Whence Principal Components?

The *Journal of Educational Psychology* as a Precursor to *Psychometrika*

Before the establishment of the journal *Psychometrika* in 1936, the main outlet for the publication of technical/mathematical material with a psychological bent was, somewhat surprisingly, the *Journal of Educational Psychology (JEdP)*. *JEdP* was founded in 1910, with an opening lead article written by E. L. Thorndike (the second President of the Psychometric Society after Thurstone). By the time the 1930s arrived, *JEdP* was dominated by authors who would later become inaugural members of the Psychometric Society as well as some of its later presidents. For example, in the 1930 volume there were quantitative articles written by the familiar names of Cureton, Dunlap, Holzinger, Spearman, Rulon, Lindquist, Edgerton, Garrett, and Carter. (We might add that in the 1930s and 40s, Jack Dunlap, one of the six founding members of the Psychometric Society, was an Editor of *JEdP*, with responsiblity for all technical/quantitative submissions.) So it may not be completely surprising that Harold Hotelling, one of the leading mathematical statisticians of the 20th century, would publish his method of principal components in *JEdP* in 1933.[1] What may be more interesting historically, however, is

[1]Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, *24*, 417–441; 498–520. (6034 citations in Google Scholar as of 4/1/2016)

The technical level of Hotelling's 1933 *JEdP* article is quite high and would be unexpected in any journal devoted mainly to substantive matters. For example, Darrell Bock in his chapter, "Rethinking Thurstone," in the book, *Factor Analysis at 100* (2007), comments on Hotelling's *JEdP* articles as follows (p. 42):

Speaking of notation, I add that although Hotelling may have derived his iterative procedure for latent roots and vectors in matrix terms, in consideration of the audience, he confined his presentation to scalar algebra. Curiously, however, he introduces a notational convention from tensor calculus — namely, that when how Hotelling came to the topic in the first place — that story is the purpose of this note.

Harold Hotelling and Truman Lee Kelley

Harold Hotelling (1895–1973) received his doctoral degree in mathematics (and economics) from Princeton in 1924. Immediately thereafter he became a Research Associate at the Stanford University Food Research Institute; from 1927 to 1931 he was an Associate Professor of Mathematics, also at Stanford. He moved to the Economics Department of Columbia University in 1931, and stayed until 1946 when he left for the University of North Carolina to found the Department of Statistics. He remained a Professor of Mathematical Statistics at North Carolina until his death. Judging from a perusal of the Harold Hotelling archives at Columbia University and those of Truman Lee Kelley at Harvard, Hotelling's work on principal components (as well as his subsequent development of canonical correlation — see, for example, Hotelling's other *JEdP* publication, "The most predictable criterion" [1935, *26*, 139–142]), was motivated by his association with Kelley. They overlapped as colleagues at Stanford from 1924 to 1931, where Kelley was a Professor of Education. Kelley moved in 1931 to the Harvard Graduate School of Education at exactly the same time that Hotelling moved to Columbia. As discussed below, this period of the early 1930s was a time of sustained interaction between Kelley and Hotelling that directly led to Hotelling's development of principal components and canonical correlation.

an equation is written as say, $b_i = a_{ij}$, it denotes the summation of the right-hand member with respect to the j subscript. This device is somewhat unsettling to anyone accustomed to seeing the summation sign in these equations. Surely, this is the only paper containing tensor notation in the entire psychological literature and perhaps the statistical literature.

The same year that both Kelley and Hotelling left Stanford for their respective East Coast positions at Harvard and Columbia (1931), also saw the formation of the Unitary Traits Committee under E. L. Thorndike, with both Kelley and Hotelling as committee members. Several excerpts are given below from a survey that discusses the work of this group written by Karl Holzinger in the *Journal of Personality* (1936, *4*, 335-343), entitled "Recent research on unitary mental traits":

