Linear algebra and multivariate analysis in statistics: development and interconnections in the twentieth century

N. H. BINGHAM and W. J. KRZANOWSKI

In memory of Sir David Cox, 15.7.1924 - 17.1.2022

Abstract

The most obvious points of contact between linear and matrix algebra and statistics are in the area of multivariate analysis. We review the way that, as both developed during the last century, the two influenced each other. We illustrate this by examining a number of key areas.

We begin with matrix and linear algebra, its emergence in the 19th century, and its eventual penetration into the undergraduate curriculum in the 20th century. We continue with a much longer similar account for multivariate analysis in statistics. We pick out the year 1936 as that of three key developments, by H. Hotelling, R. A. Fisher and P. C. Mahalanobis, and in the early post-war period three more key developments, by M. G. Kendall, M. S. Bartlett and C. R. Rao. We then turn to some special results in linear algebra that we need: Schur complements, and related inversion formulae. We briefly discuss four of the contributors, Fisher and Kendall named above, together with S. S. Wilks and T. W. Anderson. We close with thirteen 'case studies', showing in a range of specific cases how these general algebraic methods have been put to good use and changed the face of statistics.

Keywords linear algebra, matrix algebra, multivariate analysis in statistics, multivariate normal distribution, Gaussian Regression Formula, Schur complement, singular values decomposition.

0. Time-line

Our main focus is on the 20th century, though we touch on the 18th and 19th (and of course are influenced by the first part of the 21st). Both authors were born in the 1940s and began publishing in the 1970s. We allow ourselves to use the term 'modern', imprecise and subjective as it is, as being as useful in this context as in general usage. It might be loosely read here as 'within living memory'. For example, we say (§4.3) of Anderson's 1958 book that we regard it as 'the first of the unambiguously modern textbooks we cite'. We invite readers to form their own views here.

1. Introduction

As both subject areas changed beyond recognition in the last century, we begin by outlining their development, discussing linear algebra in §2 and multivariate analysis in §3. In §4 we comment briefly on four major contributors to the area: Fisher, Kendall, Wilks and Anderson. We turn in §5 to special results in linear algebra: what is now called the *Schur complement* (the crucial ingredient for, e.g., the Gaussian Regression Formula of §6.2), and various inversion formulae for partitioned matrices (e.g. the (Sherman-Morrison-)Woodbury formula), crucial for studying the sensitivity of a statistical analysis to error in one data point. In §6 we discuss thirteen specific instances – 'case studies' – of this algebra-statistics interplay. We conclude in §7.

2. Matrix and linear algebra: development and absorption into the undergraduate curriculum

The theory of determinants certainly precedes that of matrices (the reverse of the logical order). Muir, in Volume 1 of his 5-volume history (Muir 1906), discusses seven authors in his Chapter II (1693-1779, from Leibniz to Bézout) (including Vandermonde, for more on whom see Ycart (2013)), and two more in Chapter III (1784-1812), before discussing Gauss, who introduced the term in his *Disquisitiones arithmeticae* of 1801. Our determinant notation is due to Cayley in 1841.

The term matrix was introduced in 1858 by Cayley (Hawkins 1975a), and he and Sylvester worked extensively on matrix theory, but the subject was by then well established; see Hawkins (1975b) for details and references. One source was quadratic forms, in analytic geometry and number theory. Another was (what we now call) spectral theory, for instance in the work of Cauchy of 1829 on eigenvalues. Inverse matrices date from Jacobi in 1834 (Turnbull 1928, 77). An idea of the state of the theory in the early 20th century can be gained from the books Turnbull (1928) (showing the influence of the 19th century interest in invariants) and Turnbull and Aitken (1932) (on canonical forms).

A good summary of the relevant developments 1750-1900 is given by Farebrother (1999). He observes how the problem of fitting linear relationships, ubiquitous in statistics from Legendre, Gauss and Laplace on in the method of least square (and in Boscovich's L_1 precursor of this L_2 theory:

see Eisenhart (1961)) helped to motivate the development of matrix algebra and linear programming. A study of the emergence of matrix theory as a subject in its own right, with particular reference to the inter-war years, is given by Brechenmacher (2010).

J. W. Gibbs, in his lecture notes at Yale of 1881 and 1884 (published 1901), did much to spread the use of vector methods. An early stimulus for these was Grassmann's *Ausdehnungslehre* of 1844. Vector methods were championed in the UK by O. Heaviside (to whom we owe, e.g., the use of bold face for vectors). For a good brief account of the struggles between the proponents of vectors and quaternions (W. R. Hamilton, 1843), see the Historical Introduction in (Weatherburn 1921). In a paper written while he was still a student, Fisher applied vector methods to geometry (Fisher 1913). Unfortunately, vector methods were still scorned by some as late as 1937; see e.g. the Preface of (Ramsey, 1937).

An insight into the our theme a century earlier is provided by the following passage from (Todhunter 1869) acknowledging input from Cayley (thanks to Steve Stigler for this):
'Addition. It has been pointed out to me by Professor Cayley, that the evaluation of the integral K in Art. 7 may be effected at a single step by a simultaneous transformation of the variables ....

' I have however preserved the methods given in the memoir, because they require less knowledge of the theory of determinants or of linear transformations'.

Modern algebra may be said to stem from van der Waerden's *Moderne Algebra* of 1932, though that has only one chapter on linear algebra. The standard text on general modern algebra in the UK in the mid-1960s when the authors were students was Birkhoff and MacLane (1953). Standard works on matrix and linear algebra from that time include Gantmacher (1959) and Mirsky (1955).

An insight into the undergraduate curriculum, at least in the UK, around this period is given in Mirsky's obituary (Burkill et al 1986). When Leon Mirsky (1918-1983) arrived to take up his post in Sheffield: 'However, when A G Walker, who was appointed to Daniell's former chair in 1947, asked him to give a lecture course in linear algebra (which had hardly figured in his undergraduate curriculum), he immediately became fascinated by this novel subject. The result was his textbook on linear algebra ....'

Students of the 1960s will remember the Oliver & Boyd series University Mathematical Texts published in Edinburgh: good, short, cheap (7/6 old

money, or 37.5p) and so compact they would go in a jacket pocket. Several authors wrote two of these, including A C Aitken (1895-1967), who wrote the first two: *Determinants and matrices* (Aitken 1939a) and *Statistical mathematics* (Aitken 1939b). It has been shrewdly observed that the reader would not know these books were by the same author, nor that the two areas had anything to do with one another. Their interplay is our main theme here. (For background on Aitken's very interesting life, see his obituary (Bartlett and Whittaker 1968).)

The more recent literature distinguishes between *matrix* theory, where coordinate representations are needed (see for example Horn and Johnson (1985; 1991)), and *linear algebra*, where they are not. But the area containing both has long been recognised as being as much a part of the undergraduate mathematics curriculum as calculus or differential equations.

3. Multivariate analysis in statistics: development and absorption into the undergraduate curriculum

3.1. *Multivariate analysis: origins and early development*

The idea of analysing data collected on each of p variables for each of n individuals already existed at the start of the 20th century. In cases where all p variables featured in the analysis, however, one of them generally carried special importance, while the others served mainly to help either explain the special one or predict its future likely behaviour. This was generally done using the well-understood techniques of correlation and regression.

The first appearance of a technique for inspecting the overall features of a data set in which no variable carried any special emphasis was Karl Pearson's (Pearson 1901). This paper established for future generations of multivariate analysts the notion of a set of p observations as the coordinates of a point in Euclidean space — a representation already familiar from (multiple) regression, but without any axis now carrying any special importance, and closeness of fit being orthogonal distance from point to line or plane, rather than distance measured parallel to a particular axis.

