A tutorial history of least squares with applications to astronomy and geodesy ☆

Yves Nievergelt [*]

Department of Mathematics, Eastern Washington University, MS-32, Cheney, WA 99004-2431, USA

Abstract

This article surveys the history, development, and applications of least squares, including ordinary, constrained, weighted, and total least squares. The presentation includes proofs of the basic theory, in particular, unitary factorizations and singular-value decompositions of matrices. Numerical examples with real data demonstrate how to set up and solve several types of problems of least squares. The bibliography lists comprehensive sources for more specialized aspects of least squares. © 2000 Elsevier Science B.V. All rights reserved.

Keywords: History; Constrained; Ordinary; Total; Weighted; Least squares

1. Introduction

The mathematical concept of least squares is the basis for several methods to fit certain types of curves and surfaces to data. Problems of fitting curves and surfaces have a history spanning several millenia, which is outlined in Section 2 to set in perspective the contribution of least squares to their solution. The citations provided here include page numbers from Dreyer's book [13] to identify the original texts. Examples of such problems include the determination of the shape and size of celestial bodies and of their trajectories.

These problems were still without satisfactory solutions near the end of the eighteenth century A.D., at the time of the development of the concepts of problems of least squares and their solution with normal equations; see Section 3. (For greater detail, see Stewart's translation [16] of Gauss's work.)

For computations with floating-point or other approximate arithmetics, normal equations can exhibit a sensitivity to errors in data or in rounding larger than the sensitivity of methods with unitary factorizations. These factorizations also provide methods to solve problems of constrained and total least squares, as explained in Sections 4 and 5.

For the state-of-the-art in computing with least squares near the end of the second millenium A.D., Björk [1], Dennis Jr., and Schnabel [12], and Lawson and Hanson [32] present algorithms to solve least-squares problems, and Higham [22] also treats the analysis of sensitivity to errors. Van Huffel and Vandewalle [50] focus on total least-squares. These references also contain extensive bibliographies. To compute solutions of practically significant problems, the usual recommendation is to use one of the professionally maintained libraries of computer programs, for instance, `netlib` (`http://www.netlib.org/lapack/`).

2. An ancient history of curve and surface fitting

2.1. Fitting surfaces: the shapes of the earth and of the moon

One fitting problem consists in estimating the shape of the earth. Early in the first millenium B.C., several shapes were fitted to various combinations of religious canons, philosophical doctrines, and observations of the rôles of air, earth, fire, and water. Types of surfaces fitted to such ideas included a *circular disc* (Thales of Miletus, about 640–562 B.C. [13, p. 11]), an *infinite plane* (Xenophanes of Kolophon, about 570–475 B.C. [13, p. 18]), and a *sphere* (Parmenides of Elea, early in the fifth century B.C. [13, p. 20]). The type of surface was also fitted to observations of inequalities reported by travelers. For example, the star Canopus remained invisible to a traveler in Greece, became just visible above the horizon at Rhodes, and then appeared higher and higher above the horizon as the traveler went further and further south [13, p. 20]. Also while sailing toward the setting sun, mariners in the north saw the sun on their left, but mariners in the south saw the sun on their right [13, p. 39]. From the fifth century B.C., in Greece and India, the type of surface fitted to such observations was a sphere [13, pp. 39, 242].

Similarly, for the shape of the moon, a sphere fitted the observation that the lighted side of the moon always faces the sun (Parmenides [13, p. 21]; Anaxagoras of Klazomenæ, about 500–428 B.C. [13, p. 32]).

With the shape settled to be a sphere arises the problem of estimating its size.

To estimate the circumference of the earth, Posidonius of Apameia (about 135–50 B.C.) referred to a result attributed to Archimedes (287–212 B.C.) and Dikæarchus of Messana (about 285 B.C.), using two stars seen from two cities; see Fig. 1. The cities are Lysimachia in Thrace, and Syene in Upper Egypt, which lie 20 000 stadia apart from each other. The first star, γ Draconis, appears at the zenith (vertical direction) above Lysimachia. The second star, in the constellation Cancer, appears at the zenith above Syene. The difference between the declinations (angular elevations from the celestial equator) of the two stars is $1/15$ of a full circle, which is thus the difference between the vertical directions at the two cities. Therefore, the circumference of the earth is $15 * 20\,000 = 300\,000$ stadia, corresponding to approximately $100\,000$ stadia for the earth's diameter [13, pp. 173–174]. (Though Archimedes and Apollonius already knew the approximations $\pi \approx 22/7$ and $\pi \approx 3.1416$
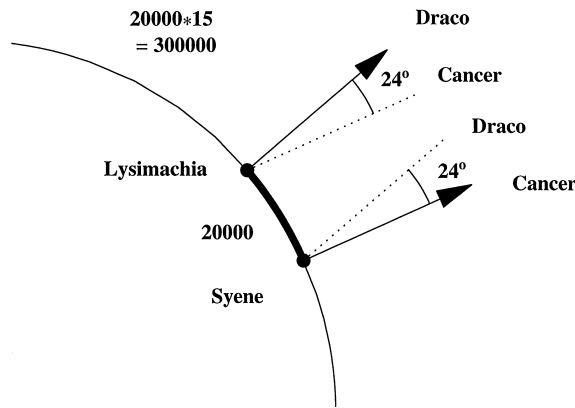
Fig. 1. Posidonius's estimate of the earth's circumference. Stars appear in the same direction from every point on earth. Two stars make an angle $2\pi/15$. One of them is at the zenith above Syene, the other is at the zenith at Lysimachia. Therefore, 15 times the distance from Syene to Lysimachia equals the earth's circumference.
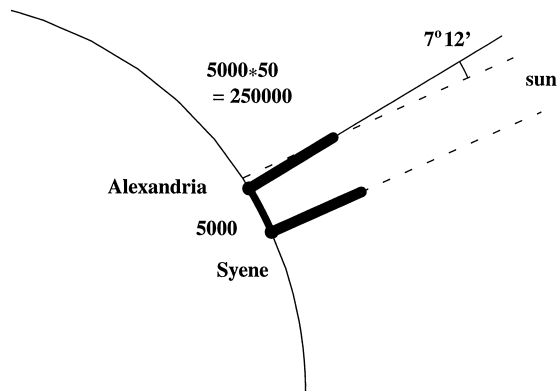

Fig. 2. Eratosthenes's estimate of the earth's circumference. The sun rays appear parallel on earth. They are vertical at Syene. At Alexandria, with a vertical stick they make an angle 1/50 of a full circle. Therefore, 50 times the distance from Syene to Alexandria equals the earth's circumference.

[49, pp. 185–186], the approximation $\pi \approx 3$ was then common for practical purposes not only in Greece but also in Babylon, in Egypt [49, p. 173], and in China [49, p. 196].)

With a different procedure, Eratosthenes of Alexandria (276–194 B.C.) used the shadows of vertical rods in two cities; see Fig. 2. At the summer soltice, at Syene the rod casts no shadow, so that the sun rays fall vertically, while at Alexandria the sun rays and the vertical rod make and angle equal to 1/50 of a full circle. (According to van der Waerden, the computation of this angle from measurements of the lengths of the rod and of its shadow proceeded through the Theorem of Pythagoras and tables of sines [49, p. 214].) Because Syene lies 5000 stadia away from Alexandria, it follows that the circumference of the earth is about $50*5000=250\,000$ stadia. Kleomedes corroborated this results through the same procedure at the same locations but at the winter soltice. Table 1 shows comparisons with the World Geodetic System WGS-84 [23].

Table 1
Comparisons of estimates of the earth's polar circumference and radius

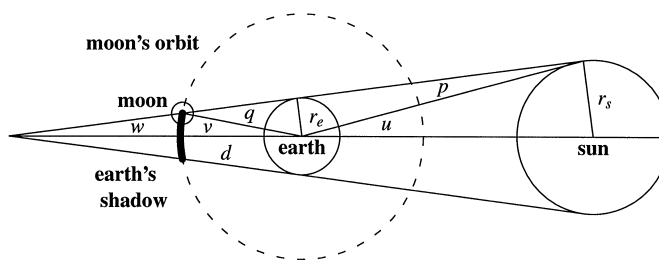| Source	Circumference	Radius
Archimedes, Dikæarchus,	300 000 stadia	50 000 stadia
Posidonius, 3rd century B.C.	(47 250 000 m)	(7 875 000 m)
Eratosthenes, 2nd century B.C.	250 000 stadia	
(1 stade ≈ 157.5 m)	(39 375 000 m)	(6 562 500 m)
WGS-84, 1984 A.D.		$a = 6\,378\,137.000\,00$ m
		$b = 6\,356\,752.314\,25$ m
		$e^2 = 0.006\,694\,379\,990\,13$
Mathematica `4aEllipticE[`e^2`]`	40 007 862.917 27 m	

Fig. 3. Hipparchus's estimate of the radii of the moon and its orbit.