> When Professor Spearman conceived the idea that the arrangement of a set of intercorrelations could be used to determine factors underlying a set of variables, he opened up an objective method in psychology that has been gathering momentum ever since. After the publication of *Abilities of Man*, in 1927, interest in factor theory began to spread widely throughout America, engaging the attention of such workers as Professor Truman Kelley and Professor T. V. Moore. In a book entitled *Crossroads in the Mind of Man* (1928) Professor Kelley dealt largely with group factors and new methods for their evaluation. These two volumes laid the immediate foundation for the formation of the Unitary Traits Committee in 1931.
>
> Professor E. L. Thorndike, for years a passive onlooker of methods of factorization, now became an active promoter. Through his influence a committee was formed to study methods of factorization and apply them if possible to large bodies of data. Professor Thorndike named the committee the Unitary Traits Committee and with his characteristic symbolism, "U. T. C. for short."
>
> The Problems and Plans Committee of the American Council on Education empowered Professor Thorndike to act as chairman of this committee and secured a grant of money from the Carnegie Corporation for the purpose of preparing a plan to study unitary differential traits. The early members of this committee included Professors E. L. Thorndike, Charles Spearman, T. L. Kelley, Clark Hull, Karl Lashley, and Karl J. Holzinger. At later meetings Professors T. V. Moore, Henry Garrett, and Harold Hotelling were added to the committee.

...

The sub-committees were organized as follows:

1. Mathematical theory and techniques and the improvement of methods of analysis. T. L. Kelley and Harold Hotelling.

...

During the early meetings of the Unitary Traits Committee some criticism was made of existing methods of factorization, chiefly those of Professor Kelley in *Crossroads in the Mind of Man*. Professor Kelley was already at work amending these techniques, and enlisted the aid of Professor Harold Hotelling to further this work. As a mathematical statistician Professor Hotelling was of great service to the committee. He contributed many valuable suggestions at meetings, and the factorization technique now known as the Method of Principle [sic] Components.

The remainder of this note can be seen as a series of interesting subtopics (or at least we hope they are) concerning the introduction of "the method of principal components" in *JEdP* (1933). Several of these observations result from private correspondence and material from the Unitary Traits Committee available in archives for Kelley and Hotelling at Harvard and Columbia, respectively.

Hotelling as a Quantitative Consultant for Psychology

For a period of time in the late 1920s and 30s, Harold Hotelling was a favored mathematician to consult when a particularly vexing quantitative derivation task was at hand. Acknowledgments to Hotelling appeared regularly in *JEdP* in the early 1930s; others occurred in several books from around that same time. For example, in Kelley's *Interpretation of Educational Measurements* (1927), we have the footnote (p. 213): "I am indebted to Dr. Harold Hotelling for a suggestion which readily led to the evaluation of this determinant." Or, in Kelley's *Crossroads in the Mind of Man* (1928), we have the following in the actual text (p. 54): "Dr. Harold Hotelling has kindly provided the following set of necessary conditions which are more readily investigated than are the 12 sufficient equations in Formula 35."[2]

In John Flanagan's thesis under Kelley at Harvard in 1935, *Factor Analysis in the Study of Personality*, there is the following paragraph about Hotelling developing the method of principal components at the behest of the Unitary Traits Committee:

> This brings us directly to the last method of multiple-factor analysis which we shall consider, that of Hotelling. At the request of the Unitary Traits Committee, Hotelling attacked the problem of obtaining a serviceable solution to the problem proposed by Kelley in 1928 [in *Crossroads in the Mind of Man*], "first, a determination, having tests A, B, C, of what the independent mental traits are; and secondly an experimental construction of new tests measuring these independent traits." As we have just noted, Hotelling's least-squares conditions are identical to those in one of the solutions presented by Thurstone. Dr. Hotelling, however, has supplied a very neat iterative solution for the k^{th} order determinant involved which makes the solution comparatively short.

The role of the Unitary Traits Committee in facilitating the development of the method of principal components is confirmed by the beginning footnote in Hotelling's paper in *JEdP* (1933):

> A study made in part under the auspices of the Unitary Traits Committee and the Carnegie Corporation.
>
> The author is indebted to Professor Truman L. Kelley, who was responsible for the initiation of this study and the propounding of many of the questions to which answers are here attempted; also to Professors L. L. Thurstone,

[2]It might also be noted that Hotelling was an inaugural member of the Psychometric Society based on the membership roster published in 1936. For some unknown reason, however, he was no longer a member as of March, 1939.

Clark V. [sic; it should be L.] Hull, C. Spearman, and E. L. Thorndike, who raised some of the further questions treated.

