University of Maryland College Park

Institute for Advanced Computer Studies TR–92–31
Department of Computer Science TR–2855

On the Early History of the Singular Value Decomposition[*]

G. W. Stewart[†]

March 1992

ABSTRACT

This paper surveys the contributions of five mathematicians — Eugenio Beltrami (1835–1899), Camille Jordan (1838–1921), James Joseph Sylvester (1814–1897), Erhard Schmidt (1876–1959), and Hermann Weyl (1885–1955) — who were responsible for establishing the existence of the singular value decomposition and developing its theory.

[*] This report is available by anonymous ftp at `thales.cs.umd.edu` in the directory `pub/reports`.

[†] Department of Computer Science and Institute for Advanced Computer Studies, University of Maryland, College Park, MD 20742. This work was done during while the author was visiting the Institute for Mathematics and Its Applications, University of Minnesota, Minneapolis, MN 55455.

On the Early History of the Singular Value Decomposition

G. W. Stewart

For Gene Golub on his fifteenth birthday

1. Introduction

One of the most fruitful ideas in the theory of matrices is that of a matrix decomposition or canonical form. The theoretical utility of matrix decompositions has long been appreciated. More recently, they have become the mainstay of numerical linear algebra, where they serve as computational platforms from which a variety of problems can be solved.

Of the many useful decompositions, the singular value decomposition — that is, the factorization of a matrix $\mathbf{A}$ into the product $\mathbf{U\Sigma V}^{\mathrm{H}}$ of a unitary matrix $\mathbf{U}$ a diagonal matrix $\mathbf{\Sigma}$ and another unitary matrix $\mathbf{V}^{\mathrm{H}}$ — has assumed a special role. There are several reasons. In the first place, the fact that the decomposition is achieved by unitary matrices makes it an ideal vehicle for discussing the geometry of n-space. Second, it is stable; small perturbations in $\mathbf{A}$ correspond to small perturbations in $\mathbf{\Sigma}$, and conversely. Third, the diagonality of $\mathbf{\Sigma}$ makes it easy to determine when $\mathbf{A}$ is near to a rank-degenerate matrix; and when it is, the decomposition provides optimal low rank approximations to $\mathbf{A}$. Finally, thanks to the pioneering efforts of Gene Golub, there exist efficient, stable algorithms to compute the singular value decomposition.

The purpose of this paper is to survey the contributions of five mathematicians — Eugenio Beltrami (1835–1899), Camille Jordan (1838–1921), James Joseph Sylvester (1814–1897), Erhard Schmidt (1876–1959), and Hermann Weyl (1885–1955) — who were responsible for establishing the existence of the singular value decomposition and developing its theory. Beltrami, Jordan, and Sylvester came to the decomposition through what we should now call linear algebra; Schmidt and Weyl approached it from integral equations. To give this survey context, we will begin with with a brief description of the historical background.

Is is an intriguing observation that most of the classical matrix decompositions predated the widespread use of matrices: they were cast in terms of determinants, linear systems of equations, and especially bilinear and quadratic forms. Gauss is the father of this development. Writing in 1823 [15, §31], he describes his famous elimination algorithm as follows.

> Specifically, the function Ω [a quadratic function of x, y, z, etc.] can be reduced to the form
>
> $$\tfrac{u^0 u^0}{\mathcal{A}^0} + \tfrac{u' u'}{\mathcal{B}'} + \tfrac{u'' u''}{\mathcal{C}''} + \tfrac{u''' u'''}{\mathcal{D}'''} + \text{etc.} + M,$$
>
> in which the divisors $\mathcal{A}^0$, $\mathcal{B}'$, $\mathcal{C}''$, $\mathcal{C}'''$, etc. are constants and u^0, u', u'', u''', etc. are linear functions of x, y, z, etc. However, the second function, u', is independent of x; the third, u'', is independent of x and y; the fourth, u''' is independent of x, y, and z, and so on. The last function $u^{(\pi-1)}$ depends only on the the last of the unknowns x, y, z, etc. Moreover, the coefficients $\mathcal{A}^0$, $\mathcal{B}'$, $\mathcal{C}''$, etc. multiply x, y, z, etc. in u^0, u', u'', etc. respectively.

From this we easily see that Gauss's algorithm factors the matrix of the quadratic form $\mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{x}$ into the product $\mathbf{R}\mathbf{D}^{-1}\mathbf{R}$, where $\mathbf{D}$ is diagonal and $\mathbf{R}$ is upper triangular with the diagonals of $\mathbf{D}$ on its diagonal. Gauss's functions u^0, u', u'', etc. are the components of the vector $\mathbf{u} = \mathbf{R}\mathbf{x}$.

Gauss was also able to effectively obtain the inverse of a matrix by a process of *eliminatio indefinita,* in which the system of equations $\mathbf{y} = \mathbf{A}\mathbf{x}$ is transformed into the inverse system $\mathbf{x} = \mathbf{B}\mathbf{y}$. Gauss's skill in manipulating quadratic forms and systems of equations made possible his very general treatment of the theory and practice of least squares.

Other developments followed. Cauchy [6, 1829] established the properties of the eigenvalues and eigenvectors of a symmetric system (including the interlacing property) by considering the corresponding homogeneous system of equations. In 1846, Jacobi [25] gave his famous algorithm for diagonalizing a symmetric matrix, and in a posthumous paper [26, 1857] he obtained the LU decomposition by decomposing a bilinear form in the style of Gauss. Weierstrass [50, 1868] established canonical forms for pairs of bilinear functions — what we should today call the generalized eigenvalue problem. Thus the advent of the singular value decomposition in 1873 is seen as one of a long line of results on canonical forms.

We will use modern matrix notation to describe the early work on the singular value decomposition. Most of it slips as easily into matrix terminology as Gauss's description of his decomposition; and we shall be in no danger of anachronism, provided we take care to use matrix notation only as an expository device, and otherwise stick close to the writer's argument. The greatest danger is that the use of modern notation will trivialize the writer's accomplishments by making them obvious to our eyes. On the other hand, presenting them in the original scalar form would probably exaggerate the obstacles these people had to overcome, since they were accustomed, as we are not, to grasping sets of equations as a whole,

With a single author, it is usually possible to modernize notation in such a way that it corresponds naturally to what he actually wrote. Here we are dealing with several authors, and uniformity is more important than correspondence with the original. Consequently, throughout paper we will be concerned with the singular value decomposition

$$\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^{\mathrm{T}},$$

where $\mathbf{A}$ is a real matrix of order n,

$$\mathbf{\Sigma} = \mathrm{diag}(\sigma_1, \sigma_2, \ldots, \sigma_n)$$

has nonnegative diagonal elements arranged in descending order of magnitude, and

$$\mathbf{U} = (\mathbf{u}_1\ \mathbf{u}_2\ \ldots\ \mathbf{u}_n) \quad \text{and} \quad \mathbf{V} = (\mathbf{v}_1\ \mathbf{v}_2\ \ldots\ \mathbf{v}_n)$$

are orthogonal. The function $\|\cdot\|$ will denote the Frobenius norm defined by

$$\|\mathbf{A}\|^2 = \sum_{i,j} a_{ij}^2 = \sum_i \sigma_i^2.$$

In summarizing the contributions I have followed the principle that if you try to say everything you end up saying nothing. Most of the works treated here are richer than the following sketches would indicate, and the reader is advised to go to the sources for the full story.

2. Beltrami [5, 1873]

Together Beltrami and Jordan are the progenitors of the singular value decomposition, Beltrami by virtue of first publication and Jordan by the completeness and elegance of his treatment. Beltrami's contribution appeared in the *Journal of Mathematics for the Use of the Students of the Italian Universities,* and its purpose was to encourage students to become familiar with bilinear forms.