The estimate of the same circumference by different observers through different procedures or through repeated measurements hints at some attempts to detect errors, but no records of such attempts appear to remain [13, p. 177].

2.2. Fitting curves: the radii of the moon and its orbit

Another fitting problem consists in estimating the trajectories of celestial bodies. For example, rectilinear motions fitted the poetical ideas of Xenophanes in the sixth century B.C. [13, p. 18]. A century later, Philolaus of Thebes proposed circular orbits for the earth, the moon, the planets, and the sun, all around a "central fire" reflected by the sun toward the earth; such orbits fitted coarse observations of planetary motions [13, pp. 40–49]. In the third century B.C., Aristarchus of Samos outlined a heliocentric system with a circular orbit for the earth around the sun [13, p. 137].

With the orbits settled as circles arises the problem of estimating their size.

To estimate simultaneously the distance from the earth to the moon and the radius of the moon, Hipparchus of Nicæa (second century B.C.) used a full lunar eclipse [13, pp. 183–184]; see Fig. 3. Within the measurement accuracy available then, the sun's parallax p is nearly zero. Seen from the earth, the sun sustends an angle $u = 16'36''55'''$, and the path of the moon across the earth's shadow sustends and angle $v = 41'32''17.5'''$. The ratio $180°/v \approx 260$ can also be calculated as the ratio t_1/t_2 of the time t_1 of a full revolution of the moon (29.5 days) and the time t_2 taken by the moon to cross the earth's shadow. Consequently, the parallax of the earth's shadow on the moon is nearly $q = u + v = 58'09''12.5'''$. Therefore, the ratio d/r_e of the distance d from the earth to

Table 2
Comparisons of estimates of the radii of the moon and its orbit

Source	Moon's radius	Orbit's radius
Hipparchus, 2nd century B.C.	$r_m = r_e/3.5$	$d = 59.1 r_e$
	(1 875 000 m)	(387 843 750 m)
Hipparchus, 2nd century B.C.,	$r_m = r_e/3.4$	$d = 60\frac{5}{6} r_e$
reported by Kleomedes.	(1 930 147 m)	(399 218 750 m)
[26, p. 476], 1984 A.D.	1 738 000 m	384 400 000 m
	($b/r_m \approx 3.658$)	($d/b \approx 60.47$)

the moon and the earth's radius r_e is $1/\sin(q) = 1/\sin(u + v) = 59.1$. Moreover, the diameter of the earth's shadow at distance d from the earth equals about $d * v/180° = d * t_2/t_1$. A measurement of the time t_3 from the moment the moon touches the earth's shadow to the moment it disappears in it then gives an estimate of the radius of the moon r_m in the form $2r_m/(d * t_2/t_1) = t_3/t_2$, whence $r_m = (d/2) * (t_3/t_1) = (59.1 r_e/2) * (t_3/t_1) = r_e/3.5$.

According to Ptolemy's account, Hipparchus attempted to measure a lower bound and an upper bound for the sun's parallax p. The results just presented correspond to the lower bound 0. Kleomedes's report of another result from Hipparchus, $d = 60\frac{5}{6} r_e$ [13, pp. 183–184], corresponds to the upper bound $2'44''$. Such bounds hint at attempts to detect the maximum error.

With Eratosthenes's measure of the earth's radius, Hipparchus's results give 387 843 750 m for the distance of the moon, and 1 875 000 m for the radius of the moon. Table 2 shows comparisons with textbook values [26, p. 476].

2.3. Fitting curves and surfaces: planetary orbits and earth's geoid

It was also considerations of maximum errors, of the order of $8'$ between Tycho Brahe's observations of Mars and Copernicus's heliocentric model, which led Johann Kepler to abandon circles for the orbits, and finally (about 18 December 1604 A.D.) to substitute *ellipses* with a focus at the sun, along which planets sweep equal areas in equal times [13, pp. 389–392]. In 1687, Isaac Newton outlines in the *Philosophiae Naturalis Principia Mathematica* a proof that Kepler's laws are mathematically equivalent to the action of an attraction from the sun and inversely proportional to the square of the distance from the sun to the planet [39].

From Newton's law of gravitational attraction, it follows (from mathematical derivations by Newton, Ivory, Huygens, Clairaut, and Laplace [31, Book III, Section 18]) that a rotating mass of a homogenous and incompressible fluid can have the shape of an *ellipsoid* rotating around its shortest axis [20, pp. 172–175]. From 1700 through 1733, three surveys in France all suggested that the earth was an ellipsoid rotating around its *largest* axis [4, pp. 250–251]; such a surface failed to fit Newton's mathematical theory, based on Kepler's physics, itself based on Tycho Brahe's measurements. Ordered by Louis XV, a survey in Lapland and a survey in Peru in 1735 reversed the earlier results and confirmed that the earth was an ellipsoid rotating around it shortest axis [4, pp. 251–252].

The foregoing historical outline shows that for nearly three millenia, curves and surfaces were fitted to ideologies and theories. Yet errors — discrepancies between the fitted curve or surface and
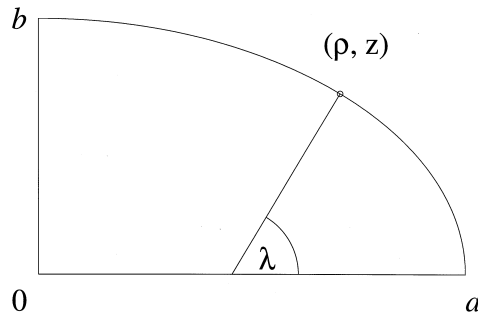
Fig. 4. The geodetic latitude of a point (ρ, z) is the angle λ between the normal to the surface through (ρ, z) and the equatorial plane.

observations — drew attention through gross departures from the theory or through unacceptable *maximum* values.

3. Weighted ordinary least squares and geodesy

3.1. Precursors: minimax and minimum average modulus

By the end of the 18th century A.D., the Marquis Pierre Simon de Laplace (1749–1827) was using a sequence of several methods to fit curves and surfaces to measurements in geodesy and astronomy. Each of his methods minimizes either the maximum residual, the average absolute residual, or the average squared residual, of a linearized model.

For example, consider the problem of fitting an ellipse to a polar cross section of the earth, with principal semi-axes of lengths $a \geqslant b > 0$. Let $e^2 := 1 - (b/a)^2$ be its squared eccentricity, and set $\sigma^2 := 1 - e^2 = (b/a)^2$. For each point $x := (x, y, z)$ on the earth's surface, the geodetic latitude of x is the angle λ between the normal to the surface at x and the equatorial plane, as in Fig. 4. With the cylindrical coordinate $\rho := \sqrt{x^2 + y^2}$, calculus gives

$$\rho = \frac{a \cos(\lambda)}{\sqrt{1 - e^2 [\sin(\lambda)]^2}}, \qquad z = \frac{a\sigma^2 \sin(\lambda)}{\sqrt{1 - e^2 [\sin(\lambda)]^2}}.$$

Hence, the differential of the arclength s along a meridian becomes

$$\mathrm{d}s = \frac{a\sigma^2}{\{1 - e^2 [\sin(\lambda)]^2\}^{3/2}} \, \mathrm{d}\lambda$$

$$= a\sigma^2 \{1 + \tfrac{3}{2} e^2 [\sin(\lambda)]^2 + \frac{3 * 5}{2 * 2 * 2!} e^4 [\sin(\lambda)]^4 + \cdots\} \, \mathrm{d}\lambda.$$

Crude approximations indicate that $e^2 < 0.01$. Beyond the first two terms,

$$\sum_{k=2}^{\infty} \prod_{\ell=1}^{k} (2\ell + 1) \frac{|e \sin(\lambda)|^{2k}}{2^k * (k!)} < \frac{15 e^4}{8} \left(1 + \frac{7 e^2}{6}\right) \sum_{k=o}^{\infty} e^k < 0.000\,25.$$

Thus, with a relative error less than 0.00025 uniformly over the earth's surface, the length Δs of an arc $\Delta\lambda$ of meridian at the geodetic latitude λ takes the following form, with $c_0 := a\sigma^2$ and $c_1 := \frac{3}{2}a\sigma^2 e^2$:

$$\frac{\Delta s}{\Delta\lambda} = c_0 + c_1[\sin(\lambda)]^2.$$

Thus, measurements of the lengths of n arcs of a meridian produce n equations.