In a four-page single-spaced letter to Kelley from Hotelling (June 2, 1932), the approach that Hotelling was to take is spelled out in some detail: "Another line of possible development in tetrad analysis (or rather factor analysis) is to take as independent factors those linear functions of a number of test scores which correspond to the principal axes of the ellipsoids of the scatter diagram." Apparently, this long letter (along with some extensive handwritten notes) served as a proposal to work for the Unitary Traits Committee for two summer months in 1932 (for $800); Kelley responded to Hotelling with a letter dated June 20, 1932:

> This letter is in confirmation of our agreement that you work for the Unitary Traits Committee for a period to two months and receive therefore a total of $800.00. It is understood between us that you are to be free to meet such other obligations during this time as incidentally arise, and that we upon our part may occasionally call upon you in the future for things not involving an extended study upon your part.
>
> I am sending a copy of this letter to Dr. Thorndike, chairman of the Committee.
>
> I am returning herewith your notes, for which please accept my thanks.

Hotelling replied on June 25, 1932 (with a notation that a copy was also sent to E. L. Thorndike):

> With your letter of June 20 this will confirm our agreement that I am to work for the Unitary Traits Committee for two months this summer.
>
> Thank you for the return of my rough notes, which I hope latter to elaborate. During the past week at Syracuse I have been discussing their contents at considerable length with L. L. Thurstone, Jack Dunlap, and Ragnar Frisch. Dunlap is going to try the method of principal axes on some tests he has made of chickens. [sic?; "children"?]

I hope to be at Blackey's Hotel at Gilmanton Iron Works early in July and to see you there. Meanwhile I am wrestling with some of the very beautiful and intricate mathematical problems involved.

This last letter is interesting for several reasons, and particularly for the three people Hotelling mentioned that he had extensive discussions with: L. L. Thurstone, Jack Dunlap, and Ragnar Frisch. The 1932 Syracuse meeting referred to was of the American Association for the Advancement of Science and its many affiliated societies (such as the American Psychological Association). At this meeting, Thurstone presented his own principal axes solution to the problem of factor analysis. As Hotelling notes in a 1933 *JEdP* footnote:

> Since this was written Professor Thurstone has kindly sent me a pamphlet he has prepared for class use, in which he uses the same geometric interpretation as in the present section, and discusses the problem from essentially the same standpoint as that taken in [Part One]. His iterative procedure appears to have no relation to that of [Part Four]. In June, 1932, Professor Thurstone presented at the Syracuse meeting of the American Association for the Advancement of Science certain of the considerations which have served as a point of departure for this paper.

Interestingly, Thurstone abandoned his first principal axes approach because he thought it did not conform to a "true" and psychologically meaningful factor analytic model. This particular debate between the use of principal components and the reliance on the factor model rages to this day.[3]

[3] Thurstone's 1935 book, *The Vectors of Mind: Multiple Factor Analysis for the Isolation of Primary Traits*, includes a Chapter IV on "The Principal Axes." Even at this early date, Thurstone concludes with a summary rejection of principal components as a viable approach to the factor model (p. 132): "These considerations make it necessary to discard the method of principal axes and also Hotelling's special case of this method as solutions to the psychological factor problem."

When I've taught modules on principal components analysis (PCA) and factor analysis (FA) in a Multivariate Analysis class, I introduce PCA with three introductory points:

(a) Principal component analysis (PCA) deals with only one set of variables without the need for catego-

The mention of Jack Dunlap in Hotelling's letter is also interesting because he was to be the Editor of *JEdP* overseeing the publication of Hotelling's 1933 contribution. Ragnar Frisch, for those who might not know, was the first recipient of the Nobel Prize in Economic Sciences in 1969; he is recognized as founding the discipline of econometrics and for coining the word pair "macroeconomics/microeconomics" in 1933.

Hotelling's Power Method

At the meeting of the Unitary Traits Committee in December of 1932, several papers were read that were devoted to numerical exam-

---

rizing the variables as being independent or dependent. There is asymmetry in the discussion of the general linear model; in PCA, however, we analyze the relationships among the variables in one set and *not* between two.