The Derivation. Beltrami begins with a bilinear form

$$f(\mathbf{x}, \mathbf{y}) = \mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{y},$$

where $\mathbf{A}$ is real and of order n. If one makes the substitutions

$$\mathbf{x} = \mathbf{U}\boldsymbol{\xi} \quad \text{and} \quad \mathbf{y} = \mathbf{V}\boldsymbol{\eta},$$

then

$$f(\mathbf{x}, \mathbf{y}) = \boldsymbol{\xi}^{\mathrm{T}} \mathbf{S} \boldsymbol{\eta},$$

where

$$\mathbf{S} = \mathbf{U}^{\mathrm{T}} \mathbf{A} \mathbf{V}. \tag{2.1}$$

Beltrami now observes that if $\mathbf{U}$ and $\mathbf{V}$ are required to be orthogonal then there are $n^2 - n$ degrees of freedom in their choice, and he proposes to use these degrees of freedom to annihilate the off diagonal element of $\mathbf{S}$.

Assume that $\mathbf{S}$ is diagonal; i.e. $\mathbf{S} = \boldsymbol{\Sigma} = \mathrm{diag}(\sigma_1, \ldots, \sigma_n)$. Then it follows from (2.1) and the orthogonality of $\mathbf{V}$ that

$$\mathbf{U}^{\mathrm{T}} \mathbf{A} = \boldsymbol{\Sigma} \mathbf{V}^{\mathrm{T}}. \tag{2.2}$$

Similarly

$$\mathbf{A} \mathbf{V} = \mathbf{U} \boldsymbol{\Sigma}. \tag{2.3}$$

Substituting the value of $\mathbf{U}$ obtained from (2.3) into (2.2), Beltrami obtains the equation

$$\mathbf{U}^{\mathrm{T}} (\mathbf{A} \mathbf{A}^{\mathrm{T}}) = \boldsymbol{\Sigma}^2 \mathbf{U}^{\mathrm{T}}, \tag{2.4}$$

and similarly he obtains

$$(\mathbf{A}^{\mathrm{T}} \mathbf{A}) \mathbf{V} = \mathbf{V} \boldsymbol{\Sigma}^2.$$

Thus the σ_i are the roots of the equations

$$\det(\mathbf{A} \mathbf{A}^{\mathrm{T}} - \sigma^2 I) = 0 \tag{2.5}$$

and

$$\det(\mathbf{A}^{\mathrm{T}} \mathbf{A} - \sigma^2 I) = 0. \tag{2.6}$$

Note that the derivation, as presented by Beltrami, assumes that $\boldsymbol{\Sigma}$, and hence $\mathbf{A}$, is nonsingular.[1]

Beltrami now argues that the two functions (2.5) and (2.6) are identical because they are polynomials of degree n that assume the same values at $\sigma = \sigma_i$ $(i = 1, \ldots, n)$ and the common value $\det^2(A)$ at $\sigma = 0$, an argument that presupposes that the singular values are distinct and nonzero.

[1] However, it is possible to derive the equations without assuming that $\mathbf{A}$ is nonsingular; e.g., $\mathbf{U}^{\mathrm{T}} \mathbf{A} \mathbf{A}^{\mathrm{T}} = \boldsymbol{\Sigma} \mathbf{V}^{\mathrm{T}} \mathbf{A}^{\mathrm{T}} = \boldsymbol{\Sigma}^2 \mathbf{U}^{\mathrm{T}}$, the first equality following on multiplying (2.2) by $\mathbf{A}^{\mathrm{T}}$, and the second on substituting the transpose of (2.3). Thanks to Ann Greenbaum for pointing this fact out.

Beltrami next states that by a well-known theorem, the roots of (2.5) are real. Moreover, they are positive. To show this he notes that

$$0 < \|\mathbf{x}^{\mathrm{T}}\mathbf{A}\|^2 = \mathbf{x}^{\mathrm{T}}(\mathbf{A}\mathbf{A}^{\mathrm{T}})\mathbf{x} = \boldsymbol{\xi}^{\mathrm{T}}\boldsymbol{\Sigma}^2\boldsymbol{\xi}, \tag{2.7}$$

the last equation following from the theory of quadratic forms. This inequality immediately implies that the σ_i^2 are positive.

There is some confusion here. Beltrami appears to be assuming the existence of the vector $\boldsymbol{\xi}$, whose very existence he is trying to establish. The vector required by his argument is an eigenvector of $\mathbf{A}\mathbf{A}^{\mathrm{T}}$ corresponding to σ. The fact that the two vectors turn out to be the same apparently caused Beltrami to leap ahead of himself and use $\boldsymbol{\xi}$ in (2.7).

Beltrami is now ready to give an algorithm to determine the diagonalizing transformation.

1. Find the roots of the equation (2.5).
2. Determine $\mathbf{U}$ from (2.4). Here Beltrami notes that the columns of $\mathbf{U}$ are determined up to factors of ± 1, which is true only if the σ_i are distinct. He also tacitly assumes that the resulting $\mathbf{U}$ will be orthogonal, which also requires that the σ_i be distinct.
3. Determine $\mathbf{V}$ from (2.2). This step requires that $\boldsymbol{\Sigma}$ be nonsingular.

Discussion. From the foregoing it is clear that Beltrami derived the singular value decomposition for a real, square, nonsingular matrix having distinct singular values. His derivation is the one given in most textbooks, but it lacks the extras needed to handle degeneracies. It may be that in omitting these extras Beltrami was simplifying things for his student audience, but a certain slackness in the exposition suggests that he had not thought the problem through.

3. Jordan [28, 29, 1874]

Camille Jordan can fairly be called the codiscoverer of the singular value decomposition. Although he published his derivation a year after Beltrami, it is clear that the work is independent. In fact, the “Mémoire sur les formes bilinéaires” treats three problems, of which the the reduction of a bilinear form to a diagonal form by orthogonal substitutions is the simplest.[2]

The Derivation. Jordan starts with the form

$$P = \mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{y}$$

and seeks the maximum and minimum of P subject to

$$\|\mathbf{x}\|^2 = \|\mathbf{y}\|^2 = 1. \tag{3.1}$$

The maximum is determined by the equation

$$0 = dP = d\mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{y} + \mathbf{x}^{\mathrm{T}}\mathbf{A}d\mathbf{y}, \tag{3.2}$$

which must be satisfied for all $d\mathbf{x}$ and $d\mathbf{y}$ that satisfy

$$d\mathbf{x}^{\mathrm{T}}\mathbf{x} = 0 \quad \text{and} \quad d\mathbf{y}^{\mathrm{T}}\mathbf{y} = 0. \tag{3.3}$$

3 Jordan then asserts that "equation (3.2) will therefore be a combination of the equations (3.3)," from which one obtains[3]

$$\mathbf{A}\mathbf{y} = \sigma\mathbf{x} \tag{3.4}$$

and

$$\mathbf{x}^{\mathrm{T}}\mathbf{A} = \tau\mathbf{y}^{\mathrm{T}}. \tag{3.5}$$

From (3.4) it follows that the maximum is

$$\mathbf{x}^{\mathrm{T}}(\mathbf{A}\mathbf{y}) = \sigma\mathbf{x}^{\mathrm{T}}\mathbf{x} = \sigma.$$

Similarly the maximum is also τ, so that $\sigma = \tau$.