Modern analysts of course recognise such subspaces of closest fit as *principal component* subspaces, but the description had not yet been coined, and users had perforce to treat the technique as purely descriptive. Inferential techniques need underlying population probability models, but only the multivariate normal (see §6.1 below) existed at this stage (having been comprehensively treated in Edgeworth (1892; 1893); see Chapter 9 of Stigler

(1986)), so inferential progress was limited. The breakthrough did not come until Wishart's generalization of the chi-squared distribution to what is now known as the *Wishart distribution* (Wishart 1928). This opened the floodgates to a decade of extraordinary developments, initially by three statisticians working independently in three different continents, who developed a set of distinct but interrelated techniques that still provide the basic tools of much modern multivariate analysis.

3.2. *The year 1936: Hotelling, Fisher, Mahalanobis*

First out of the starting blocks was Harold Hotelling (1895-1973) in America, who published first the multivariate generalization T^2 of the t-test (Hotelling 1931), then a follow-up to Pearson (1901) with a systematic development of *principal component analysis* (Hotelling 1933), and finally the technique of *canonical correlation analysis* (Hotelling 1936). These three papers form cornerstones of most traditional modern undergraduate lecture courses in multivariate analysis. Hard on Hotelling's heels came R A (Sir Ronald) Fisher (1890-1962) in England, with his derivation of the linear discriminant function (Fisher 1936), and P C Mahalanobis (1893-1972) in India, with his squared distance measure between populations, now known universally as Mahalanobis distance (Mahalanobis 1936) (see Stigler (2018) for his relationship with Fisher). Both of these were also published in 1936, surely making this a truly memorable year in the development of the subject.

Derivations in all these papers were done predominantly 'longhand' through term-by-term formulae and sets of simultaneous equations. The one flash of genuine advanced mathematics was the differential-geometric argument deployed by Mahalanobis — but no signs of matrices anywhere. Tentative first steps in the use of matrices occurred in the 1930s in connection with density functions and parameters of the multivariate normal and Wishart distributions in Wishart and Bartlett (1933), Aitken (1935; 1936), Bartlett (1938) and Ledermann (1939), but by the outbreak of the second world war these were the only instances of matrices appearing in multivariate work.

3.3. *The post-war period; Bartlett 1947*

After a further note in the sequence of papers concerned with the multivariate normal distribution in Lawley (1942), three publications in the immediate post-war years brought multivariate analysis firmly into its more recognisable modern shape. Kendall published the first edition of his *Advanced Theory of Statistics, Volume 1* in 1943, and of Volume 2 in 1946; each con-

tained a chapter on multivariate analysis (Chapters 15, 27; see Kendall and Stuart (1977; 1979)). Then Bartlett read his paper (Bartlett 1947) to the research section of the Royal Statistical Society; it had sections on multivariate analysis of variance, canonical reduction of the general regression problem, discriminant functions, and the general sampling theory of canonical roots. Finally, Rao read his paper (Rao 1948) on 'the utilisation of multiple measurements in problems of biological classification' to the same section of the Society, among other features extending Fisher's concept of two-group discrimination to more groups via canonical analysis. Rao's approach for the analysis was to represent the groups in the subspace of maximum average Mahalanobis distance between all pairs. A more tractable practical solution was soon provided by Bryan (1951), where successive canonical axes carry progressively decreasing proportions of between-group relative to within-group variation. This solution quickly became popularised as *canonical variate analysis*. It is worth noting that at the start of his paper Bartlett warns that he has not hesitated to use matrix and vector algebra or associated geometrical representation, and this mindset was adopted by all later authors. Consequently, from this date vectors, matrices and their associated operations became fundamental to all publications in multivariate analysis.

3.4. *Factor and Cluster analyses*

It is also from around this time that two essentially multivariate techniques that had evolved in specific disciplines became gradually subsumed into the statistical fold. One of these was *factor analysis*, which had its origins in the broadly social and behavioural sciences, perhaps more specifically in Education and Psychology, where reasonable models postulated that the observed values for a particular individual on measurable variables (for example IQ tests) were made up of a combination of values that individual had on a range of underlying but unobservable (*latent*) 'factors' (such as numerical ability and verbal ability). Attempts had been made since the early 1900s to extract an individual's score on each factor using various ad-hoc methods, until Lawley started to apply statistical principles such as maximum likelihood to the estimation problem (Lawley 1940; 1941). A summary of the state of play was provided by Lawley and Maxwell (1963, updated in 1971). A more recent and wider survey of latent variable and factor models is given in Bartholomew and Knott (1999).

The other technique that was brought into the multivariate canon at this time was *cluster analysis* or, more generally, *classification*. This is concerned

with subdividing a set of n individuals into distinct groups such that within a group the individuals were all 'similar' to each other in their responses to the p measured variables, and 'different' from individuals in other groups. Such clustering is of interest in various disciplines, for example ecology, biology, market research, and many different ad-hoc computational methods already existed. Fundamental to all of them was the specification of an initial measure of dissimilarity between two individuals, from which the $(n \times n)$ matrix of inter-object dissimilarities would provide the starting point of the calculations. A major advance was provided by Gower (1966), who showed that an eigendecomposition of an appropriately scaled dissimilarity matrix led to a subspace in which the individuals could be optimally represented. This brought the analysis into line with those of principal component and canonical variate analyses, and enabled groupings to be examined pictorially. He called the technique principal coordinate analysis, now better known as *metric scaling*. The similarity across all these techniques arises because they are all all based on optimising criteria of a particular form, as shown by Krzanowski (1971).

3.5. *Post 1960; computing and research*

After the wartime developments, computers had begun to appear in universities in the 1950s, and this led to the setting up of computing laboratories and computer science departments. There was now a real prospect of writing software to carry out all the techniques mentioned above, but computers still had limited capacity and were quite slow in operation. So it was paramount to write highly efficient programs that used as little computer storage as possible. This required expertise in numerical analysis as well as in programming, and for the commonly needed techiques like eigenvalue decomposition texts such as that by Wilkinson (1965) were invaluable. So was the statistical algorithms section which started in 1968 in the *Journal of the Royal Statistical Society Series C (Applied Statistics)*, and which over the next 20 years provided users with all manner of tailor-made routines.

Much statistical research is inevitably computationally intensive, and many techniques were developed from the 1960s onwards that could not be conducted without much computing effort. An early example is nonmetric multidimensional scaling – see Kruskal (1964a; 1964b). In general problems of inference, Bayesian methods have now come very much to the fore, with ready access to Markov Chain Monte Carlo (MCMC) methods for carrying out the inference in a wide variety of situations. On specifically multivari-

ate applications, early computer-intensive procedures such as bootstrapping, jackknifing and cross-validation have now been overtaken by neural networks, support vector machines, random forests and a whole panoply of pattern recognition techniques to help with data mining of enormous data sets. A description and discussion of all these techniques can be found in Webb and Copsey (2011).

3.6. *Teaching of Statistics*

From about the end of the 19th century, teaching of statistics (in those universities where it was done) had typically been split between one person in an applied department, to cover practical statistical methods, and another in the mathematics department, to cover theoretical statistics. For example, at Reading Harold Sanders covered statistics in the Faculty of Agriculture and Horticulture, while Arthur Bowley lectured in mathematics and economics (Curnow 2006), and at Cambridge G Udny Yule was followed by John Wishart in the Faculty of Agriculture, while Maurice Bartlett lectured in the Faculty of Mathematics (Whittle 1993). A department of applied statistics had existed at University College London since 1912, and the Cambridge Statistical Laboratory was established in 1947. By the 1960s, the pressure to expand became great in many other universities, and most of the newly created departments of statistics date from this time. This was also a time of revision of many mathematics degree programmes, with introduction of much 'modern mathematics' into the syllabus. So it is not surprising to find many degrees whose titles include statistics appearing from this date in university handbooks. These included joint degrees with mathematics, computer science, economics, physics, and other sciences, as well as single-subject statistics degrees. The increasingly available software programs and specifically statistics-orientated 'packages', such as SPSS, BMDP, Genstat, R and others, meant that inclusion of multivariate analysis in the syllabus was both possible and desirable. It would typically be included as a final-year course, usually as an option within a large set of options.

