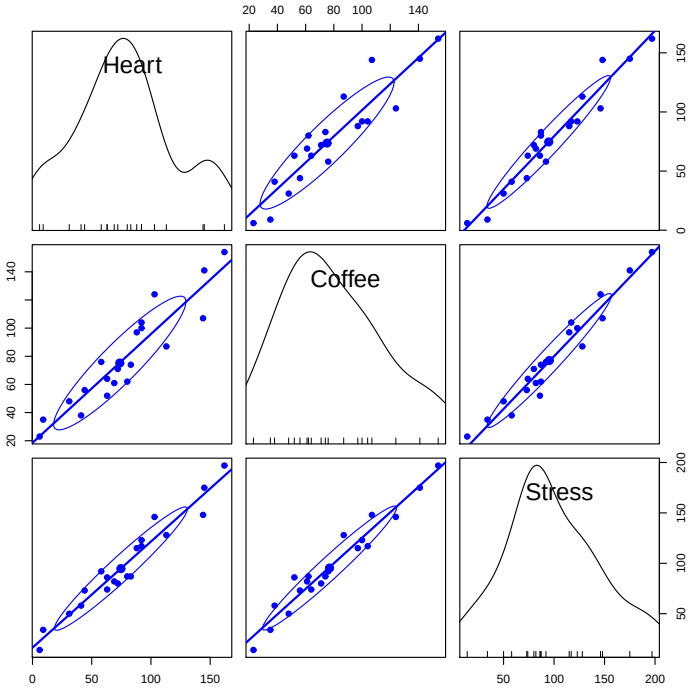

Figure 13: Scatterplot matrix, showing the relations between Heart (y), Coffee (x_1) and Stress (x_2), with linear regression lines and 68% data ellipses for the marginal bivariate relations.

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -7.7943     5.7927  -1.346    0.196
Coffee       -0.4091     0.2918  -1.402    0.179
Stress        1.1993     0.2244   5.345 5.36e-05 ***
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

Residual standard error: 10.36 on 17 degrees of freedom
Multiple R-squared: 0.9462,     Adjusted R-squared: 0.9399
F-statistic: 149.6 on 2 and 17 DF,  p-value: 1.620e-11
```

Figure 14 shows the relation 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 13, but with a standard (40%) data ellipse. The right panel shows 95% confidence regions and intervals in $\boldsymbol{\beta}$ space, determined as

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

where d is the number of dimensions for which we want coverage, ν is the residual degrees of freedom for

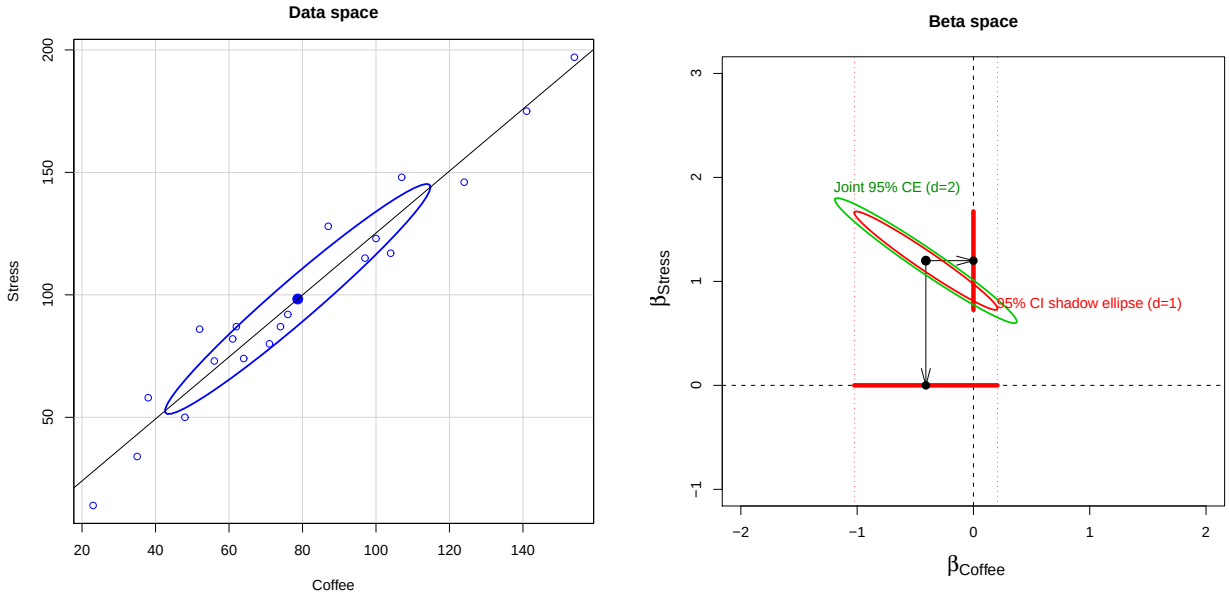

Figure 14: Data space and $\boldsymbol{\beta}$ space representations of Coffee and Stress. Left: Standard (40%) data ellipse Right: Joint 95% confidence ellipse (green) for $(\beta_{\text{Coffee}}, \beta_{\text{Stress}})$, CI ellipse (red) with 95% univariate shadows.

s_e and $\boldsymbol{S}_X$ is the covariance matrix of the predictors.

Thus, the green ellipse in Figure 14 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 ellipse.
- The shadows of this ellipse on the horizontal and vertical axes give Scheffé joint 95% confidence intervals for each parameter, with protection for "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 14, which is scaled by $\sqrt{F^{.95}_{1,\nu}} = t^{.975}_{\nu}$. The shadows of this 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 relation between the data ellipse and the confidence ellipse. These have exactly the same shape, but (with equal coordinate scaling of the axes), the confidence ellipse is exactly a 90^o rotation of the data ellipse. In directions in data space where the data ellipse is wide— where we have more information

about the relation 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).