Jordan now observes that σ is determined by the vanishing of the determinant

$$D = \begin{vmatrix} -\sigma\mathbf{I} & \mathbf{A} \\ \mathbf{A}^{\mathrm{T}} & -\sigma\mathbf{I} \end{vmatrix}$$

[2] The other two are to reduce a form by the same substitution of both sets of variables and to reduce a pair of forms by two substitutions, one for each set of variables. Jordan notes that the former problem had been considered by Kronecker [31, 1866] in a different form, and the latter by Weierstrass [50, 1868].

[3] Jordan's argument is not very clear. Possibly he means to say that for some constants σ and τ we must have $d\mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{y} + \mathbf{x}^{\mathrm{T}}\mathbf{A}d\mathbf{y} = \sigma d\mathbf{x}^{\mathrm{T}}\mathbf{x} + \tau d\mathbf{y}^{\mathrm{T}}\mathbf{y}$, from which the subsequent equations follow from the independence of $d\mathbf{x}$ and $d\mathbf{y}$.

of the system (3.4)–(3.5). He shows that this determinant contains only even powers of σ.

Now let σ_1 be a root of the equation $D = 0$, and let the equations (3.4) and (3.5) be satisfied by $\mathbf{x} = \mathbf{u}$ and $\mathbf{y} = \mathbf{v}$, where $\|\mathbf{u}\|^2 = \|\mathbf{v}\|^2 = 1$. (Jordan notes that one can find such a solution, even when it is not unique.) Let

$$\hat{\mathbf{U}} = (\mathbf{u}\ \mathbf{U}_*) \quad \text{and} \quad \hat{\mathbf{V}} = (\mathbf{v}\ \mathbf{V}_*)$$

be orthogonal, and let

$$\mathbf{x} = \hat{\mathbf{U}}\hat{\mathbf{x}} \quad \text{and} \quad \mathbf{y} = \hat{\mathbf{V}}\hat{\mathbf{y}}.$$

With these substitutions, let

$$P = \hat{\mathbf{x}}^{\mathrm{T}}\hat{\mathbf{A}}\hat{\mathbf{y}}.$$

In this system, P attains its maximum[4] for $\hat{\mathbf{x}} = \hat{\mathbf{y}} = \mathbf{e}_1$, where $\mathbf{e}_1 = (1, 0, \ldots, 0)^{\mathrm{T}}$. Moreover, at the maximum we have

$$\hat{\mathbf{A}}\hat{\mathbf{y}} = \sigma_1\hat{\mathbf{x}} \quad \text{and} \quad \hat{\mathbf{x}}^{\mathrm{T}}\hat{\mathbf{A}} = \sigma_1\hat{\mathbf{y}}^{\mathrm{T}},$$

which implies that

$$\hat{\mathbf{A}} = \begin{pmatrix} \sigma & \mathbf{0} \\ \mathbf{0} & \mathbf{A}_1 \end{pmatrix}.$$

Thus with $\xi_1 = \hat{x}_1$ and $\eta_1 = \hat{y}_1$, P assumes the form

$$\sigma_1\xi_1\eta_1 + P_1,$$

where P_1 is independent of ξ_1 and η_1. Jordan now applies the reduction inductively to P_1 to arrive at the canonical form

$$P = \boldsymbol{\xi}^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{\eta}.$$

Finally, Jordan notes that when the roots of the characteristic equation $D = 0$ are simple, the columns of $\mathbf{U}$ and $\mathbf{V}$ can be calculated directly from (3.1), (3.4), and (3.5).

Discussion. In this paper we see the sure hand of a skilled professional. Jordan proceeds from problem to solution with economy and elegance. His approach of using a partial solution of the problem to reduce it to one of smaller size — *deflation* is the modern term — avoids the degeneracies that complicate Beltrami's

[4] Jordan nods here, since he has not explicitly selected the largest root σ_1.

approach. Incidentally, the technique of deflation apparently lay fallow until Schur [41, 1917] used it to establish his triangular form of a general matrix. It is now a widely used theoretical and algorithmic tool.

The matrix

$$\begin{pmatrix} \mathbf{0} & \mathbf{A} \\ \mathbf{A}^{\mathrm{T}} & \mathbf{0} \end{pmatrix},$$

from which the determinant D was formed, is also widely used. Its present day popularity is due to Wielandt (see [14, p.113]) and Lanczos [32, 1958], who apparently rediscovered the decomposition independently.

Yet another consequence of Jordan's approach is the variational characterization of the largest singular value as the maximum of a function. This and related characterizations have played an important role in perturbation and localization theorems for singular values (for more see [43, §IV.4]).

4. Sylvester [44, 46, 45, 1889]

Sylvester wrote a footnote and two papers on the subject of the singular value decomposition. The footnote appears at the end of a paper in *The Messenger of Mathematics* [44] entitled "A New Proof That a General Quadric May Be Reduced to Its Canonical Form (That Is, a Linear Function of Squares) by Means of a Real Orthogonal Substitution." In the paper Sylvester describes an iterative algorithm for reducing a quadratic form to diagonal form. In the footnote he points out that an analogous iteration can be used to diagonalize a bilinear form and says that he has "sent for insertion in the C. R. of the Institute a Note in which I give the rule for effecting this reduction." The rule turns out to be Beltrami's algorithm. In a final *Messenger* paper [45], Sylvester presents both the iterative algorithm and the rule.

The Rule. Here we follow the *Messenger* paper. Sylvester begins with the bilinear form

$$B = \mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{y}$$

and considers the quadratic form

$$M = \sum_i \left(\frac{dB}{dy_i}\right)^2$$

(which is $\mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{A}^{\mathrm{T}}\mathbf{x}$, a fact tacitly assumed by Sylvester). Let $M = \sum \lambda_i \boldsymbol{\xi}_i^2$ be the canonical form of M. If B has the canonical form $B = \sum \sigma_i \boldsymbol{\xi}_i \boldsymbol{\eta}_i$, then $\sum[\sigma_i \boldsymbol{\xi}]^2$ is orthogonally equivalent to $M = \sum \lambda_i \boldsymbol{\xi}_i^2$, which implies that $\lambda_i = \sigma_i^2$ in some order.

To find the substitutions, Sylvester introduces the matrices $\mathbf{M} = \mathbf{A}\mathbf{A}^{\mathrm{T}}$ and $\mathbf{N} = \mathbf{A}^{\mathrm{T}}\mathbf{A}$ and asserts that the substitution for $\mathbf{x}$ is the substitution that diagonalizes $\mathbf{M}$ and substitution for $\mathbf{y}$ is the one that diagonalizes $\mathbf{N}$. In general, this is true only if the singular values of $\mathbf{A}$ are distinct.

In his *Comptes Rendu* note Sylvester gives the following rule for finding the coefficients of the $\mathbf{x}$-substitution corresponding to a singular value σ. Strike a row of the matrix $\mathbf{M} - \sigma^2\mathbf{I}$. Then the vector of coefficients is the vector of minors of order $n-1$ of the reduced matrix normalized so that their sum of squares is one. Coefficients of the $\mathbf{y}$-substitution may be obtained analogously from $\mathbf{N} - \sigma\mathbf{I}$. This only works if the singular value σ is simple.

Infinitesimal iteration. Sylvester first proposed this method as a technique for showing that a quadratic form could be diagonalized, and he later extended it to bilinear forms. It is already intricate enough for quadratic forms, and we will confine ourselves to a sketch of that case.

Sylvester proceeds inductively, assuming that he can solve a problem of order $n-1$. Thus for $n = 3$ he can assume the matrix is of the form

$$\mathbf{A} = \begin{pmatrix} a & 0 & f \\ 0 & b & g \\ f & g & c \end{pmatrix},$$

the zeros being introduced by the induction step. His problem is then to get rid of f and g without destroying the zeros previously introduced.