(b) As always, everything can be done computationally without the Multivariate Normal (MVN) assumption; we are just getting descriptive statistics. When significance tests and the like are desired, the MVN assumption becomes indispensable. Also, MVN gives some very nice interpretations for what the principal components are in terms of our constant density ellipsoids.

(c) Finally, it is probably best if you are doing a PCA, not to refer to these as "factors". A lot of blood and ill-will has been spilt and spread over the distinction between component analysis (which involves linear combinations of *observable* variables), and the estimation of a factor model (which involves the use of underlying latent variables or factors, and the estimation of the factor structure). We will get sloppy ourselves later, but some people really get exercised about these things.

In introducing FA, I begin with four introductory points:

(a) In a principal component approach, the emphasis is completely on linear combinations of the observable random variables. There is no underlying (latent) structure of the variables that I try to estimate. Statisticians generally love models and find principal components to be somewhat inelegant and nonstatistical.

(b) The issue of how many components should be extracted is always an open question. With explicit models having differing numbers of "factors," we might be able to see which of the models fits "best" through some formal statistical mechanism.

(c) Depending upon the scale of the variables used (i.e., the variances), principal components may vary and there is no direct way of relating the components obtained on the correlation matrix and the original variance-covariance matrix. With some forms of factor analysis, such as maximum likelihood (ML), it is possible to go between the results obtained from the covariance matrix and the correlations by dividing or multiplying by the standard deviations of the variables. In other words, we can have a certain type of "scale invariance" if we choose, for example, the maximum likelihood approach.

(d) If one wishes to work with a correlation matrix and have a means of testing whether a particular model is adequate or to develop confidence intervals and the like, it is probably preferable to use the ML approach. In PCA on a correlation matrix, the results that are usable for statistical inference are limited and very strained generally (and somewhat suspect).

ples of Hotelling's iterative strategy for obtaining the principal components of a correlation matrix. The procedure proposed by Hotelling would today be referred to as (a repeated use of) the power method for finding the dominant eigenvalue of a matrix. Bodewig (*Matrix Calculus*, 1956, p. 250) attributes the power method to von Mises in 1929, as published in a rather obscure German language periodical. However, because of the close date to Hotelling's own use of a power method and his not referencing von Mises (but he did so later in an *Annals of Mathematical Statistics* article in 1943, entitled "Some new methods in matrix calculation"), the power method itself might just as well be attributed to Hotelling. In fact, Hotelling's repeated use of the power method to find *all* the eigenvalues and eigenvectors of a matrix, involve what have now become well-known as Hotelling deflations: these are outer products of an eigenvector with itself, weighted by the eigenvalue, and subtracted from the starting matrix. We give a summary of this process taken from Morrison (1967; *Applied Multivariate Analysis*):

Let $\mathbf{A}$ be the $p \times p$ matrix of real elements. It is not necessary that $\mathbf{A}$ be symmetric. Order the characteristic roots λ_i of $\mathbf{A}$ by their absolute values:

$$|\lambda_1| > |\lambda_2| \geq \cdots \geq |\lambda_p|$$

and denote their respective characteristic vectors as $\mathbf{a}_1, \ldots, \mathbf{a}_p$. Initially we shall require that only $|\lambda_1| > |\lambda_2|$. Let $\mathbf{x}_0$ be any vector of p real components, and form the sequence: $\mathbf{x}_1 = \mathbf{A}\mathbf{x}_0; \ldots \mathbf{x}_n = \mathbf{A}\mathbf{x}_{n-1} = \mathbf{A}^n\mathbf{x}_0$ of vectors. Then if the successive $\mathbf{x}_i$ are scaled in some fashion, the sequence of standardized vectors will converge to the characteristic vector $\mathbf{a}_1$. Probably the most convenient scaling is performed by dividing the elements by their maximum, with normalization to unit length merely reserved for the last, or exact, vector. Since $\mathbf{A}\mathbf{a}_1 = \lambda_1\mathbf{a}_1$ the characteristic root itself can be found by dividing any element of $\mathbf{A}\mathbf{a}_1$ by the corresponding element of $\mathbf{a}_1$. The same iterative procedure can be used to compute any distinct characteristic root of $\mathbf{A}$. To extract the second largest root and its vector we normalize the first characteristic vector $\mathbf{a}_1$ to unit length, form the $p \times p$ matrix $\lambda_1 \mathbf{a}_1 \mathbf{a}_1^{'}$ and subtract it from $\mathbf{A}$ to give the residual matrix $\mathbf{A}_1 = \mathbf{A} - \lambda_1 \mathbf{a}_1 \mathbf{a}_1^{'}$. [A Hotelling deflation]