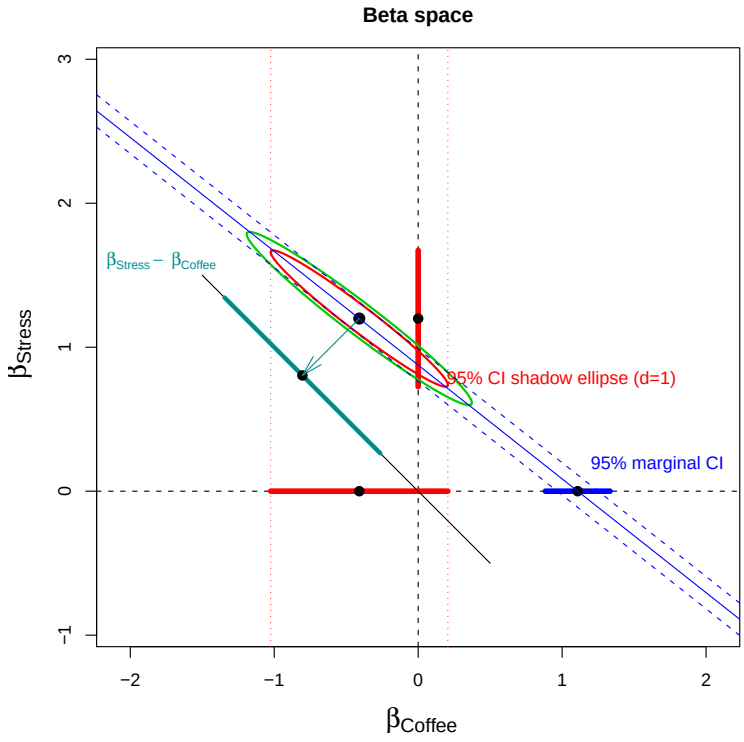

Figure 15: Joint 95% confidence ellipse for $(\beta_{\text{Coffee}}, \beta_{\text{Stress}})$, together with the 1D marginal confidence interval for β_{Coffee} ignoring Stress (thick blue line), and a visual confidence interval for $\beta_{\text{Stress}} - \beta_{\text{Coffee}} = 0$ (dark cyan).

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 13 showed the strong marginal relation. This can be seen in Figure 15 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 horozontal 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. It 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) within certain sub-spaces. Let's assume that Coffee and Stress were measured on commensurable scales, so it makes sense to ask if they have equal impacts on Heart disease, in the joint model that includes them both. Figure 15 also shows an auxilliary 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 Ellipsoids in added variable plots

In contrast to the marginal, bivariate views of the relations of several predictors to a response (e.g., such as shown in the top row of Figure 13), added-variable plots (aka partial regression residual plots) show the partial relations between the response and each predictor, where the effects of all other predictors have been controlled or adjusted for.

ToDo: Finish this section, or decide to abandon it.

5 Multivariate linear models: HE plots

Multivariate linear models (MvLM s) 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,

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

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

For *any* such hypothesis of the form 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 crossproducts (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' 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 give 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' Λ	$\Lambda = \prod^s \frac{1}{1+\lambda_i} = \prod^s (1 - \rho_i)$	geometric	$\eta^2 = 1 - \Lambda^{1/s}$
Pillai trace	$V = \sum^s \frac{\lambda_i}{1+\lambda_i} = \sum^s \rho_i$	arithmetic	$\eta^2 = \frac{V}{s}$
Hotelling-Lawley trace	$H = \sum^s \lambda_i = \sum^s \frac{\rho_i}{1-\rho_i}$	harmonic	$\eta^2 = \frac{H}{H+s}$
Roy maximum root	$R = \lambda_1 = \frac{\rho_i}{1-\rho_i}$	supremum	$\eta^2 = \frac{\lambda_1}{1+\lambda_1} = \rho_1$

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 in 3D) which express the size of co-variation 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 16 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 6. Panel (b) shows the HE plot for these variables from the one-way MANOVA model $\boldsymbol{y}_{ij} = \boldsymbol{\mu}_i + \boldsymbol{e}_{ij}$ 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 $\boldsymbol{H} = \widehat{\boldsymbol{Y}}^\mathsf{T}\widehat{\boldsymbol{Y}}$ 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 9(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 ellipses are then exactly the data ellipses of the fitted values, and correspond visually to multivariate analogs of univariate effect size measures (e.g., $(\bar{y_1} - \bar{y_2})/s_e$ where s_e=within group standard deviation).

For significance scaling, it turns out to be most visually convenient to use Roy's largest root test 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

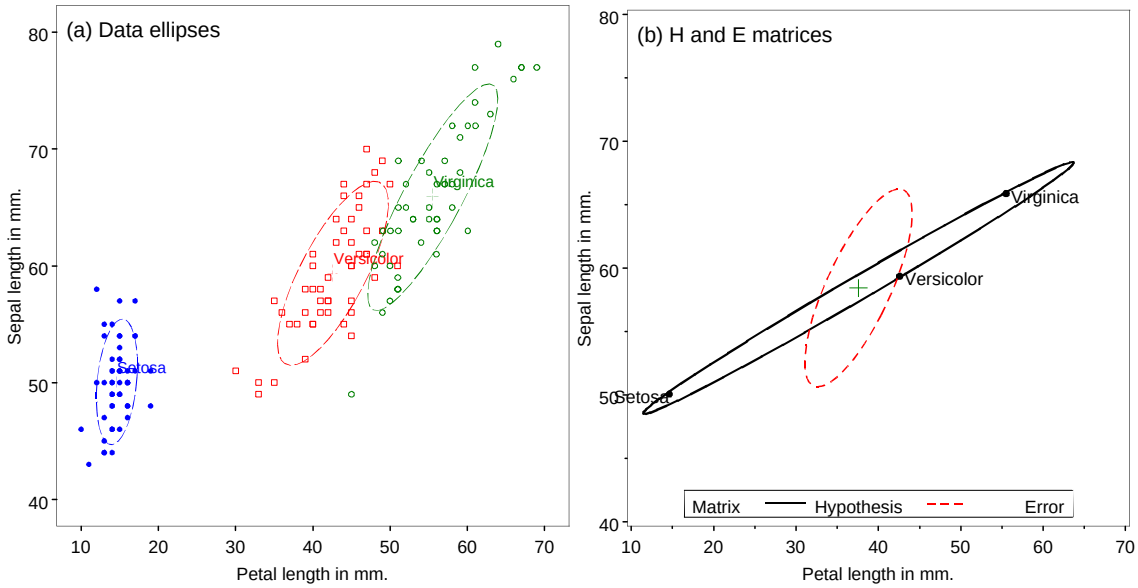

Figure 16: (a) Data ellipses and (b) corresponding HE plot for sepal length and petal length in the Iris dataset. The $\boldsymbol{H}$ ellipse is the data ellipse of the fitted values defined by the group means, $\bar{\boldsymbol{y}}_{i\cdot}$. The $\boldsymbol{E}$ ellipse is the data ellipse of the 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.