Sylvester proposes to make an "infinitesimal orthogonal substitution" of the form

$$\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 1 & \epsilon & \eta \\ -\epsilon & 1 & \theta \\ -\eta & -\theta & 1 \end{pmatrix} \begin{pmatrix} \xi_1 \\ \xi_2 \\ \xi_3 \end{pmatrix},$$

where the off diagonal quantities are so small that powers higher than the first can be neglected. Then the the $(2,1)$- and $(1,2)$-elements of the transformed matrix are

$$(a-b)\epsilon - f\theta - g\eta, \tag{4.1}$$

while the change in $f^2 + g^2$ is given by

$$\frac{1}{2}\delta(f^2 + g^2) = (a-c)f\eta + (b-c)g\theta.$$

If either of $(a-c)f$ or $(b-c)g$ is nonzero, η and θ can be chosen to decrease f^2+g^2. If $(a-b)$ is nonzero, ϵ may then be chosen so that (4.1) is zero; i.e., so that the zero previously introduced is preserved. Sylvester shows how special cases like $a=b$ can be handled by explicitly deflating the problem.

Sylvester now claims that an infinite sequence of these infinitesimal transformations will reduce one of f or g to zero, or will reduce the problem to one of the special cases.

Discussion. These are not easy papers to read. The style is opaque, and Sylvester pontificates without proving, leaving too many details to the reader. The mathematical reasoning harks back to an earlier, less rigorous era.

The fact that Sylvester sent a note to *Comptes Rendu*, the very organ where Jordan announced his results a decade and a half earlier, makes it clear that he was working in ignorance of his predecessors. It also suggests the importance he attached to his discovery, since a note in *Comptes Rendu* was tantamount to laying claim to a new result.

Sylvester was also working in ignorance of the iterative algorithm of Jacobi [25, 1846] for diagonalizing a quadratic form. The generalization of this algorithm to the singular value decomposition is due to Kogbetliantz [30].

It is not clear whether Sylvester intended to ignore second order terms in his iteration or whether he regards the diagonalization as being composed of an (uncountably) infinite number of infinitesimal transformation. Though the preponderance of his statements favor the latter, neither interpretation truly squares with everything he writes. In the first, small, but finite, terms replace the zeros previously introduces, so that a true diagonalization is not achieved. The second has the flavor of some recent algorithms in which discrete transformations are replaced by continuous transformations defined by differential equations (for applications of this approach to the singular value decomposition see [7, 9]). But Sylvester does not give enough detail to write down such equations.

5. Schmidt [39, 1907]

Our story now moves from the domain of linear algebra to integral equations, one of the hot topics of the first decades of our century. In his treatment of integral equations with unsymmetric kernels, Erhard Schmidt (of Gram–Schmidt fame) introduced the infinite dimensional analogue of the singular value decomposition. But he went beyond the mere existence of the decomposition by showing how it can be used to obtain optimal, low-rank approximations to an operator. In doing so he transformed the singular value decomposition from a mathematical curiosity to an important theoretical and computational tool.

Symmetric Kernels. Schmidt's approach is essentially the same as Beltrami's; however, because he worked in infinite dimensional spaces of functions he could not appeal to previous results on quadratic forms. Consequently, the first part of his paper is devoted to symmetric kernels.

Schmidt begins with a kernel $A(s,t)$ that is continuous and symmetric on $[a,b]\times[a,b]$. A continuous, nonvanishing function $\varphi(s)$ satisfying

$$\varphi(s) = \lambda \int_a^b A(s,t)\varphi(t)\,dt$$

is said to be an eigenfunction of A corresponding to the eigenvalue λ. Note that Schmidt's eigenvalues are the reciprocals of ours.

Schmidt then establishes the following facts.

1. The kernel A has at least one eigenfunction.
2. The eigenvalues and their eigenfunctions are real.
3. Each eigenvalue of A has at most a finite number of linearly independent eigenfunctions.
4. The kernel A has a complete, orthonormal system of eigenfunctions; that is, a sequence $\varphi_1(s)$, $\varphi_2(s)$, ... of orthonormal eigenfunctions such that every eigenfunction can be expressed as a linear combination of a finite number of the $\varphi_j(s)$.[5]
5. The eigenvalues satisfy

   $$\int_a^b\int_a^b \bigl(A(s,t)\bigr)^2\,ds\,dt \ge \sum_i \frac{1}{\lambda_i^2},$$

   which implies that the sequence of eigenvalues is unbounded.

Unsymmetric Kernels. Schmidt now allows $A(s,t)$ to be unsymmetric and calls any nonzero pair $u(s)$ and $v(s)$ satisfying

$$u(s) = \lambda \int_a^b A(s,t)v(t)\,dt$$

[5] This usage of the word "complete" is at variance with today's usage, in which a sequence is complete if its finite linear combinations are dense.

and

$$v(t) = \lambda \int_a^b A(s,t)u(s)\,ds$$

a pair of adjoint eigenfunctions corresponding to the eigenvalue λ.[6] He then introduces the symmetric kernels

$$\bar{A}(s,t) = \int_a^b A(s,r)A(t,r)\,dr$$

and

$$\underline{A}(s,t) = \int_a^b A(r,s)A(r,t)\,dr$$

and shows that if $u_1(s)$, $u_2(s)$, $\ldots$ is a complete orthonormal system for $\bar{A}(s,t)$ corresponding to the eigenvalues λ_1^2, λ_2^2, $\ldots$ then the sequence defined by

$$v_i(t) = \lambda_i \int_a^b A(s,t)u(s)\,ds, \qquad i = 1, 2, \ldots$$

is a complete orthonormal system for $\underline{A}(s,t)$. Moreover, for $i = 1, 2, \ldots$ the functions $u_i(s)$ and $v_i(s)$ form an adjoint pair for $A(s,t)$.

Schmidt then goes on to consider the expansion of functions in series of eigenfunctions. Specifically, if

$$g(s) = \int_a^b A(s,t)h(t)\,dt,$$

then

$$g(s) = \sum_i \frac{u_i(s)}{\lambda_i} \int_a^b h(t)v_i(t)\,dt,$$

and the convergence is absolute and uniform. Finally, he shows that if g and h are continuous then

$$\int_a^b \int_a^b A(s,t)g(s)h(t)\,ds\,dt = \sum_i \frac{1}{\lambda_i} \int_a^b g(s)u_i(s)\,ds \int_a^b h(t)v_i(t)\,dt, \tag{5.1}$$

an expression which Schmidt says “corresponds to the canonical decomposition of a bilinear form.”

[6] Again the usage differs from ours, but now in two ways. We work with the reciprocal of λ, calling it a singular value, and we distinguish between the singular values of a matrix and its eigenvalues.

The Approximation Theorem. Up to now, our exposition has been cast in the language of integral equations, principally to keep issues of analysis in the foreground. These issues are not as important in what follows, and we will therefore return to matrix notation, taking care, as always, to follow Schmidt's development closely.

The problem Schmidt sets out to solve is that of finding the best approximation to $\mathbf{A}$ of the form

$$\mathbf{A} \cong \sum_{i=1}^{k} \mathbf{x}_i \mathbf{y}_i^{\mathrm{T}}$$

in the sense that

$$\|\mathbf{A} - \sum_{i=1}^{k} \mathbf{x}_i \mathbf{y}_i^{\mathrm{T}}\| = \min .$$

In other words, he is looking for the best approximation of rank not greater than k.