Example 1. With lengths in double toises $(1/0.256537\text{ m})$ and angles in grads $(2\pi = 400^\circ)$, Laplace considered the following system [31, Book III, Section 41]:

$\Delta s/\Delta\lambda = c_0 + c_1[\sin(\lambda)]^2$;	location;	latitude λ;	arc $\Delta\lambda$;
$25538.85 = c_0 + c_1 * 0.00000$;	Peru;	00.0000°;	3.4633°;
$25666.65 = c_0 + c_1 * 0.30156$;	Good Hope;	37.0093°;	1.3572°;
$25599.60 = c_0 + c_1 * 0.39946$;	Pennsylvania;	43.5556°;	1.6435°;
$25640.55 = c_0 + c_1 * 0.46541$;	Italy;	47.7963°;	2.4034°;
$25658.28 = c_0 + c_1 * 0.52093$;	France;	51.3327°;	10.7487°;
$25683.30 = c_0 + c_1 * 0.54850$;	Austria;	53.0926°;	3.2734°;
$25832.25 = c_0 + c_1 * 0.83887$;	Lapland;	73.7037°;	1.0644°.

The problem then consisted in fitting c_0 and c_1 to this linear system.

Laplace's first method aimed at determining the ellipsoid that *minimizes the maximum error* between the fitted ellipsoid and the measurements [31, Book III, Section 39]. From this first method he concluded that the earth's surface was not exactly an ellipsoid but the maximum error was within the measurement accuracy, with a flattening $f := 1 - (b/a) = 1/277$ [31, Book III, Section 41], which corresponds to a squared eccentricity $e^2 < 0.007207 < 0.01$.

Laplace's second method aimed at determining the ellipsoid that *minimizes the average absolute values of the errors* subject to the constraint that *the sum of the errors equal zero*; the result yielded what he considered the most probable ellipsoid [31, Book III, Section 40].

The second method presented several difficulties. Firstly, the "most probable" estimate depends on the probability distribution of the errors and can fail to coincide with the minimum average absolute error [24, pp. 400–401]. Secondly, Laplace's method did not lend itself to the methods of power series, and no efficient algorithm existed to determine the solutions (until George B. Dantzig's simplex algorithm in the 1950s [6,10,11]). Finally, for an overdetermined system of linear equations with a matrix of any rank, Laplace's method can lead to multiple solutions filling an entire polytope [6, p. 219].

Example 2. Consider the following system $Ax = b$ with maximal rank:

$$x + y = 4,$$
$$x - y = 0,$$
$$x - y = 2,$$
$$x + y = 6.$$

The residuals $r = Ax - b$ add to zero at $x = 3$. Setting $x = 3$ gives

$$\{|(3 + y) - 4| + |(3 - y)| + |(3 - y) - 2| + |(3 + y) - 6|\}/4$$

$$= \{|y - 1| + |y - 3|\}/2$$

$$= \begin{cases} 2 - y > 1 & \text{if } y < 1, \\ 1 = 1 & \text{if } 1 \leqslant y \leqslant 3, \\ y - 2 > 1 & \text{if } 3 < y. \end{cases}$$

The average reaches its minimum everywhere on the segment $\{3\} \times [1,3]$.

Each of Laplace's numerical examples of a minimization of the average absolute error consists of an *odd* number of equations [31, Book III, Sections 41–42]. In contrast, for the determination of orbits of celestial bodies, Laplace used an *even* number of linearized equations, corresponding to measurements at times scattered symmetrically about a central time t_0:

$$t_0 - t_k, t_0 - t_{k-1}, \ldots, t_0 - t_1, t_0 + t_1, \ldots, t_0 + t_{k-1}, t_0 + t_k.$$

This produces a peculiar type of linear system, where the first column of coefficients $A(; 1)$ is perpendicular to the second column of coefficients $A(; 2)$, as in Example 2. For such systems, Laplace did not minimize the average absolute error. Instead, in effect, he computed the dot product of the system with the transposed column $A(; 1)^*$ and solved for x, and then computed the dot product of the system with $A(; 2)^*$ and solved for y [31, Book II, Section 37].

Example 3. Consider the system $Ax = b$ from Example 2:

$$\begin{pmatrix} 1 & 1 \\ 1 & -1 \\ 1 & -1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 4 \\ 0 \\ 2 \\ 6 \end{pmatrix},$$

$$\begin{pmatrix} 4x \\ 4y \end{pmatrix} = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & -1 & -1 & 1 \end{pmatrix} \begin{pmatrix} 1 & 1 \\ 1 & -1 \\ 1 & -1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & -1 & -1 & 1 \end{pmatrix} \begin{pmatrix} 4 \\ 0 \\ 2 \\ 6 \end{pmatrix} = \begin{pmatrix} 12 \\ 8 \end{pmatrix}.$$

Thus, $x = 3$ and $y = 2$.

For the peculiar type of linear system in Examples 2 and 3, Laplace's method amounts to solving the normal equations for the least squares solution.

3.2. Weighted ordinary least squares

Around 1800, Laplace, Legendre, and Gauss were fitting functional forms to data through various types of least squares. Laplace's method applied to systems with mutually orthogonal columns. Legendre (1752–1833) published the method of normal equations in 1805 [33]. In 1821–1823, Gauss published the method of weighted least squares to solve linear systems $Ax = b$ with a matrix A with

n linearly independent columns and $m \geqslant n$ rows [16]. Though Gauss did not employ a notation with matrices, a derivation of weighted least squares with matrices can proceed as follows [47, p. 144].

The problem consists in determining a linear function u of n variables $a_1, \ldots, a_n$, which amounts to determining coefficients $x_1, \ldots, x_n$ so that

$$u(a_1, \ldots, a_n) = a_1 x_1 + \cdots + a_n x_n.$$

The data consist of m points $A(i;) = (a_{i,1}, \ldots, a_{i,n})$, arranged as the rows of the matrix A, and of the value b_i if u at each point. The problem then amounts to fitting coefficients $x_1, \ldots, x_n$ to the linear system $Ax = b$:

$$a_{1,1} x_1 + \cdots + a_{1,n} x_n = b_1,$$
$$\vdots$$
$$a_{m,1} x_1 + \cdots + a_{m,n} x_n = b_\mathrm{m}.$$

The data can also include estimates of the precision of the measurements, in the form of the reciprocal of the variance of each measurement, as investigated by Gauss, or, more generally, in the form of the inverse V^{-1} of the covariance matrix V of the measurements, as investigated by Aiken [2]. Specifically, if b_i represents the average $E(B_i)$ of a random variable B_i, estimated by the average of several observations, then $V_{i,j} = E[(B_i - b_i)(B_j - b_j)]$ is the covariance of B_i and B_j. The solution X of the linear system $AX = B$ is then also a random variable. The problem solved by Gauss consists in finding a linear transformation L such that $LA = I$, to solve for $\tilde{x} = I\tilde{x} = LA\tilde{x} = Lb$, such that $\tilde{x} = Lb$ minimizes the covariance

$$U = E[(X - \tilde{x})(X - \tilde{x})^*].$$

Gauss showed that $\tilde{x}$ is also the solution of the weighted least-squares system

$$WAx = Wb$$

with a matrix of weights W such that $W^* W = V^{-1}$, and then

$$L = (A^* V^{-1} A)^{-1} A^* V^{-1}.$$

Indeed, for every matrix K such that $KA = I$,

$$U = E[(X - \tilde{x})(X - \tilde{x})^*]$$

$$= E[(X - KB)(X - KB)^*]$$

$$= E\{[X - KAX - K(B - AX)][X - KAX - K(B - AX)]^*\}$$

$$= E\{[K(B - AX)][K(B - AX)]^*\}$$

$$= KE\{(B - AX)(B - AX)^*\}K^*$$

$$= KVK^*$$

$$= LVL^* + (K - L)VL^* + LV(K - L)^* + (K - L)V(K - L)^*.$$

The two middle terms equal zero, because of the condition $KA = I$ and the definition of L. The last term, $(K - L)V(K - L)^*$, is hermitian positive semidefinite. Hence, for each vector z,

$$z^* U z = z^* LVL^* z + z^*(K - L)V(K - L)^* z \geqslant z^* LVL^* z,$$

with z^*Uz minimum for $K:=L$. Moreover, the formula for U simplifies to

$$U = (A^*V^{-1}A)^{-1},$$

which is thus the covariance matrix of the weighted least squares solution X.