The demands of teaching prompted the appearance of a number of books on matrix theory for statisticians. We mention Searle (1982), Magnus and Neudecker (1988), and Harville (1997).

4. Notes on major contributors

4.1. *Fisher*

Fisher avoids matrices, though he makes extensive use of tables for displaying data. Indeed, rather like his contemporary Sir Harold Jeffreys, Fisher regarded himself as a Cambridge-trained applied mathematician, and seems to have relied on the mathematical machinery he learned at Cambridge in the early years of the last century, eschewing such things as measure theory and matrix algebra, relying instead on his genius, the promptings of the task in hand, and his great geometric insight (which he needed because of his poor eyesight). Lesser mortals need matrix algebra, and use it very effectively. One wonders how mathematicians did as well as they did without it for so long.

For background on Fisher, see the biography by his daughter Joan Fisher Box (Box 1978).

4.2. *Kendall*

M G (Sir Maurice) Kendall (1907-1983), whose career was spent outside academia, wrote *A course in multivariate analysis* (Kendall 1957). This does indeed use matrices, but largely for the purpose of displaying data. By contrast, his books *The advanced theory of statistics* (originally two volumes, 1943 on and 1946 on, by Kendall alone, the three-volume version with Stuart, 1958 on), whose numerous editions span the period 1943-1979, make heavy use of matrix algebra, at least in later editions.

Despite their value as works of reference, the deficiencies of Kendall's books are manifest, and well documented in *Mathematical Reviews*, to which we refer for scholarly and balanced criticism.

4.3. *Wilks and Anderson*

S S (Sam) Wilks (1906-1964) was one of the founding fathers of multivariate analysis, his Λ statistic proposed in Wilks (1932) matching Hotelling's T^2 in longevity of usefulness in hypothesis testing. He spent much of his career at Princeton, where he founded and led the department of statistics. He supervised the PhD of T W Anderson, whose book (Anderson 1958) we regard as the first of the unambiguously modern textbooks we cite. It thus seems odd at first sight that the later book (Wilks 1962) (whose last chapter is on multivariate analysis) does not use modern matrix notation, while Anderson's earlier book does. This was commented on in the (favourable) review of his book in *Mathematical Reviews*: 'The common matrix and vector notation and ready-made matrix manipulations are used very little.' (Much the same can be said of Cramér (1946), below.)

The explanation lies in Wilks's career choices. His 1962 book is the second by that title; the earlier 1943 version (Princeton University Press; favourably reviewed in *Mathematical Reviews* by Neyman) was his source for the later one. Wilks chose to focus on teaching, administration and academic leadership, at the expense of his own research. For more on Wilks's remarkable life and career, see Anderson's fine obituary of him (Anderson 1965).

5. Special results in linear algebra

We turn now to certain special results in linear algebra, crucially important in statistics, particularly multivariate analysis, needed in §6 below.

5.1. *Schur complements*

In 1917 Issai Schur (1875-1941) wrote the first part of an important two-part paper on analysis, which contained (p.215-216) what was named the *Schur complement* in 1968 by Emily Virginia Haynsworth (1916-1985). The origins of this can be traced back to Laplace in 1812 and Sylvester in 1851 (Puntanen and Styan 2005).

For a partitioned matrix

$$M = \begin{pmatrix} P & Q \\ R & S \end{pmatrix}$$

with P non-singular, the *Schur complement* of P in M is

$$M/P := S - RP^{-1}Q.$$

The *Schur determinant lemma* (Schur 1917, Hilfssatz) (which holds even for P singular) states that for P, Q, R, S $n \times n$ matrices such that P and R commute, and M the $2n \times 2n$ matrix above, then

$$\det\ M = \det(PS - RQ).$$

When P is non-singular, this gives the *Schur determinant formula*

$$\det(M) = \det(P).\det(M/P) = \det(P).\det(S - RP^{-1}Q).$$

Of course, no special role is played by P here. If instead one of Q, R, S is assumed non-singular, its Schur complement in M is

$$M/Q := R - RQ^{-1}P, \quad M/R := Q - PR^{-1}S, \quad M/S := P - QS^{-1}R.$$

If all four are non-singular, one has *Aitken's four-complements formula* (Aitken 1939a, 138-139):

$$M^{-1} = \begin{pmatrix} P & Q \\ R & S \end{pmatrix}^{-1} = \begin{pmatrix} (M/S)^{-1} & (M/Q)^{-1} \\ (M/R)^{-1} & (M/P)^{-1} \end{pmatrix}.$$

One also has *Aitken's block-diagonalization formula*: with M as above,

$$\begin{pmatrix} I & 0 \\ -RP^{-1} & I \end{pmatrix} \begin{pmatrix} P & Q \\ R & S \end{pmatrix} \begin{pmatrix} I & -P^{-1}Q \\ 0 & I \end{pmatrix} = \begin{pmatrix} P & 0 \\ 0 & M/P \end{pmatrix}.$$

If P is also partitioned, with 'top left-hand corner' P_{11}, one has the *Haynsworth quotient property* (Puntanen and Styan 2005, (6.0.25))

$$M/P = (M/P_{11})/(P/P_{11}).$$

5.2. *Inversion formulae*

Continuing from the above: the *Banachiewicz inversion formula* of 1937 states that for P and M/P both non-singular,

$$M^{-1} = \begin{pmatrix} P & Q \\ R & S \end{pmatrix}^{-1} = \begin{pmatrix} P^{-1} & 0 \\ 0 & 0 \end{pmatrix} + \begin{pmatrix} -P^{-1}Q \\ I \end{pmatrix} (M/P)^{-1}(-RP^{-1}\ I).$$

The *Duncan inversion formula* of 1944 (Puntanen and Styan 2005, (0.8.3)) states that if M/S is non-singular,

$$(M/S)^{-1} = (P - QS^{-1}R)^{-1} = P^{-1} + P^{-1}Q(S - RP^{-1}Q)^{-1}RP^{-1}.$$

Taking $S = I$ gives the *(Sherman-Morrison-)Woodbury formula* of 1950 (Golub and Van Loan 1996, 50); (Sherman and Morrison 1950); (Woodbury 1950)):

$$(P - QR)^{-1} = P^{-1} + P^{-1}Q(I - RP^{-1}Q)^{-1}RP^{-1}.$$

If $Q = q$ is a column-vector, $R = r^T$ a row-vector and $Q = q$ a scalar, the Duncan inversion formula becomes

$$(P - s^{-1}qr^T)^{-1} = P^{-1} + \frac{P^{-1}qr^TP^{-1}}{s - r^TP^{-1}q},$$

which (with $s = -1$) is *Bartlett's inversion formula* (Bartlett 1951).

For e_i the ith unit (column) vector, $u_i := P^{-1}e_i$ is the ith column of P^{-1},

and similarly $v_j^T = e_j^T P^{-1}$ is its jth row. Taking $Q = e_i$, $r = ke_j$ with scalar k gives

$$(P + ke_i e_j^T)^{-1} = P^{-1} + \frac{u_i v_j^T}{1 + kp^{ji}},$$

where $P^{-1} = (p^{ij})$ (venerable notation: see (Einstein, 1916, §8), (Turnbull, 1928, V.3), and below). This picks out the effect on the inverse of adding k to p_{ij}. Bartlett gives a short matrix proof, expanding the inverse in a formal geometric series, and checking the result by multiplying out. He credits the result to (Fisher, 1936). There (IV, The analogy of partial regression) Fisher in effect considers the sensitivity of his linear discriminant function to error or change in a data point, in the context of his famous iris data set.