Schmidt begins by noting that if

$$\mathbf{A}_k = \sum_{i=1}^{k} \sigma_i \mathbf{u}_i \mathbf{v}_i^{\mathrm{T}}, \tag{5.2}$$

then

$$\|\mathbf{A} - \mathbf{A}_k\|^2 = \|\mathbf{A}\|^2 - \sum_{i=1}^{k} \sigma_i^2 .$$

Consequently, if it can be shown that for arbitrary $\mathbf{x}_i$ and $\mathbf{y}_i$

$$\|\mathbf{A} - \sum_{i=1}^{k} \mathbf{x}_i \mathbf{y}_i^{\mathrm{T}}\| \geq \|\mathbf{A}\|^2 - \sum_{i=1}^{k} \sigma_i^2, \tag{5.3}$$

then $\mathbf{A}_k$ will be the desired approximation.

Without loss of generality we may assume that the vectors $\mathbf{x}_1, \ldots, \mathbf{x}_k$ are orthonormal. For if they are not, we can use Gram–Schmidt orthogonalization to express them as linear combinations of orthonormal vectors, substitute these expressions in $\sum_{i=1}^{k} \mathbf{x}_i \mathbf{y}_i^{\mathrm{T}}$, and collect terms in the new vectors.

Now

$$\begin{aligned}\|\mathbf{A} - \sum_{i=1}^{k} \mathbf{x}_i \mathbf{y}_i^{\mathrm{T}}\| &= \operatorname{trace}\Big((\mathbf{A} - \sum_{i=1}^{k} \mathbf{x}_i \mathbf{y}_i^{\mathrm{T}})^{\mathrm{T}}(\mathbf{A} - \sum_{i=1}^{k} \mathbf{x}_i \mathbf{y}_i^{\mathrm{T}})\Big) \\ &= \operatorname{trace}\Big(\mathbf{A}^{\mathrm{T}}\mathbf{A} + \sum_{i=1}^{k} (\mathbf{y}_i - \mathbf{A}^{\mathrm{T}}\mathbf{x}_i)(\mathbf{y}_i - \mathbf{A}^{\mathrm{T}}\mathbf{x}_i)^{\mathrm{T}} - \sum_{i=1}^{k} \mathbf{A}^{\mathrm{T}}\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\mathbf{A}\Big)\end{aligned}$$

Since $\mathrm{trace}\Big((\mathbf{y}_i - \mathbf{A}^{\mathrm{T}}\mathbf{x}_i)(\mathbf{y}_i - \mathbf{A}^{\mathrm{T}}\mathbf{x}_i)^{\mathrm{T}}\Big) \geq 0$ and $\mathrm{trace}(\mathbf{A}\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}) = \|\mathbf{A}\mathbf{x}_i\|^2$, the result will be established if it can be shown that

$$\sum_{i=1}^{k} \|\mathbf{A}\mathbf{x}_i\|^2 \leq \sum_{i=1}^{k} \sigma_i^2.$$

Let $\mathbf{V} = (\mathbf{V}_1\ \mathbf{V}_2)$, where $\mathbf{V}_1$ has k columns, and let $\mathbf{\Sigma} = \mathrm{diag}(\mathbf{\Sigma}_1, \mathbf{\Sigma}_2)$ be a conformal partition of $\mathbf{\Sigma}$. Then

$$\begin{aligned}\|\mathbf{A}\mathbf{x}_i\|^2 = \sigma_k^2 &+ \Big(\|\mathbf{\Sigma}_1\mathbf{V}_1^{\mathrm{T}}\mathbf{x}_i\|^2 - \sigma_k^2\|\mathbf{V}_1^{\mathrm{T}}\mathbf{x}_i\|^2\Big)\\ &-\Big(\sigma_k^2\|\mathbf{V}_2^{\mathrm{T}}\mathbf{x}_i\|^2 - \|\mathbf{\Sigma}_2\mathbf{V}_2^{\mathrm{T}}\mathbf{x}_i\|^2\Big)\\ &-\sigma_k^2\Big(1 - \|\mathbf{V}^{\mathrm{T}}\mathbf{x}_i\|\Big)\end{aligned} \tag{5.4}$$

Now the last two terms in (5.4) are clearly nonnegative. Hence

$$\begin{aligned}\sum_{i=1}^{k}\|\mathbf{A}\mathbf{x}_i\|^2 &\leq k\sigma_k^2 + \sum_{i=1}^{k}\Big(\|\mathbf{\Sigma}_1\mathbf{V}_1^{\mathrm{T}}\mathbf{x}_i\|^2 - \sigma_k^2\|\mathbf{V}_1^{\mathrm{T}}\mathbf{x}_i\|^2\Big)\\ &= k\sigma_k^2 + \sum_{i=1}^{k}\sum_{j=1}^{k}(\sigma_j^2 - \sigma_k^2)|\mathbf{v}_j^{\mathrm{T}}\mathbf{x}_i|^2\\ &= \sum_{j=1}^{k}\Big(\sigma_k^2 + (\sigma_j^2 - \sigma_k^2)\sum_{i=1}^{k}|\mathbf{v}_j^{\mathrm{T}}\mathbf{x}_i|^2\Big)\\ &\leq \sum_{j=1}^{k}\Big(\sigma_k^2 + (\sigma_j^2 - \sigma_k^2)\Big)\\ &= \sum_{j=1}^{k}\sigma_j^2,\end{aligned}$$

which establishes the result.

Discussion. Schmidt's two contributions to the singular value decomposition are its generalization to function spaces and his approximation theorem. Although Schmidt did not refer to earlier work on the decomposition in finite dimensional spaces, the quote following (5.1) suggests that he knew of its existence. Nontheless, his contribution here is substantial, especially since he had to deal with many of the problems of functional analysis without modern tools.

An important difference in Schmidt's version of the decomposition is the treatment of null-vectors of $\mathbf{A}$. In his predecessors' treatments they are part of the substitution that reduces the bilinear form $\mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{y}$ to its canonical form. For Schmidt they are not part of the decomposition. The effect of this can be seen in the third term of (5.4), which in the usual approach is zero but in Schmidt's approach can be nonzero.

The crowning glory of Schmidt's work is his approximation theorem, which is nontrivial to conjecture and hard to prove from scratch. Schmidt's proof is certainly not pretty — we will examine the more elegant approach of Weyl in the next section — but it does establish what can properly be termed the fundamental theorem of the singular value decomposition.

6. **Weyl** [51, 1912]

An important application of the approximation theorem is the determination of the rank of a matrix in the presence of error. If $\mathbf{A}$ is of rank k and $\tilde{\mathbf{A}} = \mathbf{A} + \mathbf{E}$, then the last $n - k$ singular values of $\tilde{\mathbf{A}}$ satisfy

$$\tilde{\sigma}_{k+1}^2 + \cdots + \tilde{\sigma}_n^2 \le \|E\|^2, \tag{6.1}$$

so that the defect in rank of $\mathbf{A}$ will be manifest in the size of its trailing singular values.

The inequality (6.1) is actually a perturbation theorem for the zero singular values of a matrix. Weyl's contribution to the theory of the singular value decomposition was to develop a general perturbation theory and use it to give an elegant proof of the approximation theorem. Although Weyl treated integral equations with symmetric kernels, in a footnote on Schmidt's contribution he states, "E. Schmidt's theorem, by the way, treats arbitrary (unsymmetric) kernels; however, our proof can also be applied directly to this more general case." Since here we are concerned with the more general case, we will paraphrase Weyl's development as he might have written it for unsymmetric matrices.