statistic.[6] 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.[7] 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 16(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 test 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 sum to zero.

As an important special case, for an overall effect with df_h degrees of freedom (and balanced sample sizes), a set of df_h pairwise orthogonal $(1 \times q)$ $\boldsymbol{L}$ matrices ($\boldsymbol{L}_i^{\mathsf{T}} \boldsymbol{L}_j = 0$) gives rise to a set of df_h rank 1 $\boldsymbol{H}_i$

[6] 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 this gives the critical value for an α-level test: $\lambda_\alpha = (df_1/df_2) F^{1-\alpha}_{df_1, df_2}$

[7] Other multivariate tests (Wilks' Λ, Hotelling-Lawley trace, Pillai trace) also have geometric interpretations in HE plots (e.g., Wilks' 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.

matrices that additively decompose the overall hypothesis SSCP matrix,

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

exactly as the univariate SS_H may be decomposed in an ANOVA. Each of these rank 1 $\boldsymbol{H}_i$ 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. The subhypotheses will then have hypothesis ellipses (of dimension rank($\boldsymbol{H}_i$)) that are conjugate with respect to the hypothesis ellipse for the joint hypothesis, provided the estimators for the subhypotheses are statistically independent.

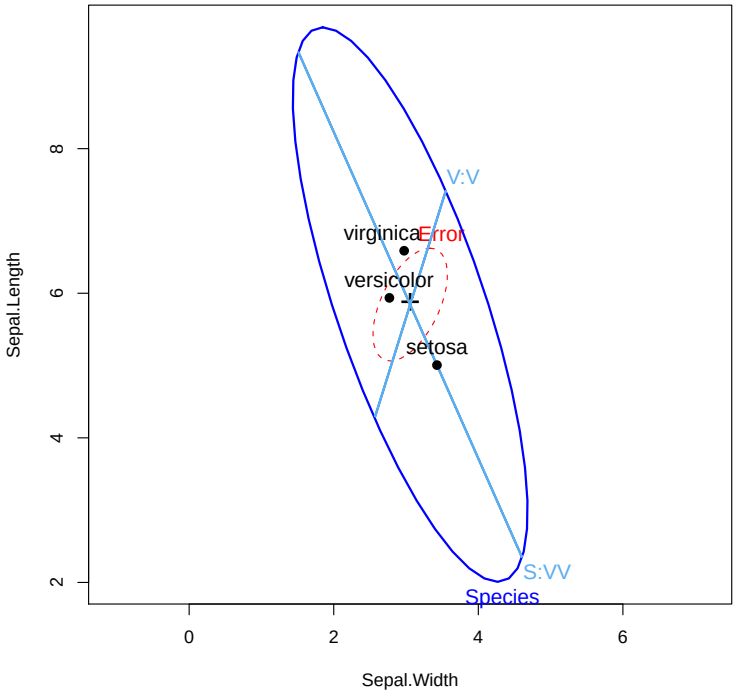

Figure 17: $\boldsymbol{H}$ and $\boldsymbol{E}$ matrices for sepal width and sepal length in the iris data, together with $\boldsymbol{H}$ matrices for testing two orthogonal contrasts in the species effect.

To illustrate, we show in Figure 17 an HE plot for the sepal width and sepal length variables in the iris data, corresponding to panel (1:2) in Figure 6. Overlayed on this plot are the 1 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 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 relation between covariation leading toward rejection of a hypothesis relative to error covariation for two variables in data space. To visualize these relations 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.

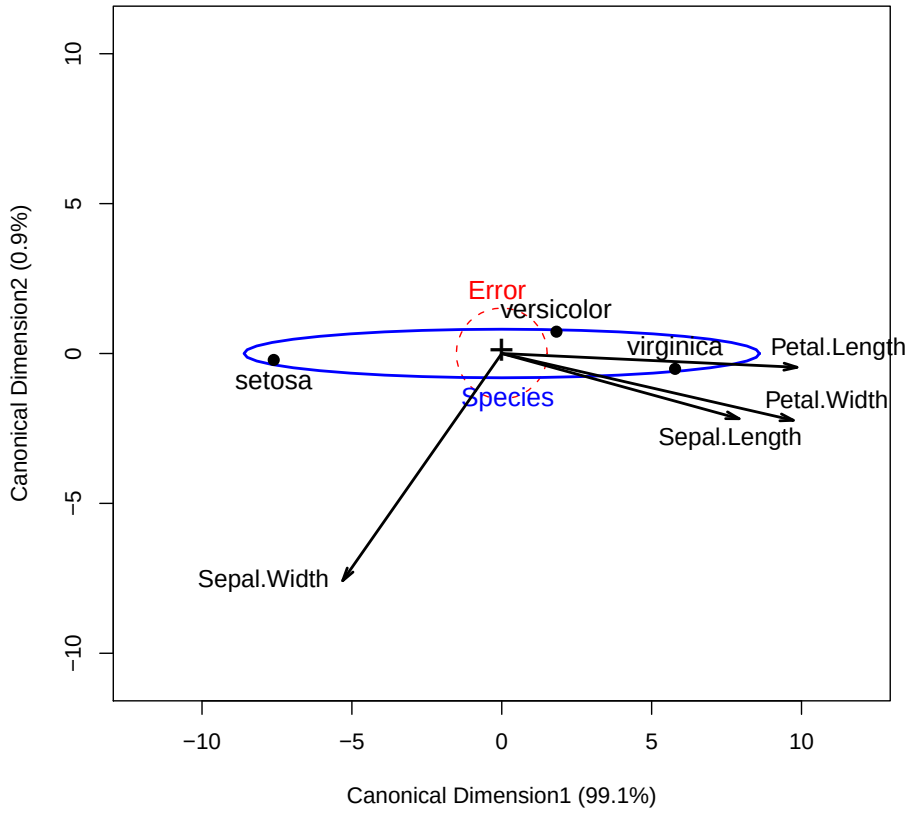

Figure 18: Canonical HE plot for the Iris data. In this plot, the $\boldsymbol{H}$ ellipse is shown using effect-size scaling to preserve resolution, and the variable vectors have been multiplied by a constant to approximately fill the plot space. The projections of the variable vectors on the coordinate 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 $\boldsymbol{y}_{ij}$, 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, $\lambda_i, i = 1, \ldots, s$, and 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 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 18 shows the canonical HE plot for the Iris data, the view in canonical space corresponding to Figure 17 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 17 in data space, showing the 4-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 2-dimensional, while $\boldsymbol{E}$ is 4D. Applying the transformation given by $\boldsymbol{E}^{-1}$ as in Figure 4 and projecting into the non-zero dimensions would give a view equivalent to the canonical HE plot in Figure 18. 