We singled out Bartlett (1947) in §3 above; we return to it and Fisher (1936) in §6.7 below.

Note: Tensors. The 'venerable notation' above regarding superscripts and subscripts derives from tensor calculus, and through it, general relativity, areas in which, far from being a notational device, it is conceptually crucial. The term tensor was introduced by Voigt in 1898. Shortly after, G Ricci-Curbastro (1853-1925) and T Levi-Civita (1873-1941) published their memoir on 'absolute differential calculus' (now known as tensor calculus or Ricci calculus); see Ricci and Levi-Civita (1900, I.3, 134). This machinery was decisive for Einstein's general relativity; see Einstein (1916, (16), 787).

Tensor products are ubiquitous in linear and multilinear algebra, and functional analysis; see e.g. Landsberg (2012). Tensor methods are also widely used in statistics; see McCullagh (2017).

6. Case studies

6.1. *The multivariate normal (multinormal, MVN) distribution*

Any textbook on multivariate analysis will contain a treatment of the multinormal distribution, in p dimensions ('p for parameter'; we reserve n for the sample size), with mean vector μ and covariance matrix Σ, $N_p(\mu, \Sigma)$ say. It is instructive to compare them, technically and notationally.

Technically, the most important distinction lies in the definition, and how it handles the distinction between the full-rank case, where Σ has rank p and the distribution has support $\mathbb{R}^p$, and the singular case, where Σ is singular and the support has lower dimension. These are best handled together by taking the defining property to be that all linear combinations of coordinates

be univariate normal, as in Rao (1965). The two cases are handled briefly at the end of Cramér (1937, Chapter X), in scalar notation and in the language of quadratic forms, and at chapter length, with matrix notation, in Cramér (1946, Chapter 24). Here we find a good insight into Cramér's own thinking, or his assessment of his readership's thinking, at that time: he gives (Cramér 1946, (11.12.1a)) the p-fold integral for the multinormal characteristic function as we would, in matrix notation, and continues 'or in ordinary notation', to give it again in scalar notation and the language of quadratic forms.

The crux here is the formulae for the density and characteristic function of $N_p(\mu, \Sigma)$. The first of these is (allowing for notation) due to Edgeworth in 1892-3, so we follow Stigler (1986, Chapter 9) and call this *Edgeworth's theorem*. It involves the *inverse* of the covariance (or dispersion) matrix Σ; this is called the *precision matrix* $K := \Sigma^{-1}$ (or concentration matrix: 'K for Konzentration'). Then

$$\begin{aligned} f(x) &= |2\pi\Sigma|^{-1/2} \exp\{-\frac{1}{2}(x-\mu)^T\Sigma^{-1/2}(x-\mu)\} \\ &= |K/2\pi|^{1/2} \exp\{-\frac{1}{2}(x-\mu)^T K^{1/2}(x-\mu)\}, \end{aligned}$$

$$\phi(t) = \exp\{it^T\mu + \frac{1}{2}t^T\Sigma t\}.$$

It is a pleasure to compare the neat, self-contained efficiency of the exposition of this material in any modern book with anything pre-war; see e.g. Anderson (1958), Rao (1965), Mardia, Kent and Bibby (1979), Krzanowski (1988), Bingham and Fry (2010).

6.2. *The Gaussian Regression Formula (GRF)*

The essence of regression is conditioning, and the simplest setting is that of the multinormal above, with Σ, K, μ (conformably) partitioned:

$$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}, \quad K = \begin{pmatrix} K_{11} & K_{12} \\ K_{21} & K_{22} \end{pmatrix}, \quad \mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}.$$

The *Gaussian Regression Formula (GRF)* gives the conditional distribution of x_1 given x_2 as

$$x_1|x_2 \sim N(\mu_1 - K_{11}^{-1}K_{12}(x_2-\mu_2), K_{11}^{-1}) :$$

$$x_1|x_2 \sim N(\mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2-\mu_2), \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}). \qquad (GRF)$$

In particular, the conditional mean is *linear* in x_2, and the conditional variance is independent of x_2:

$$cov(x_1|x_2) = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} = \Sigma/\Sigma_{22},$$

the *Schur complement* of Σ_{22} in Σ. Its more traditional name in statistics is the *partial correlation matrix*; see Kendall and Stuart (1979, Chapter 27), in notation

$$cov(x_1|x_2) = \Sigma/\Sigma_{22} = \Sigma_{11.22} \quad (\text{or } \Sigma_{11.2}).$$

This alone justifies a place for the Schur complement in the canon of multivariate analysis (it also, of course, further justifies use of the concentration matrix).

The result goes back to Pearson (1903; 1911-12) as *Pearson's selection formulae* (selection being the term then used for conditioning). His work was simplified and extended in Aitken (1935; 1936), leading on to Bartlett (1938), Ledermann (1939), Lawley (1942), and Wishart (1955). This brings us close to our first modern textbook source, by T W Anderson (1918-2016) (Anderson 1958, Theorem 2.5.1).

The GRF can be found in verbal form in Kendall and Stuart (1977, Exercise 15.1). It appears in more detail in Kendall and Stuart (1979, §27.6) (first edition 1961). The theory goes through even with generalized (Moore-Penrose) inverses, as in Rao (1965); see Puntanen and Styan (2005, §6.2.2) for details.

6.3. *Covariance and concentration matrices: independence and conditional independence*

That two coordinates x_i, x_j of a random vector $x \sim N_p(\mu, \Sigma)$ are *independent* if and only if the *correlation* $\sigma_{ij} = 0$ – that is, that independence and uncorrelatedness are the same for Gaussians – is immediate from Edgeworth's formula above for the density, and indeed from Galton's earlier work on the bivariate normal (Stigler 1986, Chapter 8) of 1886 (from which regression and correlation stem). For, one may reduce to the bivariate case by conditioning on (or 'selecting', in Pearson's terminology above; see Stigler (2012, p8)) everything else. The two variables are independent if and only if their joint density, of the form $f(x, y) = c \exp\{-Q(x, y)\}$ with Q a quadratic form, *factorises*, the condition for which is that the off-diagonal coefficients, i.e. the correlations, vanish. All this was well known from Galton's time (so all the more from Pearson's, above).

That x_i, x_j are *conditionally independent given all the others* if and only if $k_{ij} = 0$ with $K = (k_{ij})$ the *concentration* matrix, is much more recent. Now taking x_1 as the random 2-vector and x_2 as the $(p-2)$-vector we condition on, the GRF shows us that the conditional density $f_{1|2}$ of $x_1|x_2$ has (conditional) covariance matrix K_{11}^{-1}. So one has conditional independence if and only if K_{11}^{-1} is diagonal, that is (the matrices being 2×2), K_{11} is diagonal, that is, $k_{12} = 0$.

This result is due to A. P. Dempster ((Dempster, 1972); see also (Dempster, 1969)); as it needs a name let us call it *Dempster's theorem*. Despite its simplicity, it is of great importance in modern statistics because of the role it plays in *graphical models* (see e.g. Lauritzen (1996, Proposition 5.2)). *Sparseness* properties of concentration matrices are highly revealing about structure, as well as being numerically very convenient (we note in passing the great importance of numerical linear algebra, here and elsewhere; see e.g. Golub and Van Loan (1996), Wilkinson (1965)). Conditional independence statements of this sort are important in Markov properties on undirected graphs (Hammersley-Clifford theorem; Lauritzen (1996, 3.2.1)), multivariate normal models (Lauritzen 1996, 5.1.3), covariance selection models ((Lauritzen 1996, 5.2), following A P Dempster), etc. They come into their own with Gaussian Markov random fields; see Rue and Held (2005), Bingham and Symons (2022).