Example 4. For the system in Example 1, Laplace weighted each equation by the number of degrees $\Delta\lambda$ in the corresponding measured arc. There was no known correlation between the measurements from the different teams assigned to measure different arcs. Thus the weight matrix W is diagonal with the corresponding values of $\Delta\lambda$ on its diagonal:

$$W = \text{diagonal } (3.4633, 1.3572, 1.6435, 2.4034, 10.7487, 3.2734, 1.0644).$$

The weighted least-squares solution (computed through the command LSQ on a Hewlett–Packard's HP48GX calculator [21, pp. 14,15]) is

$$c_0 = 25534.47,$$
$$c_1 = 242.81.$$

Hence

$$e^2 = \tfrac{2c_1}{3c_0} = 0.006\,339\ldots,$$
$$\sigma^2 = 1 - e^2 = 0.993\,661\ldots,$$
$$f = 1 - b/a = 1 - \sigma = 0.003\,175\ldots,$$
$$a = \tfrac{c_0}{\sigma^2} = 25697.38^R\ldots = 100\,170.25 \; m.$$

Laplace gives $f = 1/277 = 0.003\,610$, though Bowditch's calculations of Laplace's method lead to $f = 1/250 = 0.004$ [31, Book III, Section 41]. The values from WGS-84 are $f = 0.003\,352\,810\,664\,74$ and $a = 6\,378\,137$ m [23, p. xxiii]. Finally,

$$U = (A^*V^{-1}A)^{-1} = (A^*W^*WA)^{-1} = \begin{pmatrix} 0.007\,701 & -0.147\,686 \\ -0.147\,686 & 0.309\,706 \end{pmatrix}.$$

The method of weighted least squares assigns weights only to the measured values b of the function u, but not to be coordinates of the points $(a_{i,1},\ldots,a_{i,n})$. In Laplace's application, this would correspond to treating the measurements of the lengths of arcs of meridians as random variables, but considering the determinations of the geodetic latitudes as exact. Allowances for adjustments of all data require different methods, as explained below in Section 6.

4. Unitary factorizations and constrained least squares

4.1. Householder symmetries and unitary factorizations

To solve linear systems, Gaussian elimination performs a linear transformation known as a shear that maps a column of coefficients $r = A(\;;j)$ to a multiple of a canonical basis vector e_j, which "eliminates" the coefficients below the jth row. Yet shears alter Euclidean distances, in particular,
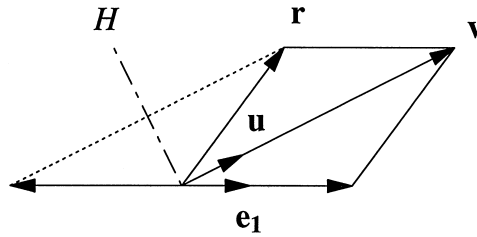
Fig. 5. A Householder symmetry maps r to $-\text{sign}(r_1)\|r\|_2 e_1$.

they do not reveal which vector lies closest to the "right-hand side" of the system. In contrast, one of the strategies for solving least squares problems consists in replacing Gaussian elimination by a type of linear elimination that preserves Euclidean distances, for instance, Modified Gram-Schmidt (MGS) orthogonalization [22, Section 19.3, 43], Givens rotations, or Householder symmetries [1,32].

Householder symmetries involves the function sign: $\mathbb{C} \to \mathbb{C}$ defined by

$$\text{sign}(z) := \begin{cases} z/|z| & \text{if } z \neq 0, \\ 1 & \text{if } z = 0. \end{cases}$$

For each non-zero vector $r \in \mathbb{C}^m \setminus \{0\}$, a Householder symmetry reflects r onto a multiple $-\text{sign}(r_1)\|r\|_2 e_1$ of the basis vector e_1 across the hyperplane $H \subset \mathbb{C}^m$ that passes through the origin 0 perpendicularly to the bisectrix of the angle formed by r and $\text{sign}(r_1)e_1$, which lies in the direction of $v := r + \text{sign}(r_1)\|r\|_2 \cdot e_1$, as in Fig. 5. The choice of $\text{sign}(r_1)$ minimizes rounding inaccuracies, so that if $r \neq 0$ then $v := r + \text{sign}(r_1)\|r\|_2 e_1 \neq 0$, because $|v_1| = |r_1 + \text{sign}(r_1)\|r\|_2| \geq \|r\|_2 > 0$. The hyperplane H is then perpendicular to the unit vector $u := (1/\|v\|_2)v$.

A Householder symmetry S thus amounts to subtracting from r twice its projection along u, so that $S(r) = r - 2\langle r, u \rangle u$, which leads to Algorithm 1.

Algorithm 1. Data: *any non-zero vector $r \in \mathbb{C}^m \setminus \{0\}$*
(1) $s := sign(r_1)$.
(2) $v := r + s\|r\|_2 e_1$.
(3) $v := 1/\{\|r\|_2(\|r\|_2 + |r_1|)\}$.

Result. $S(Z) = Z - vv(v^*Z)$ *for every* $Z \in \mathbb{M}_{m \times n}(\mathbb{C})$.

Proposition 5 verifies that Algorithm 1 produces a Householder symmetry.

Proposition 5. *The transformation S defined by algorithm 1 reflects r onto $S(r) = -s \cdot \|r\|_2 \cdot e_1$. Moreover, the matrix S of S is hermitian and unitary.*

Proof. With $S(r) = r - vv(v^*r)$ defined as in Algorithm 1,

$$v^*r = \sum_{j=1}^m \overline{v_j}r_j = \overline{v_1}r_1 + \sum_{j=2}^m \overline{v_j}r_j = (\overline{r_1} + \overline{s}\|r\|_2)r_1 + \sum_{j=2}^m \overline{r_j}r_j$$

$$= \|r\|_2^2 + \overline{s}\|r\|_2 r_1 = \|r\|_2(\|r\|_2 + |r_1|) = 1/v,$$

$$S(r) = r - v \cdot v \cdot (1/v) = r - v = r - (r + s \cdot \|r\|_2 \cdot e_1) = -s\|r\|_2 e_1.$$

Moreover, $\|v\|_2^2 = 2/v$:

$$\|v\|_2^2 = \sum_{j=1}^{m} |v_j|^2 = |r_1 + s\|r\|_2|^2 + \sum_{j=2}^{m} |r_j|^2$$

$$= |s|^2 \|r\|_2^2 + (r_1\bar{s} + \overline{r_1}s)\|r\|_2 + |r_1|^2 + \sum_{j=2}^{m} |r_j|^2$$

$$= \|r\|_2^2 + 2|r_1| \cdot \|r\|_2 + \|r\|_2^2 = 2\|r\|_2(\|r\|_2 + |r_1|) = 2v.$$

Consequently, the Householder symmetry S has a hermitian matrix, $S^* = S$:

$$S^* = (I - vv^*)^* = I^* - v(v^*)^* v^* = S.$$

Finally, the Householder symmetry S is a unitary transformation, $S^*S = I$:

$$S^*S = (I - vv^*) \cdot (I - vv^*) = I - 2Ivv^* + vv^*vv^*$$

$$= I - 2Ivv^* + v^2v\|v\|_2^2 v^* = I - 2vv^* + 2vv^* = I. \qquad \square$$

Applied to the first column $r := A(\ ;1)$ of any rectangular matrix $A \in \mathbb{M}_{m \times n}(\mathbb{C})$, the Householder symmetry S produces zeros under the first entry $(SA)_{1,1} = -s\|A(\ ;1)\|_2$, and transforms the subsequent columns into $(SA(\ ;2), \dots, SA(\ ;n))$. By induction, Householder symmetries $S_1, \dots, S_n$ (such that each S_k modifies only entries on or below the kth row) produce a unitary — but not necessarily hermitian — matrix $Q = S_n^* \cdots S_1^*$, and an upper-triangular matrix R, with

$$A = QR.$$

4.2. Solving least-squares problems with orthogonal factorizations

Consider a linear system $Ax = b$ with $n \leqslant m$ linearly independent columns in $A \in \mathbb{M}_{m \times n}(\mathbb{C})$. If $A = QR$ with Q unitary and R upper triangular, then

$$Ax = b,$$

$$Q^*Ax = Q^*b,$$

$$Rx = Q^*b,$$

where multiplication Q preserves Euclidean distances, whence

$$\|Ax - b\|_2 = \|Rx - Q^*b\|_2.$$

Because R has n linearly independent columns and has only zeros below the rth row, $\|Rx - Q^*b\|_2$ reaches a minimum if and only if x is the unique solution $\tilde{x}$ of the first n equations. Moreover,

$$\|R\tilde{x} - Q^*b\|_2 = \|((Q^*b)_{n+1}, \dots, (Q^*b)_m)\|_2.$$

For a matrix $A \in \mathbb{M}_{m \times n}(\mathbb{C})$ with rank $r \leqslant \min\{m, n\}$ and columns that *need not* be linearly independent, there exists a unitary factorization

$$AP = QR.$$

The matrix $P \in \mathbb{M}_{n \times n}(\mathbb{C})$ permutes the columns of A so that the first r columns of AP are linearly independent. Householder symmetries then yield a unitary factorization of the first r columns,

$$[AP(\ ;1), \ldots, AP(\ ;r)] = Q[R(\ ;1), \ldots, R(\ ;r)],$$

and $R = Q^*(AP)$ contains only zeros below the rth row. With the change of coordinates $z := P^{-1}x$, there is then an affine subspace of dimension $n - r$ of least-squares solutions to the system

$$Ax = b,$$

$$(Q^*AP)(P^{-1}x) = Q^*b,$$

$$Rz = Q^*b.$$

One solution z results from setting $z_{r+1} := \cdots := z_n := 0$ and solving the $z_1, \ldots, z_r$. the shortest least-squares solution x is then the orthogonal projection of any solution z on the orthogonal complement of the null space of R, in other words, on the row space of R.