The Location of Singular Values. The heart of Weyl's development is a lemma concerning the singular values of a perturbed matrix. Specifically, if $\mathbf{B}_k = \mathbf{X}\mathbf{Y}^{\mathrm{T}}$, where $\mathbf{X}$ and $\mathbf{Y}$ have k columns (i.e., $\operatorname{rank}(\mathbf{B}_k) \le k$), then

$$\sigma_1(\mathbf{A} - \mathbf{B}_k) \ge \sigma_{k+1}(\mathbf{A}), \tag{6.2}$$

where $\sigma_i(\cdot)$ denotes the ith singular value of its argument.

The proof is simple. Since $\mathbf{Y}$ has k columns, there is a linear combination

$$\mathbf{v} = \gamma_1\mathbf{v}_1 + \gamma_2\mathbf{v}_2 + \cdots + \gamma_{k+1}\mathbf{v}_{k+1}$$

of the first $k+1$ columns of $\mathbf{V}$ (from the singular value decomposition of $\mathbf{A}$) such that $\mathbf{Y}^{\mathrm{T}}\mathbf{v} = 0$. Without loss of generality we may assume that $\|\mathbf{v}\| = 1$, or equivalently that $\gamma_1^2 + \cdots + \gamma_{k+1}^2 = 1$. It follows that

$$\begin{aligned}\sigma_1^2(\mathbf{A}-\mathbf{B}) &\geq \mathbf{v}^{\mathrm{T}}(\mathbf{A}-\mathbf{B})^{\mathrm{T}}(\mathbf{A}-\mathbf{B})\mathbf{v}\\ &= \mathbf{v}^{\mathrm{T}}(\mathbf{A}^{\mathrm{T}}\mathbf{A})\mathbf{v}\\ &= \gamma_1^2\sigma_1^2 + \gamma_2^2\sigma_2^2 + \cdots + \gamma_{k+1}^2\sigma_{k+1}^2\\ &\geq \sigma_{k+1}.\end{aligned}$$

Weyl then proves two theorems. The first states that if $\mathbf{A} = \mathbf{A}' + \mathbf{A}''$ then

$$\sigma_{i+j-1} \leq \sigma_i' + \sigma_j'', \tag{6.3}$$

where the σ_i' and σ_i'' are the singular values of $\mathbf{A}'$ and $\mathbf{A}''$ arranged in descending order of magnitude. Weyl begins by establishing (6.3) for $i = j = 1$:

$$\sigma_1 = \mathbf{u}_1^{\mathrm{T}}\mathbf{A}\mathbf{v}_1 = \mathbf{u}_1^{\mathrm{T}}\mathbf{A}'\mathbf{v}_1 + \mathbf{u}_1^{\mathrm{T}}\mathbf{A}''\mathbf{v}_1 \leq \sigma_1' + \sigma_1''.$$

To establish the result in general, let $\mathbf{A}_{i-1}'$ and $\mathbf{A}_{j-1}''$ be formed in analogy with (5.2). Then $\sigma_1(\mathbf{A}' - \mathbf{A}_{i-1}') = \sigma_i(\mathbf{A}')$ and $\sigma_1(\mathbf{A}'' - \mathbf{A}_{j-1}'') = \sigma_j(\mathbf{A}'')$. Moreover $\operatorname{rank}(\mathbf{A}_{i-1}' + \mathbf{A}_{j-1}'') \leq i + j - 2$. From these facts and from (6.2) it follows that

$$\begin{aligned}\sigma_i' + \sigma_j'' &= \sigma_1(\mathbf{A}' - \mathbf{A}_{i-1}') + \sigma_1(\mathbf{A}'' - \mathbf{A}_{j-1}'')\\ &\geq \sigma_1(\mathbf{A} - \mathbf{A}_{i-1}' - \mathbf{A}_{j-1}'')\\ &\geq \sigma_{i+j-1},\end{aligned}$$

which proves the theorem.

The second theorem is really a corollary of the first. Set $\mathbf{A}' = \mathbf{A} - \mathbf{B}_k$ and $\mathbf{A}'' = \mathbf{B}_k$, where, as above, $\mathbf{B}_k$ has rank k. Since $\sigma_{k+1}(\mathbf{B}_k) = 0$, we have on setting $j = k+1$ in (6.3)

$$\sigma_i(\mathbf{A} - \mathbf{B}_k) \geq \sigma_{k+i}, \qquad i = 1, 2, \ldots.$$

As a corollary to this result we obtain

$$\|\mathbf{A} - \mathbf{B}_k\|^2 \geq \sigma_{k+1}^2 + \cdots + \sigma_n^2.$$

This inequality is equivalent to (5.3) and thus establishes the approximation theorem.

Discussion. Weyl did not actually write down the development for unsymmetric kernels, and we remind the reader once again of the advisability of consulting original sources. In particular, since symmetric kernels can have negative eigenvalues as well as positive ones, Weyl wrote down three sequences of inequalities: one for positive eigenvalues, one for negative, and one — corresponding to the inequalities presented here — for the absolute values of the eigenvalues.

Returning to the perturbation problem that opened this section, if in (6.3) we make the identification $\mathbf{A} \leftarrow \tilde{\mathbf{A}}$, $\mathbf{A}' \leftarrow \mathbf{A}$, and $\mathbf{A}'' \leftarrow \mathbf{E}$, then with $j = 1$ we get

$$\tilde{\sigma}_i \leq \sigma_i + \|\mathbf{E}\|_2,$$

where $\|E\|_2 = \sigma_1(E)$. On the other hand, if we make the identifications $\mathbf{A}' \leftarrow \tilde{\mathbf{A}}$ and $\mathbf{A}'' \leftarrow -\mathbf{E}$, then we get

$$\tilde{\sigma}_i \leq \sigma_i - \|\mathbf{E}\|_2.$$

It follows that

$$|\tilde{\sigma}_i - \sigma_i| \leq \|\mathbf{E}\|_2, \qquad i = 1, 2, \ldots, n.$$

The number $\|\mathbf{E}\|_2$ is called the spectral norm of $\mathbf{E}$. Thus Weyl's result implies that if the singular values of $\mathbf{A}$ and $\tilde{\mathbf{A}}$ are associated in their natural order, they cannot differ by more than the spectral norm of the perturbation.

7. Envoi

With Weyl's contribution, the theory of the singular value decomposition can be said to have matured. The subsequent history is one of extensions, new discoveries, and applications. What follows is a brief sketch of these developments yet to come.

Extensions. Autonne [2, 1913] extended the decomposition to complex matrices. Eckart and Young [12, 1936], [13, 1939] extended it to rectangular matrices and rediscovered Schmidt's approximation theorem, which is often (and incorrectly) called the Eckart–Young theorem.

Nomenclature.[7] The term "singular value" seems to have come from the literature on integral equations. A little after the appearance of Schmidt's paper, Bateman [4, 1908] refers to numbers that are essentially the reciprocals of the eigenvalues of the kernel as singular values. Picard [37, 1910] notes that for symmetric kernels Schmidt's eigenvalues are real and in this case (but not in general) he calls them singular values. By 1937, Smithes was referring to singular values of an integral equation in our modern sense of the word. Even at this point, usage had not stabilized. In 1949, Weyl [52] speaks of the "two kinds of eigenvalues of a linear transformation," and in a 1969 translation of a 1965 Russian treatise on

[7] Parts of this passage were taken from [43, p. 35]

nonselfadjoint operators Gohberg and Krein [16] refer to the “s-numbers” of an operator. For the term “principal component,” see below.