Such things underlie much recent work on causation questions in statistics, and machine learning. For more here, we refer to Cox and Wermuth (1996, 2.12, 2.13, 3.4, 5.3.3). We recall in passing the 'conventional wisdom' we grew up with: '. . . : the presumption of causality must always be extra-statistical' (Kendall and Stuart 1979, §27.1). For later views, see Dawid (1979), Cox and Wermuth (2004), Pearl (2009a, 2009b).

6.4. *Maximum-likelihood estimation for the multivariate normal distribution*

In dealing with the multivariate normal, the calculus simplifies if one maximises the log-likelihood ℓ rather than the likelihood L.

For the multivariate normal, the maximum-likelihood estimators $\hat{\mu}$, $\hat{\Sigma}$ are the natural ones, the sample mean $\bar{x}$ and the sample covariance matrix $\bar{S}$ of the sample. For proofs (a simple trace calculation), see Anderson (1958, §3.2), Rao (1965, §8a.5), Mardia, Kent and Bibby (1979, §4.2.2), Bingham and Fry (2010, Th. 4.17).

6.5. *Mixed models in regression: Henderson's mixed-model equations*

In regression, one often encounters effects of two kinds: fixed effects, which one wishes to study (for example, the effects of age or gender on a disease), and random effects (the patients). For fixed effects, one seeks a *best linear unbiased estimator (BLUE)*; for random effects, the corresponding estimator is called the *best linear unbiased predictor (BLUP)* (see Robinson (1991)). The field of mixed models was pioneered in the US dairy industry by C. R. Henderson (1911-1989) from 1950 on, and his student S. R. Searle (1928-2013); their work is credited with producing great efficiency gains in the US dairy industry, for example in selecting bulls for breeding.

The key result is *Henderson's mixed-model equations*. See e.g. Bingham and Fry (2010, Th. 9.1) (for two proofs, one via the (Sherman-Morrison-) Woodbury formula, one via Bayes's theorem and Schur complements), Puntanen and Styan (2005, §6.3.12).

6.6. *Canonical correlation analysis*

The situation here is rather as with mixed models in regression as above, but now the focus is on *correlation* between two different kinds of fixed effects. For example, in a medical setting (Krzanowski 1988, §14.5) the fixed effects might be *simple* ones (that could be measured by a routine visit to a doctor's surgery), and *complex* ones (requiring laboratory measurement); as above, the random effects are the patients. Here one studies the relationship between the simple and complex effects, the aim being to devise a diagnostic system based on simple signs and symptoms. The correlation matrix R is partitioned conformably between the simple and complex effects. A Lagrange-multiplier argument leads to seeking the *largest* eigenvalue of $S_{22}^{-1}S_{21}S_{11}^{-1}S_{12}$ (or the same with the suffices interchanged), matrices of Schur-complement type.

6.7. *Discrimination and classification*

The starting point here is the pioneering paper (Fisher 1936), where Fisher looked for the linear combination $F = a^T x$ of the measured variables x on which to separate the means $\bar{x_1}$ and $\bar{x_2}$ of samples from two populations Π_1 and Π_2. Fisher made the prior assumption, also made in Hotelling (1931), that the variables have the same dispersion Σ in the two populations, which can be estimated by the within-sample variance-covariance matrix W. The optimum coefficients are given by maximising the standardised squared

distance between the two sample means,

$$[a^T(\bar{x_1} - \bar{x_2})]^2/a^T W a,$$

which leads to the linear combination, *Fisher's linear discriminant function*,

$$F \propto [W^{-1}(\bar{x_1} - \bar{x_2})]^T x.$$

Fisher also mentioned that F could be obtained through multiple regression of y on the variables x, with y a binary variable taking values $n_2/(n_1+n_2)$ and $-n_1/(n_1+n_2)$ in samples 1 and 2 respectively, a result proved formally in Hand (1981). Here F is the best linear function to *separate* the two samples and see how the individual values spread out, but can also used to *classify* a future observation. This is done by placing it in population 1 or 2 according as its function value is above or below some critical value c chosen with regard to such aspects as sample sizes and prior probabilities of observing values from each population.

In any practical situation the obtained F can be used to classify future observations into one of the groups. How can we estimate the chance of error in the classification? The apparent error rate is just the proportion of individuals in the two samples that are misclassified by F, but this estimate is clearly biased optimistically (because the calculated F is the one that best separates the samples). Randomly removing a portion of each sample, calculating F from the retained individuals and finding the number of removed individuals that are misclassified will give an unbiased estimate, but now of the wrong F! The best procedure is leave-one-out cross-validation (Lachenbruch and Mickey 1968) in which each individual is removed in turn, F is calculated from the rest, the omitted individual is classified, and the error rate is the proportion of such mis-classifications. This method is almost unbiased, and computation is very fast using Bartlett's inversion formula at each omission of an individual.

In the last 60 years there has been very keen interest and much research effort in designing various different types of classification function and methods of assessment and comparison of their performance. For details, see the books by Hand (1981; 1997), McLachlan (1992), Denison et al (2002) and Webb and Copsey (2011).

6.8. Singular values decomposition (SVD)

The origins of the singular values decomposition (SVD) (Krzanowski,

1988, §4.1) are given in the survey by Stewart (1993). The crucial link between the algebra and the statistics was the Eckart-Young theorem (Eckart and Young 1936), on best approximation (in matrix norm) by matrices of low rank. The statistical importance of this was pointed out in Good (1969). The SVD is important in numerical linear algebra in all contexts (including statistics) because of its numerical stability.

6.9. *Prediction theory*

When our sample is naturally ordered by *time*, that is, for a time series, the partial covariance or correlation above becomes particularly important, as the *partial autocorrelation coefficient (PACF)*. It is better here to work with complex values; then partial correlations have modulus in the unit disc $\mathbb{D}$ (by the Cauchy-Schwarz inequality). By the remarkable theorem of Verblunsky (1936), re-discovered much later in statistics (Barndorff-Nielsen and Schou 1973); Ramsey, 1974), there is a bijection (the *Verblunsky bijection*) between the set of $\mathbb{D}$-valued sequences $\alpha = (\alpha_n)_0^\infty$ and the probability measures μ on the unit circle (one-dimensional torus) $\mathbb{T}$:

$$\alpha \leftrightarrow \mu. \qquad (Ver)$$

Every time series can occur in this way, by the Kolmogorov Isomorphism Theorem of 1941; so can every autocorrelation function $\gamma = (\gamma_n)$, by Herglotz's theorem (or the Verblunsky bijection). For details and references, see Bingham (2012), for example. But α is much more convenient than γ here: α gives an *unrestricted parametrizaton*, while parametrization by γ is subject to nested determinental inequalities.

The paper (Schur, 1917) which gave us the Schur complement also showed that orthogonal polynomials on the unit circle (OPUC) arise here. These have a three-term recurrence relation involving *one* sequence, the (α_n) above – which in consequence have various names: Schur coefficients, Verblunsky coefficients, Geronimus coefficients, and PACF. (This is in contrast to orthogonal polynomials on the real line, OPRL, which have a three-term recurrence relation involving *two* sequences of constants, (a_n) and (b_n), say.) For a full account of the analytic theory, see Simon (2005), and for the probabilistic consequences, Bingham (2012).

6.10. *Non-negative matrices and Markov chains*

The theory of positive matrices rests on the Perron-Frobenius theory of O

Perron (1907) and G Frobenius (1909; 1912). It has been extensively applied to matrices of transition probabilities of Markov chains. These stem from A A Markov (1856-1922) in 1906; see his book (Markoff 1906, Anhang II). For a modern treatment, see Seneta (1981) (and the additional bibliography in the 2006 reprinting).

6.11. *Group representations*

It is difficult to over-emphasize the importance of the group concept in modern mathematics and physics, and group invariance is important in statistics; see (Eaton, 1983), (Eaton, 1989). We turn to the links with linear algebra.