Such a projection can employ a unitary factorization of R^*,

$$R^* = WT$$

with W unitary and T upper triangular. Because $T = W^*R^*$ has only zeros below its rth row, it follows that the last $n - r$ columns $w_{r+1}, \ldots, w_n$ of W form an orthonormal basis of Kernel (R), while the first r columns $w_1, \ldots, w_r$ form an orthonormal basis of its row space. Consequently,

$$\tilde{x} := (w_1 \ldots w_r) \begin{pmatrix} w_1^* \\ \vdots \\ w_r^* \end{pmatrix} z$$

minimizes $\|A\tilde{x} - b\|_2$ with the smallest norm $\|\tilde{x}\|_2$; see also [32, Chapter 14].

In principle, the permutations P can be generated during the computation of each symmetry S_k, by swapping columns $A(\ ;k)$ and $A(\ ;\ell)$ for some $\ell > k$ if $A(\ ;k)$ lies in the subspace spanned by $A(\ ;1), \ldots, A(\ ;k-1)$. However, detecting such linear dependencies and selecting a permutation amounts to computing the ranks of submatrices, which is not reliable with floating-point or other approximate arithmetics [12, p. 66]. The singular value decomposition will provide some information on the reliability of such computations.

4.3. Constrained least squares and geodesy

Such practical situations as geodesy lead to problems of least squares with linear constraints. The outline presented here expands on that of Lawson and Hanson [32, Chapter 20]. Specifically, for matrices

$$C \in \mathbb{M}_{k \times n}(\mathbb{C}),$$

$$E \in \mathbb{M}_{\ell \times n}(\mathbb{C}),$$

$$d \in \mathbb{C}^k,$$

$$f \in \mathbb{C}^\ell,$$

the problem consists in determining a vector $x \in \mathbb{C}^n$ that minimizes

$$\|Ex - f\|_2$$

subject to the constraint

$$Cx = d.$$

The strategy for solving such a problem uses an orthonormal basis $(q_1, \ldots, q_k; q_{k+1}, \ldots, q_n)$, where $(q_{k+1}, \ldots, q_n)$ is an orthonormal basis on the null space of C. The basis $(q_{k+1}, \ldots, q_n)$ provides a parametrization of the solution space of the system $Cx = d$, which reduces the problem to an unconstrained least squares problem in the subspace of $\mathbb{C}^n$ spanned by $(q_1, \ldots, q_k)$.

In the generic situation where C has k linearly independent rows and E has n linearly independent columns, C^* factors in the form

$$C^* = QR,$$

$$C = LQ^*,$$

where $Q \in \mathbb{M}_{n \times n}(\mathbb{C})$ is unitary, $R \in \mathbb{M}_{n \times k}(\mathbb{C})$ is upper triangular, and $L = R^*$ is lower triangular with linearly independent rows. Because $R = Q^* C^*$ has only zeros below the kth row, it follows that in Q^* all the rows $q_{k+1}^*, \ldots, q_n^*$ are perpendicular to all the columns of C^*. Hence, the rows $q_1^*, \ldots, q_k^*$ span the column space of C^*. Thus Q performs the required change of basis. With

$$w := Q^* x,$$

the system becomes

$$\begin{pmatrix} L \\ EQ \end{pmatrix} \begin{pmatrix} w_1 \\ w_2 \end{pmatrix} = \begin{pmatrix} C \\ E \end{pmatrix} Q(Q^* x) = \begin{pmatrix} d \\ f \end{pmatrix}.$$

Therefore, there exists exactly one solution $w_1 \in \mathbb{C}^k$ to the system

$$L \begin{pmatrix} w_1 \\ 0 \end{pmatrix} = d.$$

The initial problem thus reduces to determining $w_2 \in \mathbb{C}^\ell$ minimizing

$$\left\| \begin{pmatrix} L \\ EQ \end{pmatrix} \begin{pmatrix} w_1 \\ w_2 \end{pmatrix} - \begin{pmatrix} d \\ f \end{pmatrix} \right\|_2 = \left\| EQ \begin{pmatrix} 0 \\ w_2 \end{pmatrix} - \left[f - EQ \begin{pmatrix} w_1 \\ 0 \end{pmatrix} \right] \right\|_2.$$

The following application uses the Gauss-Bonnet Theorem.

Theorem 6 (Gauss-Bonnet). *Let D be a compact oriented domain with Euler characteristic χ on a Riemannian surface M in $\mathbb{R}^3$. Let $C = \partial D$ be the boundary of D in M, and let $\alpha_1, \ldots, \alpha_L$ be the oriented internal angles at the vertices (if any) of C. Moreover, let K be the Gaussian curvature of M, and let k_g be the geodesic curvature of C. Then*

$$\sum_{\ell=1}^{L} (\alpha_l - \pi) = \int \int_D K \, d\sigma + \int_C k_g \, ds - 2\pi\chi.$$

For a proof, see Chern's book [5, pp. 125–126].

For a triangle $D = \Delta$, with $v = 3$ vertices, $s = 3$ sides, and $f = 1$ facet, $\chi_\Delta = v - s + f = 1$. If each side lies on a geodesic on M, then $k_g = 0$, whence

$$\sum_{\ell=1}^{3} \alpha_\ell = \pi + \int \int_D K \, d\sigma.$$

Example 7. With the geodetic latitude λ and the longitude φ, the parametrization of the spheroidal earth surface takes the form

$$\rho = \frac{a \cos(\lambda)}{\sqrt{1 - e^2[\sin(\lambda)]^2}},$$

$$x = \rho \cos(\varphi),$$

$$y = \rho \sin(\varphi),$$

$$z = \frac{a\sigma^2 \sin(\lambda)}{\sqrt{1 - e^2[\sin(\lambda)]^2}}.$$

Hence, calculus gives the surface area

$$d\sigma = \frac{a^2\sigma^2 \cos(\lambda)}{\{1 - e^2[\sin(\lambda)]^2\}^2} d\varphi \, d\lambda$$

and the Gaussian curvature

$$K(\varphi, \lambda) = \frac{\{1 - e^2[\sin(\lambda)]^2\}^2}{a^2\sigma^2},$$

which is the reciprocal of the product of the radii of curvature R' in the plane of the meridian and N in the perpendicular plane [44, pp. 24, 25]:

$$R' = \frac{a(1 - e^2)}{\{1 - e^2[\sin(\lambda)]^2\}^{3/2}}, \quad N = \frac{a}{\{1 - e^2[\sin(\lambda)]^2\}^{1/2}}.$$

Thus, with Ω being the domain of the parametrization of Δ,

$$\int \int_\Delta K \, d\sigma = \int \int_\Omega \frac{\{1 - e^2[\sin(\lambda)]^2\}^2}{a^2\sigma^2} \frac{a^2\sigma^2 \cos(\lambda)}{\{1 - e^2[\sin(\lambda)]^2\}^2} d\varphi \, d\lambda$$

$$= \int \int_\Omega \cos(\lambda) \, d\varphi \, d\lambda.$$

Example 8. Gauss investigated triangulations measured by De Krayenhof, for instance, the following internal angles of a spheroidal triangle [16, Section 23, p. 149]:

$$\alpha = 50°58'15.238'' \quad \text{at Harlingen,}$$

$$\beta = 82°47'15.351'' \quad \text{at Leeuwarden,}$$

$$\gamma = 46°14'27.202'' \quad \text{at Ballum.}$$

In the plane, no such triangle exists, because the sum of the three angles $\alpha + \beta + \gamma = 179°59'57.791''$ fails to equal $180°$. On an ellipsoid, the sum of the internal angles in a geodesic triangle Δ exceeds $180°$ by the integral of the Gaussian curvature K over the triangle, which Gauss computed to be $1.749''$ for this example, so that $\alpha + \beta + \gamma = 180°0'1.749''$. In either case, it is impossible to place the three cities on a map without altering the data. One strategy consists in making the "smallest" adjustment while preserving $\alpha + \beta + \gamma = 180°$, in other words, minimizing

$$
\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \\ \gamma \end{pmatrix} - \begin{pmatrix} 50°58'15.238'' \\ 82°47'15.351'' \\ 46°14'27.202'' \end{pmatrix}
$$

subject to the linear constraint

$$
(1 \quad 1 \quad 1) \begin{pmatrix} \alpha \\ \beta \\ \gamma \end{pmatrix} = 180° + \frac{180°}{\pi} \int\int_\Delta K \, d\sigma = 180°0'1.749''.
$$