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 are two principles and procedures for deriving estimates of a parameter vector $\boldsymbol{\beta}$ of a linear model, each with its associated estimated variance-covariance matrix, e.g., $\widehat{\boldsymbol{\beta}}^A$ with covariance matrix $\widehat{\mathsf{Var}}(\boldsymbol{\beta}^A)$ and $\widehat{\boldsymbol{\beta}}^B$ with covariance matrix $\widehat{\mathsf{Var}}(\boldsymbol{\beta}^B)$. We will take method A to be OLS estimation and consider several alternatives for method B. In β space, the parameter estimates appear as points and their corresponding confidence ellipsoids have the property that they will just "kiss" (or *osculate*) along a path between the two estimates. In the examples we consider, the alternative methods B represent a convex combination of information from two sources and the path of osculation is interpretable in terms of what method B aims to achieve. The same geometric ideas can also be applied in data

space, where we can consider the data ellipsoids for two (or more) groups and find statistical interpretations of their (pairwise) path of osculation.

These problems 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 19, 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.

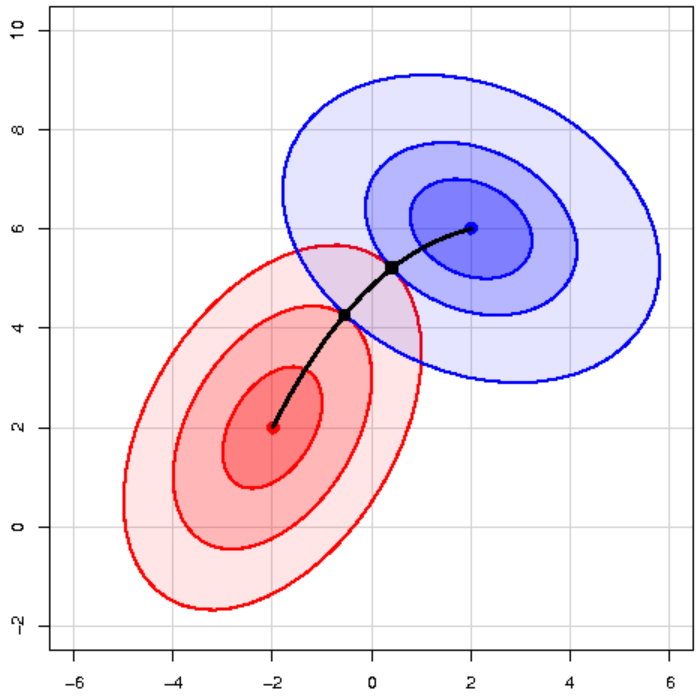

Figure 19: Locus of osculation for two ellipsoidal level curves, with centers at $\boldsymbol{m}_1 = (-2, 2)$ and $\boldsymbol{m}_2 = (2, 6)$ and shape matrices $\boldsymbol{A}_1$ and $\boldsymbol{A}_2$ given in Eqn. (15). The left ellipsoids (red) have radii=1, 2, 3. The right ellipsoids have radii=1, 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.

A 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 $\nabla f(x_1, x_2) = \left(\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}\right)$, 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, $\boldsymbol{u} \otimes \boldsymbol{v} = u_1 v_2 - u_2 v_1 = \boldsymbol{v}^\mathsf{T}\boldsymbol{C}\boldsymbol{u} = 0$, 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 $\mathcal{O} = \{\boldsymbol{x} \in \mathbb{R}^2 \,|\, \nabla f_1(\boldsymbol{x}) \otimes \nabla f_2(\boldsymbol{x}) = 0\}$, which implies

$$(\boldsymbol{x}-\boldsymbol{m}_2)^\mathsf{T}\, \boldsymbol{A}_2^\mathsf{T}\, \boldsymbol{C}\, \boldsymbol{A}_1\, (\boldsymbol{x}-\boldsymbol{m}_1) = 0 \ . \tag{16}$$

Eqn. (16) 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. (16) is necessarily zero, so the locus of osculation always passes through $\boldsymbol{m}_1$ and $\boldsymbol{m}_2$.

A visual demonstration of theory above is shown in Figure 20 (left), which overlays the ellipses in Figure 19 with contour lines (hyperbolae, here) of the vector cross product function contained in Eqn. (16). 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 20, Eqn. (16) 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.

6.1 Discriminant analysis

The right panel of Figure 20, considered in data space, provides a visual interpretation of the classical, normal theory two-group discriminant analysis problem. 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 osculation between the two ellipsoids. However, the goal in discriminant analysis 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 miss-classification.