In the more recent implementations of routines for finding the principal components of a covariance matrix (such as in Matlab), Hotelling's iterative procedure is not used. Instead, some type of Jacobi strategy for finding the eigenvalues/eigenvectors of a matrix is commonly adopted [we will come back to this topic shortly]. This replacement may be due in part to the computational difficulties one might encounter with Hotelling's approach. As Bodewig notes (1956, p. 250):

> It was R. von Mises ... who found the power method. It was a great achievement. And in many cases it gives a quick result. But it cannot be denied that in a large number of cases the convergence is extremely bad, so bad in fact that it can hardly be used at all. The convergence will be good enough only if the quotient $|\frac{\lambda_1}{\lambda 2}| > 3$. But this is only rarely the case.

We might mention that there is one prominent and current application of the power method for finding a single dominant eigenvalue/eigenvector combination — this is in Google's search engine and the use of what is called PageRank.

Hotelling's 1936 *Psychometrika* Paper: "Simplified calculation of principal components"

If Hotelling's seminal 1933 article in *JEdP* had appeared instead in *Psychometrika*, it would be, according to Google Scholar, the second most highly cited article in *Psychometrika* after Cronbach's (1951) survey on "coefficient alpha." The first co-editor of *Psychometrika*, Paul Horst, even relates how he tried to get something comparable for the first volume of *Psychometrika* (Horst and Stalnaker, 1986, p. 5 [*Psychometrika*, *51*, "Present at the birth"]):

At Proctor & Gamble we had been working with the applications of the new factor analytic methods to personnel data. I had learned of a new iterative procedure that Hotelling at Columbia had developed for finding the principal axis factors of a correlation matrix, and we were using it at Proctor & Gamble. I saw Hotelling personally at Columbia during this time, to persuade him to contribute his manuscript for the maiden issue of *Psychometrika*. I asked him whether he could give us a manuscript on his new method. He at first was markedly cool to the idea and I suspected that he was not eager to conceal his production under the cover of a dubious new journal. I then told him that I very much wanted this method published in this first issue and that, if he did not feel he could do it, I would reluctantly publish the method myself and of course give him full credit. With this, he decided to provide the manuscript himself [Hotelling, H. (1936). Simplified calculation of principal components. *Psychometrika*, *1*, 27–35], and we remained good friends as long as he lived.

The 1936 Hotelling paper referenced above is based on the simple idea that when the power method is applied to an integer power of a matrix (say, to $\mathbf{A}^2$) instead of to the original matrix (say, to $\mathbf{A}$), convergence will be faster. Unfortunately, such a conjecture appears generally unjustified. We give two quotes from Bodewig (1950, p. 134; 246) that make this point:

Hotelling [in the 1943 *Annals of Mathematical Statistics* article] therefore, proposes computing the product $\mathbf{T}$ and, then to square successively: $\mathbf{T}$, $\mathbf{T}^2$, $\mathbf{T}^4$, $\mathbf{T}^8$ ..., and then to form the vector say $\mathbf{T}^{16}\mathbf{y}^{(1)}$. This method is very elegant. *Whether it is suitable, is another matter.* (emphasis added)

Powers of Matrices: Many authors such as Kincaid, Aitken, Hammersley, and Hotelling, recommend successive squaring of $\mathbf{A}$ and iteration with $\mathbf{A}^{2\mathrm{m}}$ on $\mathbf{v}$ instead of with $\mathbf{A}$ itself. This is done in order to speed up convergence and to save work. *But this proposal cannot be defended.* (emphasis added)

Bodewig provides a formal proof of this assertion that "this proposal cannot be defended." It is based on an elaboration of the following observation: multiplying a vector $\mathbf{x}$ by a matrix $\mathbf{A}$ and that resultant vector, $\mathbf{Ax}$, by $\mathbf{A}$ again (i.e., $\mathbf{A}(\mathbf{Ax})$), requires fewer operations than multiplying $\mathbf{A}$ by $\mathbf{A}$, and then using that product matrix, $\mathbf{A}^2$, to multiply $\mathbf{x}$ (i.e., $\mathbf{A}^2\mathbf{x}$).