More generally, with n measurements $f_1, \ldots, f_n$ of quantities $\alpha_1, \ldots, \alpha_n$ subject to a constraint $\alpha_1 + \cdots + \alpha_n = d$, the system becomes

$$
\begin{pmatrix} 1^* \\ I \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_n \end{pmatrix} = \begin{pmatrix} d \\ f_1 \\ \vdots \\ f_n \end{pmatrix}.
$$

For the unitary factorization of the constraint equation,

$$
r = C^*(\ ; 1) = 1^* = (1, \ldots, 1)^*,
$$

$$
v = r + \|r\|_2 e_1 = \begin{pmatrix} \sqrt{n} + 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix},
$$

$$
\nu = \frac{1}{\|r\|_2(\|r\|_2 + |r_1|)} = \frac{1}{\sqrt{n}(\sqrt{n} + 1)},
$$

$$
S = I - \nu v v^*,
$$

$$
L = CS^* = -\sqrt{3} 1^* = (-\sqrt{n}, 0, 0) = -\sqrt{n} e_1.
$$

Consequently,

$$
L w_1 = d,
$$

$$
-\sqrt{n} w_1 = d,
$$

$$w_1 = \frac{d}{-\sqrt{n}} = \frac{180°0'1.749''}{-\sqrt{3}} = -103°55'33.072''.$$

For the least-squares system, $E = I$. Consequently,

$$EQ = IS = S = (q_1; q_2, q_3, \ldots q_n),$$

and the least-squares system takes the form

$$(q_1;\ q_2\ q_3\ \cdots\ q_n)\begin{pmatrix} 0 \\ w_2 \end{pmatrix} = f - q_1 w_1.$$

Hence,

$$\begin{pmatrix} 0 \\ w_2 \end{pmatrix} = \begin{pmatrix} q_1^*; \\ q_2^* \\ q_3^* \\ \vdots \\ q_n^* \end{pmatrix} (f - q_1 w_1).$$

Because $q_j^* q_1 = 0$ for every $j > 1$, the least-squares solution is

$$w_2 = \begin{pmatrix} q_2^* \\ q_3^* \\ \vdots \\ q_n^* \end{pmatrix} (f - q_1 w_1) = \begin{pmatrix} q_2^* \\ q_3^* \\ \vdots \\ q_n^* \end{pmatrix} (f),$$

and the first coordinate (in this example) gives the least-squares error

$$\|Ex - f\|_2 = q_1^*(f - q_1 w_1) = q_1^* f - w_1 = -(1/\sqrt{n})1^* f + d/\sqrt{n}$$

$$= \frac{1}{\sqrt{n}}\left[180° + \frac{180°}{\pi}\int\int_A K d\sigma - (f_1 + f_2 + \cdots + f_n)\right].$$

Reverting to the canonical basis through the inverse change of basis gives the solution

$$x = Sw = (q_1; q_2, q_3, \ldots, q_n)\begin{pmatrix} w_1 \\ w_2 \end{pmatrix}$$

$$= (w_1 - q_1^* f)q_1 + (q_1; q_2, q_3, \ldots, q_n)\left(\begin{pmatrix} q_1^* f \\ q_2^* \\ q_3^* \vdots \\ q_n^* \end{pmatrix}(f)\right)$$

$$= (w_1 - q_1^* f)q_1 + f$$

$$= \begin{pmatrix} 50^\circ 58' 16.557\,333'' \\ 82^\circ 47' 16.670\,333'' \\ 46^\circ 14' 28.521\,333'' \end{pmatrix} = \begin{pmatrix} \tilde{\alpha} \\ \tilde{\beta} \\ \tilde{\gamma} \end{pmatrix},$$

which add to $180^\circ 0' 1.749''$. The formula

$$x = (w_1 - q_1^* f)q_1 + f = \left(\frac{d}{\sqrt{n}} - \frac{1^* f}{\sqrt{n}} \right) \frac{1}{\sqrt{n}} 1 + f$$

shows that the measurements f are all adjusted by the average discrepancy

$$\frac{d - 1^* f}{n} = \frac{180^\circ 0' 1.749'' - 179^\circ 59' 57.791''}{3} = \frac{0^\circ 0' 3.958''}{3} = 1.319\,333\ldots''.$$

5. The singular-value decomposition and error analysis

5.1. The singular-value decomposition

Ordinary least-squares problems consist in determining the shortest vector $\tilde{x}$ that minimizes $\|A\tilde{x} - b\|_2$, perhaps subject to linear constraints. If $\tilde{b} := A\tilde{x}$, then the solution minimizes the discrepancy in the right-hand side, $\|\tilde{b} - b\|_2$, but it does *not* adjust the matrix A. In other words, $\tilde{x} = (\tilde{x}_1, \ldots, \tilde{x}_n)$ is the gradient of the linear function $u : \mathbb{C}^n \to \mathbb{C}$ that minimizes the average squared discrepancy between the measurement b_j and the value $u[A(j;\,)]$, but it does *not* minimize the Euclidean distance from the graph of u (a hyperplane) to the data $(a_{j,1}, \ldots, a_{j,n}; b_j)$ in $\mathbb{C}^{n+1}$. Such more general problems of least squares admit solutions in term of a matrix factorization called the "singular-value decomposition" that was published independently by Eugenio Beltrami in 1873 and Camille Jordan in 1874, and extended by Erhard Schmidt in 1907 and Hermann Weyl in 1912. (For greater detail on the history of the singular value decomposition consult Stewart's account [45].)

Theorem 9. *For each matrix $A \in \mathbb{M}_{m \times n}(\mathbb{C})$ of rank r, there exist unitary matrices U, V, and a diagonal matrix Σ, such that*

$$A = U \Sigma V^*$$

$$= u_1 \sigma_1 v_1^* + \cdots + u_r \sigma_r v_r^*$$

$$= \tilde{U} \tilde{\Sigma} \tilde{V}^*$$

with $\sigma_j := \Sigma_{j,j}$ and with the following features.

(U) *The matrix $U \in \mathbb{M}_{m \times m}(\mathbb{C})$ is unitary. The first r columns $(u_1, \ldots, u_r)$ of U form an orthonormal basis for the range (column space) of A. The last $m - r$ columns $(u_{r+1}, \ldots, u_m)$ of U form an orthonormal basis for the null space (kernel) of A^*.*

(V) *The matrix $V \in \mathbb{M}_{n \times n}(\mathbb{C})$ is unitary. The first r columns $(v_1, \ldots, v_r)$ of V form an orthonormal basis for the row space of A ($[\mathrm{Kernel}(A)]^\perp$). The last $n - r$ columns $(v_{r+1}, \ldots, v_n)$ of V form an orthonormal basis for the null space (kernel) of A.*

(Σ) *The matrix $\Sigma \in \mathbb{M}_{m \times n}(\mathbb{C})$ is diagonal: $\Sigma_{k,\ell} = 0$ for all $k \neq \ell$, with*

$$\sigma_1 \geqslant \sigma_2 \geqslant \cdots \geqslant \sigma_r > 0 = \sigma_{r+1} = \cdots = \sigma_{\min\{m,n\}}.$$

Moreover,

$$Av_j = \sigma_j u_j,$$

$$A^* u_j = \sigma_j v_j,$$

for every $j \in \{1, \ldots, r\}$, *and* $Av_j = 0$ *for every* $j \in \{r+1, \ldots, n\}$. *Finally,*

$$\tilde{U} = (u_1, \ldots, u_r) \in \mathbb{M}_{m \times r}(\mathbb{C}),$$

$$\tilde{V} = (v_1, \ldots, v_r) \in \mathbb{M}_{n \times r}(\mathbb{C}),$$

$$\tilde{\Sigma} = \text{diagonal}(\sigma_1, \ldots, \sigma_r) \in \mathbb{M}_{r \times r}(\mathbb{C}).$$

Proof. Let $V = (v_1, \ldots, v_r, v_{r+1}, \ldots, v_n)$ be an orthonormal basis of eigen vectors for the hermitian positive semi-definite matrix $A^*A \in \mathbb{M}_{n \times n}(\mathbb{C})$, corresponding to its eigenvalues in nondecreasing order $\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_r > 0 = \lambda_{r+1} = \cdots = \lambda_n$. Define $\sigma_j := \sqrt{\lambda_j}$, and $u_j := (1/\sigma_j)Av_j$ for every $j \in \{1, \ldots, r\}$. The remainder of the proof consists of straightforward verifications [30, Section 5.4; 46, Section 6.4]. $\square$

Definition 10. The factorization $A = U\Sigma V^*$ is the *singular-value decomposition* of A. The scalars $\sigma_1, \ldots, \sigma_r$ are the *singular values* of A. The vectors $v_1, \ldots, v_n$ are the *right singular vectors* of A. The vectors $u_1, \ldots, u_m$ are the *left singular vectors* of A.