The theory of group representations, in which groups are studied by means of matrices, matrix multiplication corresponding to the group operation, was developed by G Frobenius (1849-1917), I Schur (1875-1941) and W Burnside (1852-1927), around the period 1895-1915. The subject received a tremendous boost in 1925-26, with the development of quantum theory by Heisenberg and Schrödinger. The relevance of group representations to quantum mechanics was immediately developed by H Weyl (1885-1955) in his book of 1928 (German, 2nd edition 1931; English 1931). B L van der Waerden (1903-1996; book, 1932, German; 1974, English) and E P Wigner (1902-1995; book, 1931, German; 1959, English).

For an excellent introduction to group representations in probability and statistics, see Diaconis (1988). See Bayer and Diaconis (1992), Aldous and Diaconis (1986) for Diaconis's famous result on card-shuffling: (for a standard pack of 52 cards and the riffle shuffle) *seven shuffles suffice*. The background here involves the Diaconis 'cut-off phenomenon' and the theory of rapidly mixing Markov chains (Aldous, 1983).

6.12. *Information*

The important idea of *information* in statistics stems from Fisher's great foundational paper, (Fisher, 1922); this contains determinants and Hessians but not matrices. The *Fisher information matrix* evolved in response to the need to handle multidimensional parameters; see e.g. (Lehmann, 1983, 2.7). In 1945 the then young C. R. Rao (1920-) combined the Fisher information matrix with Riemannian geometry. This famous paper (Rao, 1945) also contains the Cramér-Rao inequality and Rao-Blackwellization; it has received much recent attention following Rao's centenary. The use of differential geometry in statistics has greatly expanded in recent times and led to the

emergence of the field of *information geometry*. For a monograph treatment, see (Amari, 2016).

6.13. *Random matrices*

As a probabilistic counterpart to §6.1-12, we close with random matrices; their very name shows clearly the influence of matrices in probability. The field is vast, fascinating and highly technical; we confine ourselves to brief mention of four aspects.

(i) Products of random matrices and Lyapunov exponents.

Interesting results on limits of products of random matrices by Furstenberg and Kesten in 1960, Kesten in 1973 and others were recognised as involving *Lyapunov exponents* (characteristic exponents). These govern the stability of dynamical systems: negative exponents indicate stability, but one positive exponent indicates exponential divergence from equilibrium, instability, even chaos. For details and references, see e.g. Guivarc'h (1980), Ledrappier (1984), Bougerol and Lacroix (1985).

(ii) The Wigner semi-circle law.

'In the 50s, Wigner suggested that the resonance lines of a heavy nucleus (their determination by analytic means being intractable) might be modelled by the spectrum of a large random matrix' (Katz and Sarnak 1999a, §2). This led to the *Wigner semi-circle law*. For the work of Wigner, Dyson, Mehta and others here, see Mehta (1967, Chapter 2), Pastur and Shcherbina (2011, §1.2).

(iii) The Tracy-Widom law.

The limit law of (appropriately normalised) extreme eigenvalues of random matrices of various types ('ensembles') was found by Tracy and Widom in 1994; it was a new type, involving the Airy kernel. See e.g. Blower (2009, §9.7, Chapter 10). Note (Baik-Deift-Johansson theorem, Blower (2009, Th. 10.4.7), Aldous and Diaconis (1999)) the non-standard centring (by $\sqrt{n}$) and scaling (by $n^{1/6}$).

(iv) Pair correlations.

In a famous chance meeting in 1974 (between the number theorist Hugh L Montgomery (1944-) and the theoretical physicist Freeman J Dyson (1923-2020), in Princeton), it emerged in conversation that the function

$$1 - \Big(\frac{\sin \pi x}{\pi x}\Big)^2,$$

which occurred in Montgomery's work on the pair correlation of zeros $\frac{1}{2} + i\gamma$

of the Riemann zeta function ζ, also occurred in Dyson's work on eigenvalues of random matrices (in the limit as $\gamma \to \infty$ for ζ and $n \to \infty$ for matrices of order n in the Gaussian Unitary Ensemble GUE respectively). This sparked an enormous amount of interest (as a possible new approach to the Riemann hypothesis (RH) from an apparently quite unrelated area), still ongoing. See e.g. Katz and Sarnak (1999a; 1999b), Keating and Snaith (2003).

Thanks to intensive computing efforts by Odlyzko and others, there is now overwhelming numerical evidence that these two phenomena are (in some sense) 'the same' (see Keating and Snaith (2003) for more on this, and the two relevant and visually indistinguishable graphs). But, no such thing has been proved, and proof is the essence of mathematics. It may well be that this matter, so closely related to RH, the resolution of which is the Holy Grail of mathematics, will remain mysterious as long as RH itself does.

7. Conclusion

To borrow the language of T S Kuhn's 1962 book *The structure of scientific revolutions*, where he speaks of paradigm shifts, we may identify three key dates here. They are: 1936, for the three key enabling technical advances (Hotelling, Fisher, Mahalanobis); 1947, for the paradigm shift to 'matrices with everything' in the research literature (Bartlett); and 1958, for that in the textbook literature (Anderson).

Our preferred metaphor for scientific change is that of a glacier, which is static or flowing, depending on how one looks at it and on what time-scale.

Acknowledgements. We thank both referees for their thorough and scholarly reports. We also thank Steve Stigler and Jim Pitman for helpful comments and references, Nick Woodhouse for comments on the physics background, and Killian Martin-Horgassan for his help with the style file.

References

Aitken, A C, 'Note on selection from a multivariate normal population', Proceedings of the Edinburgh Mathematical Society, 4 (1935), 106-110.
Aitken, A C, 'A further note on multivariate selection', Proceedings of the Edinburgh Mathematical Society, 5 (1936), 37-40.
Aitken, A C, Determinants and matrices, Edinburgh: Oliver & Boyd, 1939a (9th edition, 1956).
Aitken, A C, Statistical mathematics, Edinburgh: Oliver & Boyd, 1939b

(2nd edition, 1942).
Aldous, D J, 'Random walks on finite groups and rapidly mixing Markov chains', Séminaire de Probabilitée, XVII (1983), 243-297 (Springer: Lecture Notes in Mathematics 986).
Aldous, D J, and Diaconis, P, 'Shuffling cards and stopping times', American Mathathematical Monthly, 93 (1986), 333-348.
Aldous, D J, and Diaconis, P, 'Longest increasing subsequences: from patience to the Baik-Deift-Johansson theorem', Bulletin of the American Mathematical Society, 36 (1999), 413-432.
Amari, S-I, Information geometry and its applications, New York: Springer, 2016.
Anderson, T W, An introduction to multivariate statistical analysis, New York: Wiley, 1958 (2nd edition 1984, 3rd edition 2003).
Anderson, T W, 'Samuel Stanley Wilks, 1906-1964', Annals of Mathematical Statistics, 36 (1965), 1-23.
Barndorff-Nielsen, O E and Schou, G, 'On the parametrization of autoregressive models by partial autocorrelation', Journal of Multivariate Analysis, 3 (1973), 408-419.
Bartholomew, D J and Knott, M, Latent variable models and factor analysis, London: Arnold, 1999.
Bartlett, M S, 'Further aspects of the theory of multiple regression', Proceedings of the Cambridge Philosophical Society, 34 (1938), 33-40.
Bartlett, M S, 'Multivariate analysis', Journal of the Royal Statistical Society Series B, 10 (1947), 176-197.
Bartlett, M S, 'An inverse matrix adjustment arising in discriminant analysis', Annals of Mathematical Statistics, 22 (1951), 107-111.
Bartlett, M S, and Whittaker, J M, 'Obituary, A. C. Aitken', Biographical Memoirs of Fellows of the Royal Society, 14 (1968), 1-14.
Bingham, N H, 'Szegő's theorem and its probabilistic descendants', Probability Surveys, 9 (2012), 287-324.
Bingham, N H, and Fry, J M, Regression: Linear models in statistics, (Springer Undergraduate Mathematics Series (SUMS)), Heidelberg: Springer, 2010.
Bingham, N. H. and Symons, Tasmin L., Gaussian random fields: with and without covariances. *Theory of Probability and Mathematical Statistics* (Special Issue in memory of M. I. Yadrenko, ed. A. Olenko), to appear, 2022; arXiv:2111.11960.
Birkhoff, G, and MacLane, S, A survey of modern algebra, New York: Macmillan, 2nd edition 1953 (1st edition 1941).