Related Decompositions. Beltrami's proof of the existence of the singular decomposition shows that it is closely related to the spectral decompositions of $\mathbf{A}^{\mathrm{T}}\mathbf{A}$ and $\mathbf{A}\mathbf{A}^{\mathrm{T}}$. It can also be used to derive the polar decomposition of Autonne [1, 1902], [3, 1915], in which is a matrix is factored into the product of a Hermitian matrix and a unitary matrix.

In his investigation of the geometry of n-space, Jordan [27, 1875] introduced canonical bases for pairs of subspaces. This line of development lead to the CS (cosine-sine) decomposition of a partitioned orthogonal matrix introduced implicitly by Davis and Kahan [8, 1970], and explicitly by [42, 1977]. The CS decomposition can in turn be used to derive the generalized singular value decomposition of a matrix, either in the original form introduced by Van Loan [47, 1975] or in the revised version of Paige and Saunders [35, 1981].

Although it is not, strictly speaking, a matrix decomposition, the Moore-Penrose pseudo-inverse [34, 1920], [36, 1955] can be calculated from the singular value decomposition of a matrix as follows. Suppose that the first k singular values of $\mathbf{A}$ are nonzero while the last $n-k$ are zero, and set $\mathbf{\Sigma}^{\dagger} = \mathrm{diag}(\sigma_1^{-1}, \ldots, \sigma_k^{-1}, 0, \ldots, 0)$. Then the pseudo inverse of $\mathbf{A}$ is given by $\mathbf{A}^{\dagger} = \mathbf{U}\mathbf{\Sigma}^{\dagger}\mathbf{V}^{\mathrm{T}}$.

Unitarily Invariant Norms. A matrix norm $\|\cdot\|_{\mathrm{U}}$ is unitarily invariant if $\|\mathbf{U}^{\mathrm{H}}\mathbf{A}\mathbf{V}\|_{\mathrm{U}} = \|\mathbf{A}\|_{\mathrm{U}}$ for all unitary matrices U and V. A vector norm $\|\cdot\|_{\mathrm{g}}$ is a symmetric gauge function if $\|\mathbf{P}\mathbf{x}\|_{\mathrm{g}} = \|\mathbf{x}\|_{\mathrm{g}}$ for any permutation matrix and $\||\mathbf{x}|\|_{\mathrm{g}} = \|\mathbf{x}\|_{\mathrm{g}}$. Von Neumann [49, 1937] showed that to any unitarily invariant norm $\|\cdot\|_{\mathrm{U}}$ there corresponds a symmetric gauge function $\|\cdot\|_{\mathrm{g}}$ such that $\|\mathbf{A}\|_{\mathrm{U}} = \|(\sigma_1, \ldots, \sigma_n)^{\mathrm{T}}\|_{\mathrm{g}}$; i.e., a unitarily invariant norm is a symmetric gauge function of the singular values of its argument.

Approximation Theorems. Schmidt's approximation theorem has been generalized in a number of directions. Mirsky [33, 1960] showed that $\mathbf{A}_k$ of (5.2) is a minimizing matrix in any unitarily invariant norm. The case where further restrictions are imposed on the minimizing matrix are treated in [10, 17, 38].

Given matrices $\mathbf{A}$ and $\mathbf{B}$, The Procrustes problem, which arises in the statistical method of factor analysis, is that of determining a unitary matrix $\mathbf{Q}$ such that $\|\mathbf{A} - \mathbf{B}\mathbf{Q}\|$ is minimized (for the name see [24, 1962]). Green [20, 1952] and Schöneman [40, 1966] showed that if $\mathbf{U}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}\mathbf{B}\mathbf{V} = \mathbf{\Sigma}$ is the singular value decomposition of $\mathbf{A}^{\mathrm{T}}\mathbf{B}$, then the minimizing matrix is $\mathbf{Q} = \mathbf{V}\mathbf{U}^{\mathrm{T}}$. Rao [38, 1980] considers the more general problem of minimizing $\|\mathbf{PA} - \mathbf{BQ}\|$, where $\mathbf{P}$ and $\mathbf{Q}$ are orthogonal.

Principal Components. An alternative to factor analysis is the principal component analysis of Hotelling [22, 1933]. Specifically, if $\mathbf{x}^{\mathrm{T}}$ is a multivariate random variable with mean zero and common dispersion matrix $\mathbf{D}$, and $\mathbf{D} = \mathbf{V}\mathbf{\Sigma}\mathbf{V}^{\mathrm{T}}$ is the eigenvalue-eigenvector decomposition of $\mathbf{D}$, then the components of $\mathbf{x}^{\mathrm{T}}\mathbf{V}$ are uncorrelated with variances σ_i. Hotelling called the transformed variables "the principal components of variance" of $\mathbf{x}^{\mathrm{T}}$. If the rows of $\mathbf{X}$ consist of independent samples of $\mathbf{x}^{\mathrm{T}}$, then the expectation of $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ is proportional to $\mathbf{\Sigma}$. It follows that the matrix $\hat{\mathbf{V}}$ obtained from the singular value decomposition of $\mathbf{X}$ is an estimate $\mathbf{V}$.

Hotelling [23, 1936] also introduced canonical correlations between two sets of random variables that bears the same relation to the generalized singular value decomposition as his principal components bear to the singular value decomposition.

Inequalities Involving Singular Values. Just as Schmidt did not have the last word on approximation theorems, Weyl was not the last to work on inequalities involving singular values. The subject is too voluminous to treat here, and we refer the reader to the excellent survey with references in [21, Ch. 3]. However, mention should be made of a line of research initiated by Weyl [52, 1949] relating the singular values and eigenvalues of a matrix.

Computational Methods The singular value decomposition was introduced into numerical analysis by Golub and Kahan [18, 1965], who proposed a computational algorithm. However, it was Golub [19, 1970] who gave the algorithm that has been the workhorse of the past two decades. Recently, Demmel and Kahan [11, 1990] have proposed an interesting alternative.

8. Acknowledgment

I would like to thank Nick Higham and Hongyuan Zha for reading and commenting on the manuscript.

References

[1] L. Autonne. Sur les groupes linéaires, réels et orthogonaux. *Bulletin de la Société Mathématique de France*, 30:121–134, 1902.

[2] L. Autonne. Sur les matrices hypohermitiennes et les unitairs. *Comptes Rendus de l'Academie des Sciences, Paris*, 156:858–860, 1913.

[3] L. Autonne. Sur les matrices hypohermitiennes et sur les matrices unitaires. *Annales De L'Université de Lyons, Nouvelle Série I*, 38:1–77, 1915.

[4] H. Bateman. A formula for the solving function of a certain integral equation of the second kind. *Transactions of the Cambridge Philosophical Society*, 20:179–187, 1908.

[5] E. Beltrami. Sulle funzioni bilineari. *Giornale di Matematiche ad Uso degli Studenti Delle Universita*, 11:98–106, 1873. An English translation by D. Boley is available as University of Minnesota, Department of Computer Science, Technical Report 90–37, 1990.

[6] A. L. Cauchy. Sur l'équation á l'aide de laquelle on détermine les inégalités séculaires des mouvements des planètes. In *Oeuvres Complétes (II^e Série)*, 1829.

[7] M. Chu. A differential equation approach to the singular value decomposition of bidiagonal matrices. *Linear Algebra and Its Applications*, 80:71–79, 1986.

[8] C. Davis and W. Kahan. The rotation of eigenvectors by a perturbation. III. *SIAM Journal on Numerical Analysis*, 7:1–46, 1970.

[9] P. Deift, J. Demmel, L.-C. Li, and C. Tomei. The bidiagonal singular value decomposition and Hamiltonian mechanics. *SIAM Journal on Numerical Analysis*, 28:1463–1516, 1991.