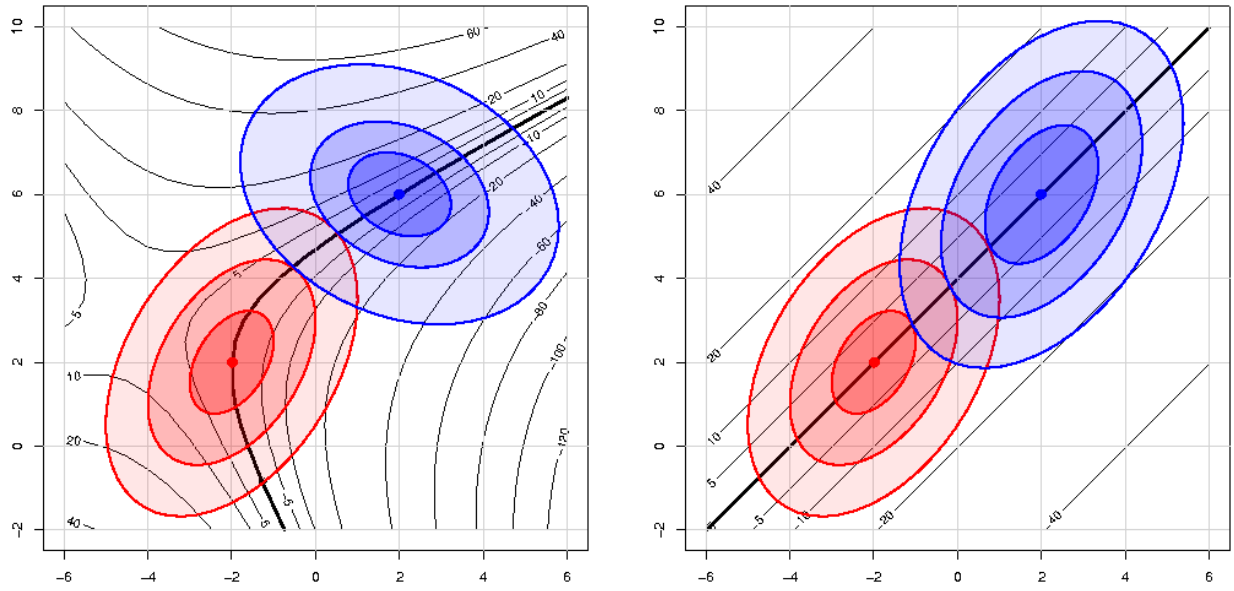

Figure 20: Locus of osculation for two ellipsoidal level curves, showing contour lines of the vector cross product function Eqn. (16). The thick black curve shows the complete locus of osculation for these two ellipses, where the cross product function equals 0. Left: with parameters as in Figure 19 and Eqn. (15). Right: with the same shape matrix, $\boldsymbol{A}_1$ for both ellipsoids.

6.2 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 approach is ridge regression which shrinks the estimates toward $\mathbf{0}$ (introducing bias) in an effort to reduce sampling variance.

Suppose the predictors have been centered at their means and the unit vector is omitted from $\boldsymbol{X}$. Then, the OLS estimates are given by

$$\widehat{\boldsymbol{\beta}}^{OLS} = (\boldsymbol{X}^\mathsf{T}\boldsymbol{X})^{-1}\boldsymbol{X}^\mathsf{T}\boldsymbol{y} \tag{17}$$

and $\beta_0 = \bar{y}$. Ridge regression replaces the standard residual sum of squares criterion with a penalized form,

$$RSS(k) = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^\mathsf{T}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) + k\boldsymbol{\beta}^\mathsf{T}\boldsymbol{\beta} \qquad (k \geq 0) \ , \tag{18}$$

whose solution is easily seen to be

$$\begin{aligned} \widehat{\boldsymbol{\beta}}_k^{RR} &= (\boldsymbol{X}^\mathsf{T}\boldsymbol{X} + k\boldsymbol{I})^{-1}\boldsymbol{X}^\mathsf{T}\boldsymbol{y} \\ &= \boldsymbol{G}\,\widehat{\boldsymbol{\beta}}^{OLS} \ . \end{aligned} \tag{19}$$

where $\boldsymbol{G} = \left[\boldsymbol{I} + k(\boldsymbol{X}^\mathsf{T}\boldsymbol{X})^{-1}\right]^{-1}$. Thus, as k increases, $\boldsymbol{G}$ decreases, driving $\widehat{\boldsymbol{\beta}}_k^{RR}$ toward $\mathbf{0}$ (Hoerl and Kennard, 1970b,a). 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 $\det(\boldsymbol{X}^\mathsf{T}\boldsymbol{X}) \approx 0$.

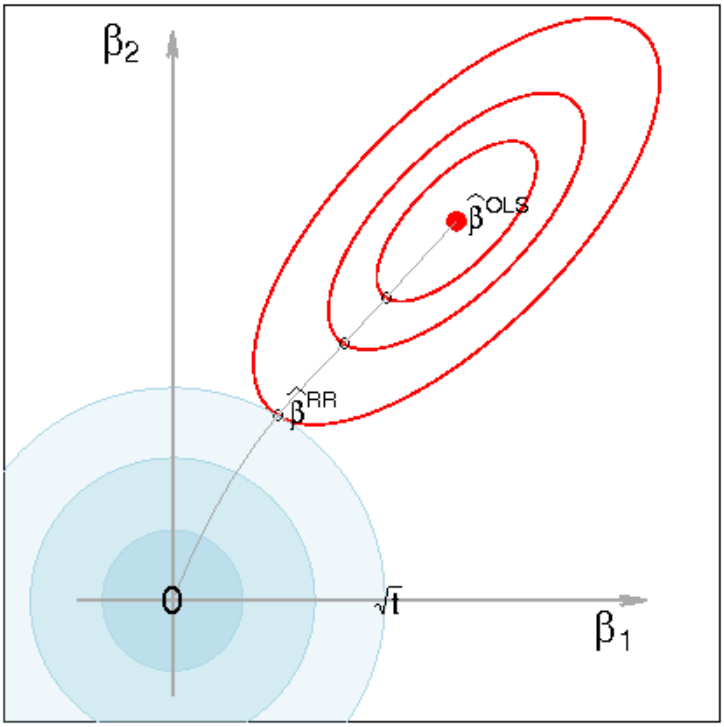

Figure 21: Elliptical contours of the OLS residual sum of squares for two parameters in a regression, together with circular contours for the constraint function, $\beta_1^2 + \beta_2^2 \leq t$. Ridge regression finds the point $\boldsymbol{\beta}^{RR}$ where the OLS contours just kiss the contstraint region.