Blower, G, Random matrices: High-dimensional phenomena, (London Mathematical Society Lecture Notes Series 367), Cambridge: Cambridge University Press 2009.
Bougerol, P, and Lacroix, J, Products of random matrices with applications to Schrödinger operators, (Progress in Probability and Statistics 8), Basel: Birkhäuser, 1985.
Box, J F, R A Fisher: The life of a scientist, New York: Wiley 1978.
Brechenmacher, F, 'Une histoire de l'universalité des matrices mathématiques', Revue de Synthèse, 131 (2010), 569-603.
Bryan, J G, 'The generalized discriminant function: mathematical foundation and computational routine', Harvard Educational Review, 21/2 (1951), 90-95.
Burkill, H, Ledermann, W, Hooley, C and Perfect, H, 'Obituary, Leon Mirsky', Bulletin of the London Mathematical Society, 18 (1986), 1945-206.
Cox, D R, and Wermuth, N, Multivariate dependencies: Models, analysis and interpretation, (Monographs in Statistics and Applied Probability 67), Boca Raton: Chapman & Hall/CRC, 1996.
Cox, D R, and Wermuth, N, 'Causality: A statistical view', International Statistical Review, 72 (2004), 285-305.
Cramér, H, Random variables and probability distributions, (Cambridge Tracts in Mathematics 36), Cambridge: Cambridge University Press, 1937 (2nd edition 1962, 3rd edition 1970).
Cramér, H, Mathematical methods of statistics, Princeton: Princeton University Press 1946 (1945, Almqvist & Wiksell, Stockholm).
Curnow, R N, Applied Statistics at the University of Reading: the first forty years, Reading: University of Reading, 2006.
Dawid, A P. 'Conditional independence in statistical theory (with discussion)', Journal of the Royal Statistical Society Series B, 41 (1979), 1-31.
Dempster, A P. Elements of continuous multivariate analysis, Reading MA: Addison-Wesley, 1969.
Dempster, A P, 'Covariance selection', Biometrics, 28 (1972), 157-175.
Denison, D G T, Holmes, C C, Mallick, B K, and Smith, A F M, Bayesian methods for nonlinear classification and regression, New York: Wiley, 2002.
Eaton, M. L., Multivariate statistics: A vector space approach. Wiley, 1983 (reprinted, Inst. Math. Stat., 2007).
Eaton, M. L., Group invariance in statistics. Inst. Math. Stat, 1989.
Eckart, C, and Young, G, 'The approximation of one matrix by another of lower rank', Psychometrika, 1 (1936), 211-218.

Edgeworth, F Y, 'Correlated averages', Philosophical Magazine:' (Series 5), 34 (1892), 190-204.
Edgeworth, F Y, 'Note on the calculation of correlation between organs', Philosophical Magazine (Series 5), 36 (1893), 350-351.
Einstein, A, 'Die Grundlage der allgemeinen Relativitätstheorie', Annalen der Physik (4) 49 (1916), 770-822.
Eisenhart, C., 'Boscovich and the combination of observations'. In Roger Joseph Boscovich (ed. L. L. Whyte), 200-212, Allen & Unwin , 1961.
Farebrother, R. W., Fitting linear relationships. A history of the calculus of observations 1750-1900. Springer, 1999.
Fisher, R. A., 'Applications of vector analysis to geometry'. Messenger of Math., 42 (1913), 161-178 (Collected Papers of R. A. Fisher, Vol. I (ed. J. H. Bennett), University of Adelaide Press, 1971).
Fisher, R A, 'The use of multiple measurements in taxonomic problems', Annals of Eugenics, 7 (1936), 179-198 (Collected Papers of R. A. Fisher, Vol. III (ed. J. H. Bennett), University of Adelaide Press, 1973).
Gantmacher, F H, The theory of matrices, Vol. 1,2, New York: Chelsea, 1959 (Russian 1953).
Golub, G H, and Van Loan, C F, Matrix computations, Baltimore: Johns Hopkins University Press, 3rd edition, 1996 (1st ed. 1983).
Good, I J, 'Some applications of the singular decomposition of a matrix', Technometrics, 11 (1969), 823-831.
Gower, J C, 'Some distance properties of latent root and vector methods used in multivariate analysis', Biometrika, 53 (1966), 325-338.
Guivarc'h, Y, 'Quelques propriétés asymptotiques des produits des matrices aléatoires', Ecole d'Eté de Probabilités de Saint-Flour VIII-1978, 177-234 (Lecture Notes in Mathematics, 774 (1980), New York: Springer).
Hand, D J, Discrimination and classification, New York: Wiley, 1981
Hand, D J, Construction and assessment of classification rules, New York: Wiley 1997.
Harville, D A, Matrix algebra from a statistician's perspective, Heidelberg: Springer, 1997.
Hawkins, T, 'The theory of matrices in the 19th century,' Proceedings of the Internatational Congress of Mathematics Vancouver 1974, 2 (1975a), 561-570.
Hawkins, T, 'Cauchy and the spectral theory of matrices', Historia Mathematica, 2 (1975b), 1-29.
Horn, R A, and Johnson, C R, Matrix analysis, Cambridge: Cambridge Uni-

versity Press, 1985 (2nd edition 2013).
Horn, R A, and Johnson, C R (1991), Topics in matrix analysis, Cambridge: Cambridge University Press, 1991.
Hotelling, H, 'The generalization of Student's ratio', Annals of Mathematical Statistics, 2(1931) 360-378.
Hotelling, H, 'Analysis of a complex of statistical variables into principal components', Journal of Educational Psychology, 24 (1933), 417-441 and 498-520.
Hotelling, H, 'Relations between two sets of variates', Biometrika, 28 (1936), 321-377.
Katz, N M, and Sarnak, P C, 'Zeros of zeta functions and symmetry', Bulletin of the American Mathematical Society, 36 (1999a), 1-26.
Katz, N M, and Sarnak, P C, Random matrices, Frobenius eigenvalues and monodromy, Providence: American Mathematical Society, 1999b.
Keating, J P, and Snaith, N C, 'Random matrices and L-functions', Journal of Physics A: Mathematical and General, 36 (2003), 2859-2881.
Kendall, M G, A course in multivariate analysis, London: Griffin, 1957.
Kendall, M G, and Stuart, A, The advanced theory of statistics, Volume 1, Distribution theory, 4th edition, London: Griffin, 1977.
Kendall, M G, and Stuart, A, The advanced theory of statistics, Volume 2, Inference and relationship, 4th edition, London: Griffin, 1979.
Kruskal, J B, 'Multidimensional scaling by optimising goodness-of-fit to a non-parametric hypothesis' Psychometrika, 29 (1964a), 1-27.
Kruskal, J B, 'Non-metric multidimensional scaling: a numerical method', Psychometrika, 29 (1964b), 115-129.
Krzanowski, W J, 'The algebraic basis of classical multivariate methods', The Statistician, 20 (1971), 51-61.
Krzanowski, W J, (1988). Principles of multivariate analysis: a user's perspective, Oxford: Oxford University Press, 1988.
Lachenbruch, P A, and Mickey, M R, 'Estimation of error rates in discriminant analysis', Technometrics, 10 (1968), 1-11.
Landsberg, J M, 'Tensors: geometry and applications' (Graduate Studies in Mathematics **128**), Providence: American Mathematical Society.
Lauritzen, S L, Graphical models, Oxford: Oxford University Press, 1996.
Lawley, D N, 'The estimation of factor loadings by the method of maximum likelihood', Proceedings of the Royal Society of Edinburgh, Section A, 60 (1940), 64-82.
Lawley, D N, 'Further investigations in factor estimation', Proceedings of the