Kelley's Approach to Principal Components

In the Holzinger survey of the work completed by the Unitary Traits Committee mentioned earlier, the following short excerpt appears:

> Very recently Professor Kelley has published a volume entitled *The Essential Traits of Mental Life* (1935). In this book he has contributed a method of factorization which appears simpler than that of Hotelling, but which gives the same results. In addition to this new technique Professor Kelley makes a comparison of current methods of factorization.

In the 1936 Hotelling paper solicited by Horst, Kelley's method of obtaining principal components is explicitly commented on as follows (p. 27):

> Another method of calculating principal components has been discovered by Professor Truman L. Kelley, which involves less labor than the original iterative method, at least in the examples to which he has applied it. How it would compare with the present accelerated method is not clear, except that some experience at Columbia University has suggested that the method here set forth is the more efficient. It is possible that Kelley's method is more suitable when all the characteristic roots are desired, but not the corresponding correlations of the variates with the components. The present method seems to the computers who have tried both to be superior when the components themselves, as well as their contributions to the total variance, are to be specified. The advantage of the present method is enhanced when, as will

often be the case in dealing with numerous variates, not all the characteristic roots but only a few of the largest are required.

A synopsis is given below of Kelley's method for finding the two principal components of a two-variable system, taken from his *Essential Traits of Mental Life*. He showed that by using this method iteratively for all pairs of variables, the complete set of principal components are retrieved:

If it is desired to create two new variables, x' and y', which are completely defined by the given variables, x and y, all that is necessary is to write $x' = a_1x + b_1y$; $y' = a_2x + b_2y$ and assign any values to $a_1, a_2, b_1,$ and b_2. Solving these equations for x and y we have

$$x = \frac{b_2x' - b_1y'}{a_1b_2 - a_2b_1}$$

$$y = \frac{a_1y' - a_2x'}{a_1b_2 - a_2b_1}$$

Of the infinite number of new sets of equivalent variables, x' and y', which can be derived by substituting different values for a_1, a_2, b_1, and b_2, that one is considered to have special merit which is a rotation of the x and y axes to the position of the major and minor axes of the ellipse. These particular new variables, which we designate x_1 and y_1, are given by the equations

$$x_1 = x\cos\theta + y\sin\theta$$

$$y_1 = -x\sin\theta + y\cos\theta$$

where θ is the angle of rotation and is given by

$$\tan 2\theta = \frac{2p}{v_1 - v_2}$$

$[p = \sigma_{12}; v_1 = \sigma_1^2; v_2 = \sigma_2^2]$ The peculiar merit of the new variables, x_1 and y_1, lies in the facts which can be immediately surmised by thinking of the elementary geometry involved.

(a) x_1 and y_1 are uncorrelated.

(b) x_1 and y_1 axes are at right angles to each other.

(c) The variance of x_1, distance from the minor axis in the direction of the major axis, is a maximum, for no other rotation of axes yields a variable with as large a variance.

(d) The variance of y_1, distance measured in the direction of the minor axis, is a minimum.

The advantage of (a), lack of correlation, need scarcely be dwelt upon, as it is the essential purpose of factorization to obtain independent measures.

The advantage of (b), orthogonality, is not quite so obvious. Though a point in two-dimensional space may be completely defined by distance from two oblique axes, nevertheless the simplicity of thought (and to create such simplicity is a basic purpose of factorization) when a point is defined in terms of perpendicular distance from two perpendicular axes, should be sufficient to commend the use of such axes.

The advantage of (c) making the variance of one of the new variables a maximum is particularly apparent when the major axis is much greater than the minor. In this case, much more about the total situation or the total field wherein variation can take place is known if variability in any other direction is known. The principle of parsimony of thought recommends a knowledge of the x_1 variable if but a single item of knowledge is available. The operation of this principle will be much more apparent when thinking of many variables, for here the variances of some of the smaller ones may be such that entire lack of knowledge of them will not be serious.