The penalized lagrangian formulation in Eqn. (18) has an equivalent form as a constrained minimization problem,

$$\widehat{\boldsymbol{\beta}}^{RR} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^{\mathsf{T}} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) \qquad \text{subject to} \qquad \boldsymbol{\beta}^{\mathsf{T}}\boldsymbol{\beta} \leq t(k) \ , \tag{20}$$

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 21. Depicted in the figure are the elliptical contours of the OLS regression sum of squares, $RSS(0)$ around $\widehat{\boldsymbol{\beta}}^{OLS}$. 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,

$$\begin{pmatrix} \boldsymbol{y} \\ \mathbf{0} \end{pmatrix} = \begin{pmatrix} \boldsymbol{X} \\ \boldsymbol{X}_k^0 \end{pmatrix} \boldsymbol{\beta}^{RR} \ , \tag{21}$$

which gives rise to the solution,

$$\widehat{\boldsymbol{\beta}}^{RR} = [\boldsymbol{X}^\mathsf{T}\boldsymbol{X} + (\boldsymbol{X}_k^0)^\mathsf{T}\boldsymbol{X}_k^0]^{-1}\boldsymbol{X}^\mathsf{T}\boldsymbol{y} \ . \tag{22}$$

But since $\boldsymbol{X}_k^0$ 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. (22) equivalent to Eqn. (19). As promised, the ridge regression estimator then reflects a weighted average of the data $[\boldsymbol{X}, \boldsymbol{y}]$ with $n(k)$ observations $[\boldsymbol{X}_k^0, \boldsymbol{0}]$ biased toward $\boldsymbol{\beta} = 0$. In Figure 21, 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. (18), consider 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 21 in which the constraint contours are ellipses aligned with the coordinate axes, with axis lengths inversely proportional to k_i. This allows for differential shrinkage of the OLS coefficients. The visual solution to the obvious modification of Eqn. (20) is again the point where the elliptical contours of $RSS(0)$ kiss the contours of the (now elliptical) contraint region.

6.2.1 Bivariate ridge trace plots

Ridge regression is touted 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, 1970a), 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 on the effects on *both* bias and variance.

In the standard linear model, confidence ellipsoids are generated from the estimated covariance matrix of the of the parameters,

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

In the ridge regression model, this becomes (Marquardt, 1970)

$$\widehat{\mathsf{Var}}(\boldsymbol{\beta}^{RR}) = \hat{\sigma}_e^2[\boldsymbol{X}^\mathsf{T}\boldsymbol{X} + k\boldsymbol{I}]^{-1}(\boldsymbol{X}^\mathsf{T}\boldsymbol{X})[\boldsymbol{X}^\mathsf{T}\boldsymbol{X} + k\boldsymbol{I}]^{-1} \ , \tag{23}$$

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

Figure 22 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, GNP.deflator (using 1954 as 100).[8] The standard linear model for these data, in R notation is `lm(Employed ~ ., data=longley)`. For each value of k, the plot shows the estimate $\widehat{\boldsymbol{\beta}}$, together with the confidence 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.

[8] 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, $\boldsymbol{X}^\mathsf{T}\boldsymbol{X}$. This 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 expected 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.

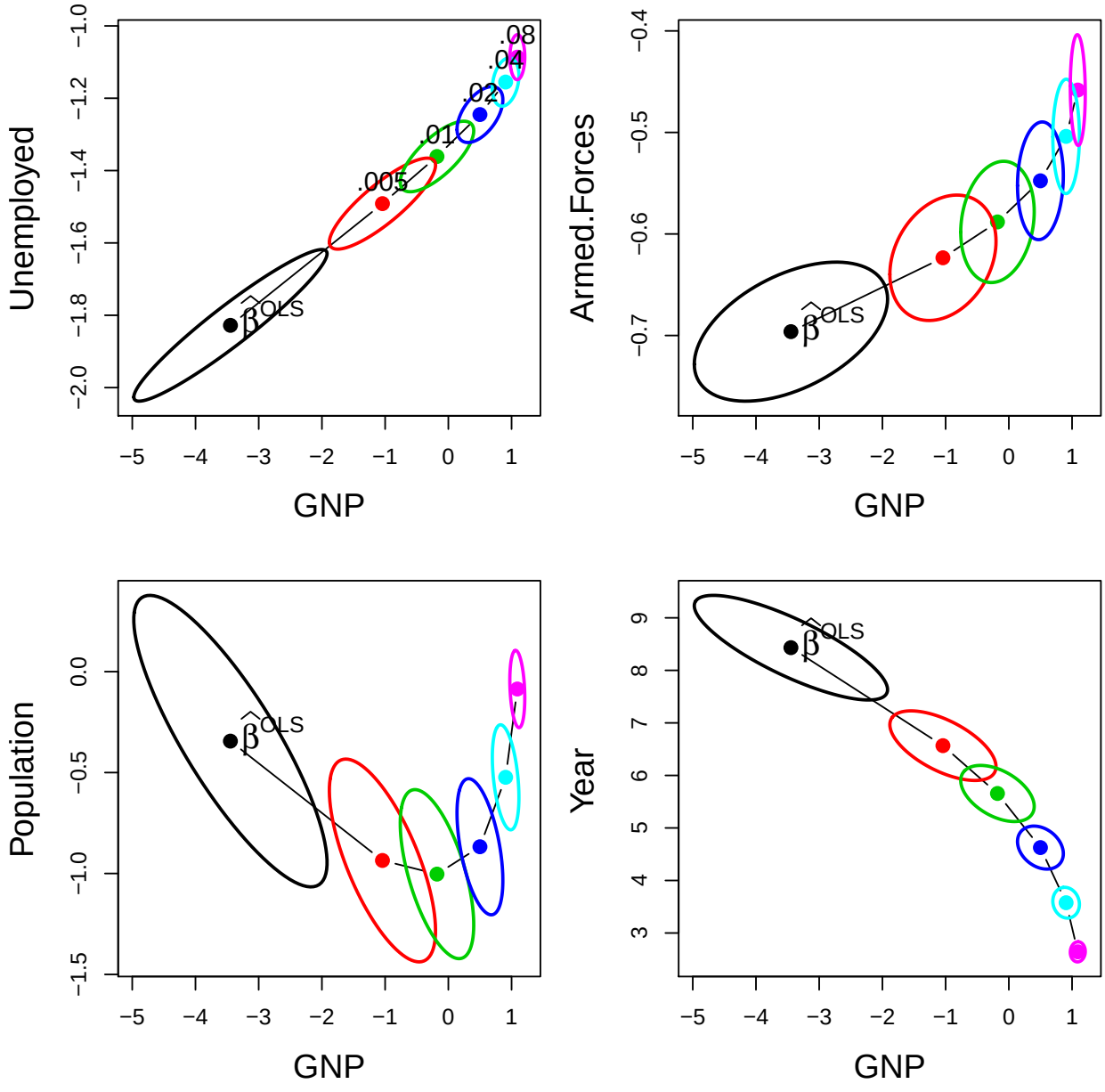

Figure 22: Bivariate ridge trace plots for the coefficients of four predictors against the coefficient for GNP in Longley's data, with $k = 0, 0.005, 0.01, 0.02, 0.04, 0.08$. In most cases the coefficients are driven toward zero, but the bivariate plot also makes clear the reduction in variance. To reduce overlap, all confidence ellipses are shown with 1/2 the standard radius.