The singular-value decomposition also provides a means to solve ordinary least-squares problems. Firstly, the product $\tilde{U}^* b$ projects b orthogonally on the column space of A, whence

$$\|Ax - \tilde{U}^* b\|_2 \leqslant \|Ax - b\|_2$$

for every $x \in \mathbb{C}^n$. Because $\tilde{U}^* b$ lies in the column space of A, there exists a solution $x \in \mathbb{C}^n$ such that $Ax = \tilde{U}^* b$. Secondly, every solution to this system differs from x by a vector in the null space of A. Consequently, the shortest solution is the orthogonal projection $x^\dagger = \tilde{V}^* x$ of x on the orthogonal complement of the null space of A. A derivation of a formula for $x^\dagger$ can proceed as follows:

$$Ax = b,$$

$$(\tilde{U}\tilde{\Sigma}\tilde{V}^*)x = b,$$

$$\tilde{\Sigma}(\tilde{V}^* x) = \tilde{U}^* b,$$

$$\tilde{V}^* x = \tilde{\Sigma}^{-1}\tilde{U}^* b,$$

$$x^\dagger = (\tilde{V}\tilde{\Sigma}^{-1}\tilde{U}^*)b.$$

Definition 11. The *pseudoinverse* of A is the matrix

$$A^\dagger := \tilde{V}\tilde{\Sigma}^{-1}\tilde{U}^*.$$

Thus, the shortest least-squares solution of $Ax = b$ is $x^\dagger := A^\dagger b$.

5.2. Norms and condition numbers of matrices

5.2.1. Norms of matrices

The concepts of norms and condition numbers for matrices provide means to estimate the propagation of errors from the data and during computations through the solutions of linear systems, as developed by Gastinel [15].

Definition 12. For each norm $\| \ \|_p$ on $\mathbb{C}^n$ and each norm $\| \ \|_q$ on $\mathbb{C}^m$, the *subordinate matrix norm* $\| \ \|_{p,q}$ on $\mathbb{M}_{m \times n}(\mathbb{C})$ is defined by

$$\|A\|_{p,q} := \max\{\|Au\|_q : u \in \mathbb{C}^n, \ \|u\|_p = 1\}$$

$$= \max\{\|Au\|_q / \|u\|_p : u \in \mathbb{C}^n, \ u \neq 0\}.$$

Example 13. With $\|x\|_\infty := \max_j |x_j|$ on $\mathbb{C}^n$ and $\|Ax\|_\infty$ on $\mathbb{C}^m$,

$$\|A\|_{\infty,\infty} = \max_{1 \leqslant i \leqslant m} \sum_{j=1}^{n} |A_{i,j}|.$$

With $\|x\|_1 := \sum_j |x_j|$ on $\mathbb{C}^n$ and $\|Ax\|_1$ on $\mathbb{C}^m$,

$$\|A\|_{1,1} = \max_{1 \leqslant j \leqslant n} \sum_{i=1}^{m} |A_{i,j}|.$$

With $\|x\|_1$ on $\mathbb{C}^n$, and $\|Ax\|_\infty$ on $\mathbb{C}^m$,

$$\|A\|_{1,\infty} = \max_{1 \leqslant i \leqslant m} \max_{1 \leqslant j \leqslant n} |A_{i,j}|.$$

(For $p \in \{1, \infty\}$ the formulae for $\|x\|_p$ and $\|A\|_{p,p}$ coincide.)

The following considerations show that $\|A\|_{2,2} = \sigma_1$ and $\kappa_{2,2}(A) = \sigma_1/\sigma_n$ is the ratio of the largest to the smallest singular values of A.

Lemma 14. *For all real numbers $\sigma_1 \geqslant \sigma_2 \geqslant \cdots \geqslant \sigma_{n-1} \geqslant \sigma_n \geqslant 0$,*

$$\min_{\|x\|_2=1} \sum_{i=1}^{n} (\sigma_i x_i)^2 = \min_{i \in \{1,\dots,n\}} \sigma_i^2 = \sigma_n^2,$$

$$\max_{\|x\|_2=1} \sum_{i=1}^{n} (\sigma_i x_i)^2 = \max_{i \in \{1,\dots,n\}} \sigma_i^2 = \sigma_1^2.$$

Proof. Solving $\sum_{i=1}^{n} x_i^2 = 1$ for x_n^2 gives $x_n^2 = 1 - \sum_{i=1}^{n-1} x_i^2$. Hence

$$\sum_{i=1}^{n} (\sigma_i x_i)^2 = \sigma_n^2 x_n^2 + \sum_{i=1}^{n-1} (\sigma_i x_i)^2$$

$$= \sigma_n^2 \left(1 - \sum_{i=1}^{n-1} x_i^2 \right) + \sum_{i=1}^{n-1} \sigma_i^2 x_i^2$$

$$= \sigma_n^2 + \sum_{i=1}^{n-1} (\sigma_i^2 - \sigma_n^2) x_i^2$$

$$\geqslant \sigma_n^2,$$

with equality if and only if $x_i = 0$ for $\sigma_i \neq \sigma_n$. Similarly, $\sum_{i=1}^{n}(\sigma_i x_i)^2 = \sigma_1^2 + \sum_{i=2}^{n}(\sigma_i^2 - \sigma_1^2)x_i^2 \leqslant \sigma_1^2$ with equality if and only if $x_i = 0$ for $\sigma_i \neq \sigma_1$.

Proposition 15. *For each matrix $A \in \mathbb{M}_{m \times n}(\mathbb{C})$, the subordinate Euclidean norm $\|A\|_{2,2} = \sigma_1$ is the largest singular value of A. Moreover, $\kappa_{2,2}(A) = \sigma_1/\sigma_n$ is the ratio of the largest to the smallest singular values of A.*

Proof. Consider a singular-value decomposition $A = U\Sigma V^*$. For each vector $x \in \mathbb{C}^n$ with $\|x\|_2 = 1$, let $w := V^*x$. Then $\|w\|_2 = \|x\|_2 = 1$. Hence,

$$\|Ax\|_2^2 = x^*A^*Ax = x^*(V\Sigma^*U^*)(U\Sigma V^*)x$$

$$= x^*V\Sigma^*\Sigma V^*x = \|\Sigma w\|_2^2 = \sum_{i=1}^{n} (\sigma_i w_i)^2 \leqslant \sigma_1^2$$

with the maximum value reached for $w = e_1$, or, equivalently, $x = v_1$. Hence, $\kappa_2(A) = \|A\|_2\|A^{-1}\|_2 = (\max_i \sigma_i)(\max_i \sigma_i^{-1}) = \sigma_1/\sigma_n$. $\square$

For norms of the type $\|x\|_p := (|x_1|^p + \cdots + |x_n|^p)^{1/p}$ with $p, q \notin \{1, 2, \infty\}$, no formula for the subordinate norm $\|A\|_{p,q}$ seems to be known [22, p. 124].

5.2.2. Condition numbers of matrices

For a square and invertible matrix $A \in \mathbb{M}_{n \times n}(\mathbb{C})$, the condition number provides lower and upper bounds on the discrepancy $\|\tilde{x} - x\|_p$ between the solution x of a linear system $Ax = b$ and any vector $\tilde{x}$. Such a vector $\tilde{x}$ can result, for instance, from an attempt at solving the system with floating-point or any other approximate arithmetic. To this end, let $\tilde{b} := A\tilde{x}$.