It is obvious from the geometry of the situation that there is but a single solution yielding variables with the properties mentioned. These constitute the components in the two-variable problem.

It is interesting to speculate where Kelley may have come up with his approach to the calculation of principal components. He gives no explicit reference for his iterative method in the *Essential Traits of Mental Life*; in fact, he opens this text (Chapter I) as follows:

A New Method of Analysis of Variables into Independent Components: Before attempting a comparison of different methods of analysis of variables into components, a new method is presented. The procedure followed is new, but the outcome is identical with that given by Hotelling's method of analysis.

One story that is at least plausible comes from a perusal of the Kelley archives at Harvard. Kelley spent a sabbatical year in the very early 1920s with Karl Pearson, who was to have a major influence on Kelley's statistical thinking. For example, in the preface to Kelley's well-received 1923 text, *Statistical Method*, there is the following acknowledgement to Karl Pearson:

I would, however, say that my greatest inspiration has been the product of that master analyst, Karl Pearson, and that the English school entire has been most contributive.

There is also a reference in *Statistical Method* (p. 363) to Karl Pearson's 1901 paper in the *Philosophical Magazine* (*2*, 559–572), "On lines and planes of closest fit to systems of points in space." As is now well-recognized, this early 1901 paper introduced "the method of principal components," although that particular terminology, introduced much later by Hotelling in 1933, was obviously not used.

The key "$\tan 2\theta$" formula in Kelley's method for finding the angle of rotation for the principal axes orientation of a two-variable system is present in Pearson (1901, p. 566). It is conceivable that Kelley could have encountered it there for the first time, but it is more likely that Kelley knew of it from his undergraduate work in mathematics at the University of Illinois in the early 1900s. Neither Pearson nor Kelley, for example, thought it necessary to include any reference for what was presumably a well-known formula in mechanics that dealt with the axes of an ellipsoid. At Illinois, Kelley did a Bachelor of Arts thesis (1909) entitled "Graphic evaluation of trigonometric functions of complex variables." (A Google search on this exact title will retrieve a copy of the thesis.) Kelley's trigonometric prowess as represented in his thesis is also well on display in his *Essential Traits of Mental Life* — an extensive set of trigonometric equations were derived by Kelley to make the iterative process work.

An interview done in 2006 with Darrell Bock in the *Journal of Educational and Behavioral Statistics* may shed some more historical light on the question of "Whence Principal Components?" The excerpts given below discuss Bock's visit to the University of Illinois in the 1950s to use the ILLIAC computer for some eigenvector/eigenvalue computations that he needed done. Note the name of the graduate student he met at Illinois, Gene Golub; Golub was soon to become a computational giant of the second half of the 20th century.

I had heard from Charles Wrigley at Michigan State University that the new ILLIAC electronic computer at Champaign-Urbana had programs for both the one- and two-matrix eigenproblems. On his advice, I phoned Kern Dickman, who had helped Charles perform a principal component analysis on the machine, and explained my needs. He invited me to come down to Urbana and bring the matrices to be analyzed with me. By that time, I had become sufficiently proficient in using punched card equipment in the business office of the University — in particular a new electronic calculating punch that could store constants and performed cumulative multiplications as fast as the cards passed through the machine.

I arrived in Urbana and found Kern; he took me directly to the computation Center to see the ILLIAC. But there was very little to see — only a photoelectric reader of teletype tape and a box with a small slit where punched tape spewed from the machine; a few dimly revealed electronic parts could be seen behind a plate-glass window. Elsewhere in the room were teletype machines for punching numbers and letters onto paper tape, printing out the characters of an existing tape, or copying all or parts of one tape to another. My first job was to key the elements of the two covariance matrices onto tape, which in spite of my best efforts to avoid errors, took most of the afternoon.