Royal Society of Edinburgh, Section A, 61 (1941), 176-185.
Lawley, D N, 'A note on Karl Pearson's selection formulae', Proceedings of the Royal Society of Edinburgh, Section A, 62 (1942), 28-30.
Lawley, D N, and Maxwell, E A, Factor analysis as a statistical method, New York: American Elsevier, 1971 (1st edition 1963).
Ledermann, W, 'Sampling distribution and selection in a normal population', Biometrika, 30 (1939), 295-304.
Ledrappier, F, 'Quelques propriétés des exposants characteristiques', Ecole d'Eté de Probabilités de Saint-Flour XII-1982 (1984), 305-396 (Lecture Notes in Mathematics, 1097, Heidelberg: Springer).
Lehmann, E H, Theory of point estimation, New York: Wiley, 1983.
Magnus, J R, and Neudecker, H, Matrix differential calculus, New York: Wiley, 1988.
Mahalanobis, P C, 'On the generalized distance in statistics', Proceedings of the National Institute of Science, India, 2 (1936), 49-55.
Mardia, K V, Kent, J T, and Bibby, J M, Multivariate analysis, London: Academic Press, 1979.
McCullagh, P, Tensor methods in statistics, 2nd ed., (Monographs in Statistics and Applied Probability 29), Boca Raton: Chapman & Hall/CRC, 2017 (1st ed. 1987).
McLachlan, G J, Discriminant analysis and statistical pattern recognition, New York: Wiley, 1992.
Markoff, A A, Wahrscheinlichkeitsrechnung, Leipzig: B G Teubner, 1912.
Mehta, M L, Random matrices and the statistical theory of energy levels, London: Academic Press, 1967 (2nd edition 1991, 3rd edition 2004).
Mirsky, L, An introduction to linear algebra, Oxford: Oxford University Press, 1955.
Muir, T, The theory of determinants in the historical order of development, Volume 1, London: Macmillan, 1906.
Pastur, L, and Shcherbina, M, Eigenvalue distributions of large random matrices. Providence: American Mathematical Society (Mathematical Surveys and Monographs, 17).
Pearl, J, Causality: models, reasoning and inference, Cambridge: Cambridge University Press, 2nd edition, 2009a (1st edition 2000).
Pearl, J, 'Causal inference in statistics: an overview', Statistics Surveys, 3 (2009b), 96-146.
Pearson, K, 'On lines and planes of closest fit to a system of points in space', Philosophical Magazine, Series 6, 2 (1901), 559-572.

Pearson, K, 'Mathematical contributions to the theory of natural selection. XI. On the influence of natural selection on the variability and correlation of organs', Philosophical Transactions of the Royal Society, 200 (1903), 1-66.
Pearson, K, 'On the general theory of the influence of selection on correlation and variation', Biometrika, 8 (1911-12), 437-443.
Puntanen, S, and Styan, G P H, 'Schur complements in probability and statistics', Chapter 6, 163-226 in Zhang, F (2005)
Ramsey, A. S., Dynamics, Part II, Cambridge University Press, 1937 (2nd ed. 1944).
Ramsey, F L, 'Characterization of the partial autocorrelation function', Annals of Statistics, 2 (1974), 1296-1301.
Rao, C R, Information and the accuracy attainable in the estimation of statistical parameters, Bulletin of the Calcutta Mathematical Society, 37 (1945), 81-91.
Rao, C R, 'The utilisation of multiple measurements in problems of biological classification', Journal of the Royal Statistical Society Supplement, 9 (1948), 159-203.
Rao, C R, Linear statistical inference and its applications, New York: Wiley, 1965 (2nd edition 1973).
Ricci, G and Levi-Civita, T, Méthodes de calcul différentiel absolu et leurs applications, Mathematische Annalen, 54 (1900), 125-201.
Robinson, G K, 'That BLUP is a good thing: The estimation of random effects', Statistical Science, 6 (1991), 15-32.
Rue, H, and Held, L, Gaussian Markov random fields, (Monographs in Statistics and Applied Probability, 104) Boca Raton: Chapman and Hall/CRC, 2005.
Schur, I, 'Über Potenzreihen, die im Innern der Einheitskreises beschränkt sind', Journal für die reine und angewandte Mathematik, 147 (1917), 205-232.
Searle, S R, Matrix algebra useful for statistics, New York: Wiley, 1982.
Seneta, E, Non-negative matrices and Markov chains, Heidelberg: Springer, 1981 (revised 2006).
Sherman, J, and Morrison, W J, 'Adjustment of an inverse matrix corresponding to a change in one element of a given matrix', Annals of Mathematical Statistics, 21 (1950), 124-127 (Abstract, 20 (1949), 62).
Simon, B, Orthogonal polynomials on the unit circle. Part 1: Classical theory, Providence: American Mathematical Society (American Mathematical Society Colloquium Publications 54.1).

Stewart, G W, 'On the early history of the singular values decomposition', SIAM Review, 35 (1993), 551-566.
Stigler, S M, The measurement of uncertainty before 1900, Harvard: Harvard University Press, 1986.
Stigler, S M, 'Studies in the history of probability and statistics L: Karl Pearson and the Rule of Three', Biometrika, 99 (2012), 1-14.
Stigler, S M, 'Mahalanobis & Fisher: mathematical statistics as a global enterprise', Sanhkyā B, 80 (2018), supplement, S167-S178.
Todhunter, I., 'On the method of least squares'. Trans. Cambridge Phil. Soc., 11 (Part II) (1869), 219; Addition, 238.
Turnbull, H W, The theory of determinants, matrices and invariants, London: Blackie, 1928 (2nd edition 1945).
Turnbull, H W, and Aitken, A C, An introduction to the theory of canonical matrices, London: Blackie, 1932.
Verblunsky, S, 'On positive harmonic functions (second paper)', Proceedings of the London Mathematical Society, 40 (1936), 290-320.
Weatherburn, C. E., Elementary vector analysis, with applications to geometry and mechanics. G. Bell and Sons, 1921 (2nd ed. 1955).
Webb, A R, and Copsey, K D, Statistical pattern recognition (3rd edition), New York: Wiley, 2011.
Whittle, P, 'A Realised Path. The Cambridge Statistical Laboratory up to 1993' (revised 2002), `www.statslab.cam.ac.uk/history-statistical-laboratory`. Statistical Laboratory, University of Cambridge, Cambridge.
Wilkinson, J H, The algebraic eigenvalue problem. Oxford: Oxford University Press, 1965.
Wilks, S S, 'Certain generalizations in the analysis of variance', Biometrika, 24 (1932), 471-494.
Wilks, S S, Mathematical statistics, New York: Wiley, 1962.
Wishart, J, 'The generalised product moment distribution in samples from a normal multivariate population', Biometrika, 20A (1928), 32-52.
Wishart, J,' Multivariate analysis', Journal of the Royal Statistical Society Series C, 4 (1955), 103-116.
Wishart, J, and Bartlett, M S, 'The generalised product moment distribution in a normal system', Proceedings of the Cambridge Philosophical Society, 29/2 (1933), 260-270.
Woodbury, M A, Inverting modified matrices, (Statistical Research Group Memorandum Report, 42) Princeton: Princeton University Press, 1950.
Ycart, B, 'A case of mathematical eponymy: the Vandermonde determinant',

Revue d'Histoire des Mathématiques, 19 (2013), 43-77.
Zhang, Fuzhen (editor), The Schur complement and its applications, (Numerical Methods and Algorithms 4), Heidelberg: Springer, 2005.