[10] J. Demmel. The smallest perturbation of a submatrix which lowers the rank and constrained total least squares problems. *SIAM Journal on Numerical Analysis*, 24:199–206, 1987.

[11] J. Demmel and W. Kahan. Accurate singular values of bidiagonal matrices. *SIAM Journal on Scientific and Statistical Computing*, 11:873–912, 1989.

[12] C. Eckart and G. Young. The approximation of one matrix by another of lower rank. *Psychometrika*, 1:211–218, 1936.

[13] C. Eckart and G. Young. A principal axis transformation for non-Hermitian matrices. *Bulletin of the American Mathematical Society*, 45:118–121, 1939.

[14] K. Fan and A. J. Hoffman. Some metric inequalities in the space of matrices. *Proceedings of the American Mathematical Society*, 6:111–116, 1955.

[15] C. F. Gauss. Theoria combinationis observationum erroribus minimis obnoxiae, pars posterior. In *Werke, IV*, pages 27–53, Königlichen Gesellshaft der Wissenschaften zu Göttingin (1880), 1823.

[16] I. C. Gohberg and M. G. Krein. *Introduction to the Theory of Linear Nonselfadjoint Operators.* American Mathematical Society, Providence, Rhode Island, 1969.

[17] G. H. Golub, A. Hoffman, and G. W. Stewart. A generalization of the Eckart-Young matrix approximation theorem. *Linear Algebra and Its Applications*, 88/89:317–327, 1987.

[18] G. H. Golub and W. Kahan. Calculating the singular values and pseudoinverse of a matrix. *SIAM Journal on Numerical Analysis*, 2:205–224, 1965.

[19] G. H. Golub and C. Reinsch. Singular value decomposition and least squares solution. *Numerische Mathematik*, 14:403–420, 1970. Also in [53, pp.134–151].

[20] B. F. Green. The orthogonal approximation of the oblique structure in factor analysis. *Psychometrika*, 17:429–440, 1952.

[21] R. A. Horn and C. R. Johnson. *Topics in Matrix Analysis.* Cambridge University Press, Cambridge, 1991.

[22] H. Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24:417–441 and 498–520, 1933.

[23] H. Hotelling. Relation between two sets of variates. *Biometrika*, 28:322–377, 1936.

[24] J. R. Hurley and R. B. Cuttell. The procrustes program: direct rotation to test a hypothesized factor structure. *Behavioral Science*, 7:258–262, 1962.

[25] C. G. J. Jacobi. Über ein leichtes Verfahren die in der Theorie der Säculärstörungen vorkommenden Gleichungen numerisch aufzulösen. *Journal für die reine und angewandte Mathematik*, 30:51–s94, 1846.

[26] C. G. J. Jacobi. Über eine elementare Transformation eines in Buzug jedes von zwei Variablen-Systemen linearen und homogenen Ausdrucks. *Journal für die reine und angewandte Mathematik*, 53:265–270, 1857, posthumous.

[27] C. Jordan. Essai sur la géométrie à n dimensions. *Bulletin de la Société Mathématique*, 3:103–174, 1875.

[28] C. Jordan. Mémoire sur les formes bilinéaires. *Journal de Mathématiques Pures et Appliquées, Deuxième Série*, 19:35–54, 1874.

[29] C. Jordan. Sur la réduction des formes bilinéaires. *Comptes Rendus de l'Académie des Sciences, Paris*, 78:614–617, 1874.

[30] E. G. Kogbetliantz. Solution of linear systems by diagonalization of coefficients matrix. *Quarterly of Applied Mathematics*, 13:123–132, 1955.

[31] L. Kronecker. Über bilineare Formen. *Sitzungberichte der Königlich Preußischen Akademie der Wissenschaften zu Berlin*, 597–613, 1866.

[32] C. Lanczos. Linear systems in self-adjoint form. *American Mathematical Monthly*, 65:665–679, 1948.

[33] L. Mirsky. Symmetric gauge functions and unitarily invariant norms. *Quarterly Journal of Mathematics*, 11:50–59, 1960.

[34] E. H. Moore. On the reciprocal of the general algebraic matrix. *Bulletin of the American Mathematical Society*, 26:394–395, 1920. Abstract.

[35] C. C. Paige and M. A. Saunders. Toward a generalized singular value decomposition. *SIAM Journal on Numerical Analysis*, 18:398–405, 1981.

[36] R. Penrose. A generalized inverse for matrices. *Proceedings of the Cambridge Philosophical Society*, 51:406–413, 1955.

[37] É. Picard. Sur un théorèm général relatif aux équations intégrales de premièr espèce et sur quelques problèmes de physique mathématique. *Rendicondi del Circolo Matematico di Palermo*, 25:79–97, 1910.

[38] C. R. Rao. Matrix approximations and reduction of dimensionality in multivariate statistical analysis. In P. R. Krishnaiah, editor, *Multivariate Analysis–V*, North-Holland, Amsterdam, 1980.

[39] E. Schmidt. Zur Theorie der linearen und nichtlinearen Integralgleichungen. I Teil. Entwicklung willkürlichen Funktionen nach System vorgeschriebener. *Mathematische Annalen*, 63:433–476, 1907.

[40] P. H. Schöneman. A generalized solution of the orthogonal procrustes problem. *Psychometrika*, 31:1–10, 1966.

[41] J. Schur. Über Potenzreihen, die im Innern des Einkeitskreise beschänkt sind. *Journal für die reine und angewandte Mathematik*, 147:205–232, 1917.

[42] G. W. Stewart. On the perturbation of pseudo-inverses, projections, and linear least squares problems. *SIAM Review*, 19:634–662, 1977.

[43] G. W. Stewart and J.-G. Sun. *Matrix Perturbation Theory.* Academic Press, Boston, 1990.

[44] J. J. Sylvester. A new proof that a general quadric may be reduced to its canonical form (that is, a linear function of squares) by means of a real orthogonal substitution. *Messenger of Mathematics*, 19:1–5, 1889.

[45] J. J. Sylvester. On the reduction of a bilinear quantic of the n^{TH} order to the form of a sum of n products by a double orthogonal substitution. *Messenger of Mathematics*, 19:42–46, 1889.

[46] J. J. Sylvester. Sur la réduction biorthogonale d'une forme linéo-linéaire à sa forme cannonique. *Comptes Rendus de l'Academie des Sciences, Paris*, 108:651–653, 1889.

[47] C. F. Van Loan. A general matrix eigenvalue algorithm. *SIAM Journal on Numerical Analysis*, 12:819–834, 1975.

[48] J. von Neumann. *Collected Works,* (A. H. Taub Editor). Pergamon, New York, 1962.

[49] J. von Neumann. Some matrix-inequalities and metrization of matrix-space. *Tomsk. Univ. Rev.*, 1:286–300, 1937. In [48, v.4, pp.205–219].

[50] K. Weierstrass. Zur Theorie der bilinearen und quadratischen Formen. *Monatshefte Akadamie Wissenshaften Berlin*, 310–38, 1868.

[51] H. Weyl. Das asymptotische Verteilungsgesetz der Eigenwert linearer partieller Differentialgleichungen (mit einer Anwendung auf der Theorie der Hohlraumstrahlung). *Mathematische Annalen*, 71:441–479, 1912.

[52] H. Weyl. Inequalities between the two kinds of eigenvalues of a linear transformation. *Proceedings of the National Academy of Sciences*, 35:408–411, 1949.

[53] J. H. Wilkinson and C. Reinsch. *Handbook for Automatic Computation. Vol. II Linear Algebra.* Springer, New York, 1971.