For this 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 confidence

ellipse. Moreover, shrinkage in variance is generally in a similar direction to the shrinkage in the coefficients.

6.3 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, $\boldsymbol{\beta}^{prior}$ and covariance matrix $\boldsymbol{A}$. We show that under reasonable conditions the Bayesian posterior estimate, $\widehat{\boldsymbol{\beta}}^{posterior}$, turns out to be a weighted average of the prior coefficients $\boldsymbol{\beta}^{prior}$ and the OLS solution $\widehat{\boldsymbol{\beta}}^{OLS}$, 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 ellipses 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 $\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}^{OLS}$ to give

$$(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^{\mathsf{T}}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) = (\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})^{\mathsf{T}}(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}) + (\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}})^{\mathsf{T}}(\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X})(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}) \tag{24}$$

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. (24) and $\Pr(\boldsymbol{\beta} \,|\, \sigma^2)$ a normal distribution,

$$\Pr(\boldsymbol{\beta} \,|\, \sigma^2) \propto (\sigma^2)^{-p} \times \exp\left(-\frac{1}{2\sigma^2}(\boldsymbol{\beta} - \boldsymbol{\beta}^{prior})^{\mathsf{T}} \; \boldsymbol{A}(\boldsymbol{\beta} - \boldsymbol{\beta}^{prior})\right) \tag{25}$$

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

$$\widehat{\boldsymbol{\beta}}^{posterior} = (\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X}\widehat{\boldsymbol{\beta}}^{OLS} + \boldsymbol{A}\boldsymbol{\beta}^{prior})(\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X} + \boldsymbol{A})^{-1} \tag{26}$$

with covariance matrix $(\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X} + \boldsymbol{A})^{-1}$. The posterior coefficients are thus 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 $\widehat{\boldsymbol{\beta}}^{prior} = \mathbf{0}$ and $\boldsymbol{A} = k\boldsymbol{I}$ and where Figure 21 provides an elliptical visualization. In Eqn. (22), 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.

> **ToDo**: Complete this section. What have I omitted in the development, necessary for this paper? Are there other geometric insights? Does it need another, Bayesian specific graph? Is there a simple expression for $\mathsf{Var}(\widehat{\boldsymbol{\beta}}^{posterior})$ that could be exploited?

6.4 Mixed models: BLUEs and BLUPs

ToDo: This section, and the incomplete example that follows is just an initial attempt, awaiting a more insightful description.

In this section we make use of the duality between data space and β space, where lines in one map into points in the other and ellipsoids to visualize the precision of estimates in the context of the general 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 OLS regression, best linear unbiased estimates (BLUEs) *within* strata and the variation of random effect estimates *between* strata.

The mixed model for hierarchical data provides a general framework for dealing with lack of independence among observations in linear models, such as occurs when students are sampled within schools, schools within counties and so forth. In these situations, the assumption of OLS that the residuals are conditionally independent is likely to be violated, because, for example, students nested within the same school are likely to have more similar outcomes than those from separate schools.

6.4.1 Example: Math achievement and SES

To illustrate, we use the classic data set from Bryk and Raudenbush (1992), Raudenbush and Bryk (2002) dealing with math achievement scores from a subsample of 7,185 students from 160 schools in a 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 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 in centered form, CSES = SES - meanSES for ease of interpretation (making the within-school intercept equal to meanSES for that school).

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. ...
>
> Sir Francis Galton, *Natural Inheritance*, London: Macmillan, 1889. ("(Elliptical)" added).

In statistical data, theory and graphical methods, one main organizing distinction can be made in *all* of these 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).