When I finished that task, Kern suggested that we should meet for dinner at his favorite watering hole in Urbana. When I arrived there I found him sitting with another person whom he introduced as Gene Golub, adding that Gene had programmed the eigenroutines for the ILLIAC. At Kern's suggestion Gene had brought along some papers for me — an introduction to programming the ILLIAC and the documentation of the eigenroutines. He said that his code was similar to that of Goldstein, who had programmed the eigen-procedures for the Maniac machine built by Metropolis at Los Alamos. It used the Jacobi iterative method, which consists of repeated orthogonal transformations of pairs of variables to reduce the elements in the off-diagonal of a real symmetric matrix to zero, all the while performing the same operation on an identity matrix. Although a given element of the matrix does not necessarily remain zero, the iterations converge to a diagonal matrix containing the eigenvalues, and the identity matrix becomes the corresponding eigenvectors.

Gene told the story that Goldstein, having heard the Jacobi method described by a colleague, stopped by John von Neumann's office to ask if the method was strictly convergent. Gazing at the ceiling for about five seconds, von Neumann replied "yes, of course." Goldstein was amazed, thinking this was another of von Neuman's fabled feats of mental calculation, but as Golub and Van Loan show in their 1996 reference, *Matrix Computations*, the proof requires only a few lines of matrix expressions, which von Neumann could have easily visualized. I already knew of this method, not as Jacobi's, but as the "method of sine and cosine transformations" described by Truman Kelley in his 1935 book, *Essential Traits of Mental Life*. He presented the method as his own creation, including a proof of convergence requiring several pages of geometric argument. Considering that Jacobi had introduced the method in the middle of the 19th-century, I wondered if Kelley had heard of it from one of his fellow professors at Harvard. But I found in his 1928 book, *Crossroads in the Mind of Man*, that he had already used sine and cosine transformations in connection with Spearman's one-factor model, and I now believe that he rediscovered Jacobi's method independently.

Bock gets this a little incorrect. Kelley did not "rediscover" Jacobi's method. He did not know, for example, that merely multiplying the pairwise orthogonal rotations together would give the eigenvectors directly as is done in Jacobi's method. But still, Kelley got very close by obtaining all of the eigenvalues of a correlation matrix at the end of his pairwise iterative process. Kelley generated the corresponding eigenvectors rather laboriously by keeping track of all the transformations carried out over the pairwise iterations as expressed in terms of the original variables.

Conclusion

So now to the opening question of "Whence principal components?" The best theoretical answer is probably Karl Pearson, given his 1901 paper mentioned earlier.[4] The numerical examples Pearson gave, however, were all extremely small and involved at most three variables. So, from a computational perspective, the answer to the question should probably be Hotelling, based upon his use of an iterative power method and the introduction of Hotelling deflations. If current computational practice is any criterion, however, Kelley could be credited with the introduction of a rudimentary Jacobi-like method. The Jacobi approach became more or less standard practice in the 1950s and 60s. As noted by Bock in the earlier excerpts, the method had been programmed by Golub for the ILLIAC computer before Bock's visit to Illinois. From the 1970s to the present, most

[4]Some other authors, however, go back much further. For example, Paul Horst in his 1992 *Chemometrics and Intelligent Laboratory Systems* paper, "Sixty years with latent variables and still more to come," traces principal components back to Cauchy, as early as 1829.

computer-implemented principal component computational routines (in Matlab, for instance) rely on a more basic singular value decomposition (SVD) algorithm developed by that same graduate student Bock met at Illinois in the 1950s, Gene Golub (see, for example, G. H. Golub and C. Reinsch, "Singular value decomposition and least squares solutions," *Numerische Mathematik*, *14*, 1970, 403–420). By way of closing, it is interesting to note that the Golub-Reinsch SVD routine relies on exactly the same type of planar rotations (but now called Givens rotations) used by Kelley in his approach to computing principal components.[5]

Lawrence Hubert, July 2016
lhubert@illinois.edu

[5]It might also be noted that Hotelling in his paper introducing canonical correlations ("The relations between two sets of variates," *Biometrika*, *28*, 1936, 321–377) relies on the same type of iterated power method for obtaining canonical correlations and canonical variates as he did in *JEdP* (1933). Kelley, in contrast, in his 1940 monograph, *Talents and Tasks: Their Conjunction in a Democracy for Wholesome Living and National Defense*, approached the canonical correlation task using planar rotations, just as he did in the *Essential Traits of Mental Life*. Also, Kelley provided a rather complete numerical example — and obviously, given the year of publication, all without any electronic computer implementation.