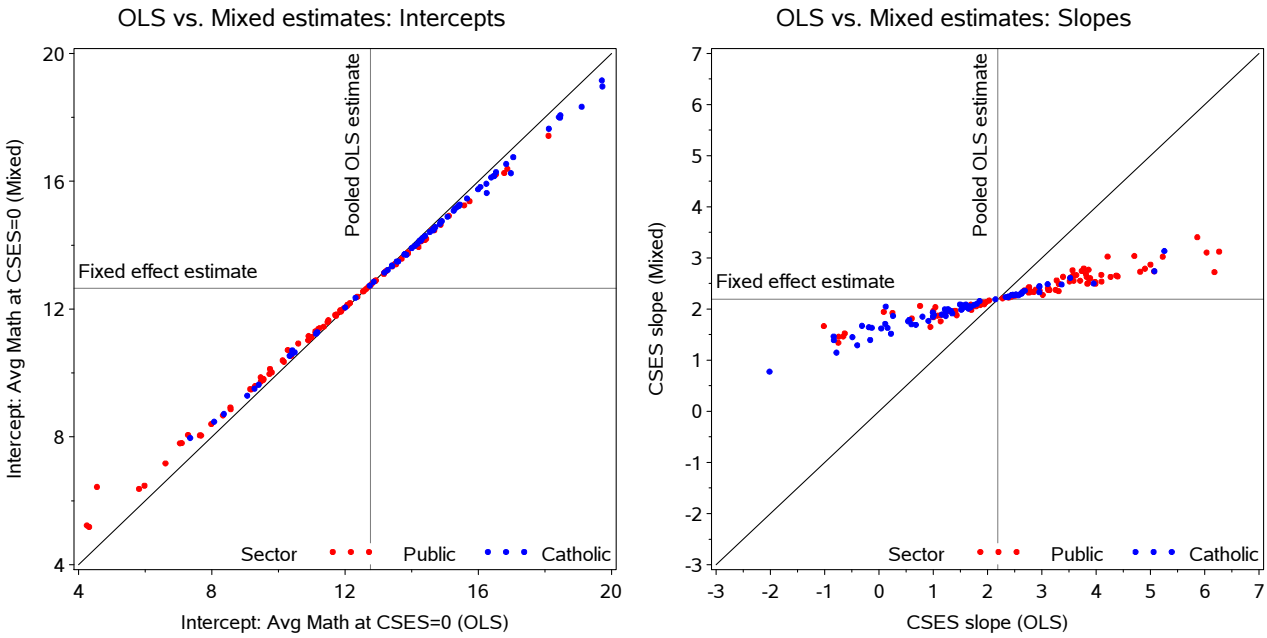

Figure 23: Comparing BLUEs and BLUPs. Each panel plots the OLS estimates from separate regressions for each school (BLUEs) versus the mixed model estimates from the random intercepts and slopes model (BLUPs). Left: intercepts; Right: slopes for CSES. The shrinkage of the BLUPs toward the OLS estimate is much greater for slopes than intercepts.

This scale[9] at least implicitly organizes much of current statistical teaching, practice, and software. But within this, the data, theory and graphical methods are often treated separately (1D, 2D, nD), without regard to geometric ideas and visualizations that help tie them together.

This paper starts from the idea that one geometric form— the ellipsoid— provides a unifying framework for many statistical phenomena, with simple representations in 1D (a line), 2D (ellipse) that extend naturally to n dimensions. The intellectual leap in statistical thinking from ONE to TWO in Galton (1886) was enormous. Galton's visual insights 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 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 relations among statistical methods, matrix algebra and geometry are not new here. To our knowledge, Dempster (1969) was the first to exploit this in a systematic fashion, establishing the connections

[9]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.

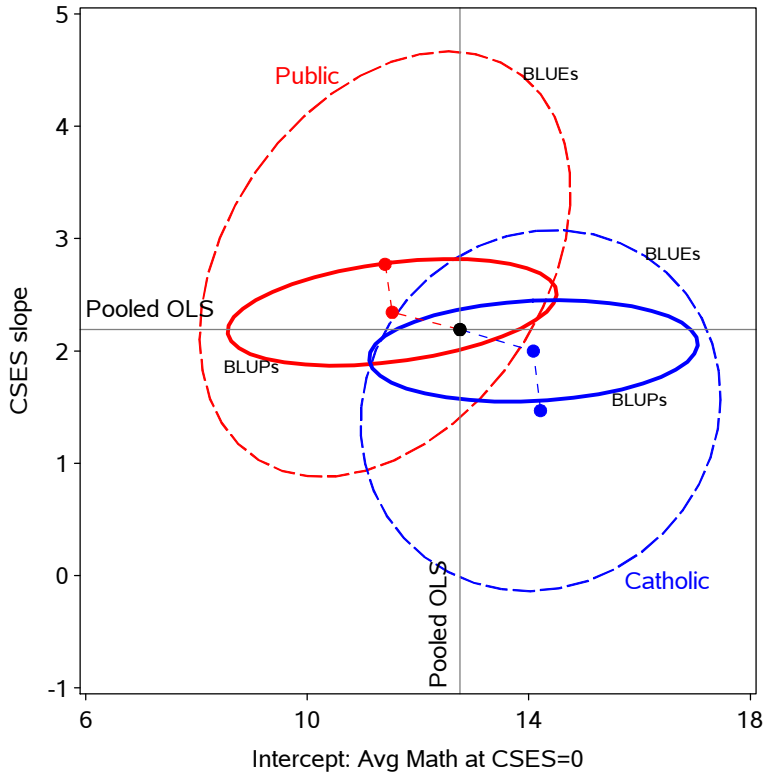

Figure 24: Comparing BLUEs and BLUPs. The plot shows ellipses of 50% coverage for the estimates of intercepts and slopes from OLS regressions (BLUEs) and the mixed model (BLUPs), separately for each sector. The centers of the ellipses illustrate how the BLUPS can be considered a weighted average of the 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.

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, leading to important visual insights.The roots of these connections go back much further— to Cramér (1946) (idea of the concentration ellipsoid), Hotelling (1933) (principal components) and, we maintain, ultimately to Galton (1886).

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 ...

Several features of the current discussion may help to present these in a new light.

... [to be completed] ...

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.

9 Acknowledgments

This work is supported by Grant OGP0138748 from the National Sciences and Engineering Research Council of Canada.

References

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

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.

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.

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

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.

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

Fox, J. (2008). *Applied Regression, Generalized Linear Models, and Related Methods*. Thousand Oaks, CA: Sage.

Freedman, D. A. (2001). Ecological inference and the ecologocal falacy. In N. J. Smelser and P. B. Baltes, eds., *Internaitonal Encyclopedia of the Social and Behavioral Sciences*, (pp. 4027–4030). Pergamon Press.

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

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

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

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

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.

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

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

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

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.

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.

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

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

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

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

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).

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.