Definition 16. *For each norm $\|\ \|_p$ on $\mathbb{C}^n$ and each norm $\|\ \|_q$ on $\mathbb{C}^m$, the condition number $\kappa_{p,q}$ is defined by*

$$\kappa_{p,q}(A) := \|A\|_{p,q}\|A^\dagger\|_{q,p}.$$

Proposition 17. *For all b, $\tilde{b}$, x, $\tilde{x}$ and A invertible with $Ax = b$ and $A\tilde{x} = \tilde{b}$,*

$$\frac{1}{\kappa_{p,q}(A)} \frac{\|\tilde{b} - b\|_q}{\|b\|_q} \leqslant \frac{\|\tilde{x} - x\|_p}{\|x\|_p} \leqslant \kappa_{p,q}(A) \frac{\|\tilde{b} - b\|_q}{\|b\|_q}.$$

Proof. Use $\|b\|_q = \|Ax\|_q \leqslant \|A\|_{p,q} \cdot \|x\|_p$ and $\|\tilde{x} - x\|_p = \|A^{-1}(\tilde{b} - b)\|_p \leqslant \|A^{-1}\|_{q,p} \cdot \|\tilde{b} - b\|_q$ [30, Section 4.4; 46, Section 4.4]. $\square$

Proposition 17 compares the solutions x and $\tilde{x}$ of two systems with right-hand sides b and $\tilde{b}$ but with the same matrix A. In contrast, with different matrices A and C the following result holds. For each invertible $A \in \mathbb{M}_{n \times n}(\mathbb{C})$, for each $C \in \mathbb{M}_{n \times n}(\mathbb{C})$. If $\|A - C\| < 1/\|A^{-1}\|$, then for each nonzero vector $b \in \mathbb{C}^n$ and for the solutions $x \in \mathbb{C}^n$ of $Ax = b$ and $w \in \mathbb{C}^n$ of $Cw = b$,

$$\frac{\|w - x\|}{\|x\|} \leqslant \frac{\kappa(A)}{1 - \kappa(A) \cdot \|A - C\|/\|A\|} \cdot \frac{\|A - C\|}{\|A\|}.$$

For a proof see [46, pp. 188–198], and for other similar error bounds see [22, Chapter 7]. Yet more generally, a theorem of Wedin for all matrices $A \in \mathbb{M}_{m \times n}(\mathbb{C})$ and $C \in \mathbb{M}_{m \times n}(\mathbb{C})$, with rank $r = n \leqslant m$, and for all vectors $b \in \mathbb{C}^m$ and $d \in \mathbb{C}^m$, if there exists a positive real ε for which

$$\kappa_2(A)\varepsilon < 1,$$

$$\|A - C\|_2 \leqslant \varepsilon \|A\|_2,$$

$$\|b - d\|_2 \leqslant \varepsilon \|b\|_2,$$

then the *least-squares* solutions $\tilde{x} \in \mathbb{C}^n$ and $\tilde{z} \in \mathbb{C}^n$ of the systems $Ax = b$ and $Cz = d$ satisfy the following inequalities [22, Chapter 19]:

$$\frac{\|x - z\|_2}{\|x\|_2} \leqslant \frac{\kappa_2(A)\varepsilon}{1 - \kappa_2(A)\varepsilon} \left\{ 2 + [1 + \kappa_2(A)] \frac{\|b - Ax\|_2}{\|A\|_2 \cdot \|x\|_2} \right\},$$

$$\frac{\|(b - Ax) - (d - Az)\|_2}{\|b\|_2} \leqslant [1 + 2\kappa_2(A)]\varepsilon.$$

The following theorem of Kahan [25, pp. 775,776], who credits Gastinel, shows that for each invertible matrix A the distance to the closest singular matrix is $1/\|A^{-1}\|$.

Theorem 18. *For every invertible matrix A and every subordinate norm*:

$$\min_{\det(S)=0} \|A - S\| = \frac{1}{\|A^{-1}\|}.$$

Proof. For each singular matrix S there exists a vector $z \neq 0$ with $Sz = 0$:

$$\|A - S\| \geqslant \frac{\|(A - S)z\|}{\|z\|} = \frac{\|Az\|}{\|z\|} = \frac{\|A^{-1}\|\|Az\|}{\|A^{-1}\|\|z\|} \geqslant \frac{\|A^{-1}Az\|}{\|A^{-1}\|\|z\|} = \frac{1}{\|A^{-1}\|}.$$

There exists a vector $y \neq 0$ with $\|A^{-1}y\| = \|A^{-1}\|\|y\|$. As in the Hahn–Banach theorem [48], choose a linear functional w dual to $A^{-1}y$, so that

$$w(A^{-1}y) = \|w\| \cdot \|A^{-1}y\| = 1,$$

let w^* be the matrix of w relative to the canonical basis, so that $w(z) = w^*z$ for every vector z, and define

$$S := A - yw^*.$$

Then S is singular, because

$$S(A^{-1}y) = (A - yw^*) \cdot (A^{-1}y) = y - y \cdot 1 = 0.$$

Moreover,

$$\|A - S\| = \max\{\|(yw^*)x\|: \|x\| = 1\}$$

$$= \max\{\|y(w^*x)\|: \|x\| = 1\}$$

$$= \|y\| \cdot \max\{w^*x: \|x\| = 1\}$$

$$= \|y\| \cdot \|w^*\|$$

$$= \|y\| \cdot \frac{1}{\|A^{-1}y\|}$$

$$= \|y\| \cdot \frac{1}{\|A^{-1}\| \cdot \|y\|}$$

$$= \frac{1}{\|A^{-1}\|}. \qquad \square$$

6. Matrix approximation and total least squares

6.1. The approximation theorems of Schmidt, Mirsky, and Weyl

A theorem of Schmidt [43], with later versions by Mirsky [34] and Weyl [52], approximates a matrix $C \in \mathbb{M}_{m \times n}(\mathbb{C})$ of rank r by a singular matrix $S \in \mathbb{M}_{m \times n}(\mathbb{C})$ of rank $s < r$ that minimizes the Frobenius norm $\|C - S\|_F$, defined by

$$\|A\|_F^2 := \sum_{i=1}^m \sum_{j=1}^n |A_{i,j}|^2 = \sum_{i=1}^m \|A(i;)\|_2^2 = \sum_{j=1}^n \|A(; j)\|_2^2.$$

All unitary matrices U and V preserve Euclidean and Frobenius norms:

$$\|UA\|_F^2 = \sum_{j=1}^n \|UA(; j)\|_2^2 = \sum_{j=1}^n \|A(; j)\|_2^2 = \|A\|_F,$$

$$\|AV\|_F = \sum_{i=1}^m \|A(i;)V\|_2^2 = \sum_{i=1}^m \|A(i;)\|_2^2 = \|A\|_F.$$

In particular, with a singular-value decomposition $A = U\Sigma V^*$,

$$\|A\|_F^2 = \|U\Sigma V^*\|_F^2 = \|\Sigma\|_F^2 = \sum_{i=1}^r \sigma_r^2 \geqslant \sigma_1^2 = \|A\|_2^2.$$

The following theorem follows Stewart's version [45, pp. 561, 562].

Theorem 19. *For each matrix $C \in \mathbb{M}_{m \times n}(\mathbb{C})$ with $\sigma_1 \geqslant \cdots \geqslant \sigma_r > 0$ and*

$$C = \sum_{i=1}^{r} \sigma_i u_i v_i^*$$

and for each matrix $S \in \mathbb{M}_{m \times n}(\mathbb{C})$ of rank $k \in \{0, \ldots, r\}$,

$$\|C - S\|_F^2 \geqslant \sum_{i=k+1}^{r} \sigma_i^2,$$

with the minimum $\sigma_{k+1}^2 + \cdots + \sigma_r^2$ reached for

$$S = \sum_{i=1}^{k} \sigma_i u_i v_i^*.$$

Proof. If $k = r$ then the theorem holds because $S = C$. Henceforth, assume that $k < r$. Also, for each matrix A, let $\sigma_i(A)$, $u_i(A)$, and $v_i(A)$ be the ith singular value and singular vectors of A, and define

$$A_k := \sum_{i=1}^{k} \sigma_i(A) u_i(A) v_i^*(A).$$

The following argument shows that $\sigma_1(C - S) \geqslant \sigma_{k+1}(C)$. If S has rank k then S has a singular value decomposition $S = \sum_{i=1}^{k} \tau_i w_i z_i^* = W \tau Z^*$. Moreover, the linear space $Z^\perp$ perpendicular to $z_1, \ldots, z_k$ has dimension $n - k > n - (k + 1) \geqslant n - r$. Because the column space V spanned by $v_1, \ldots, v_{k+1}$ has dimension $k + 1$, it follows that $Z^\perp \cap V \neq \{0\}$. Thus, there exists a non-zero vector of coefficients $\gamma \in \mathbb{C}^{k+1}$, for instance with $\|\gamma\|_2 = 1$, such that $x := V_\gamma = \sum_{i=1}^{k+1} \gamma_i v_i \in Z^\perp \cap V$, whence $0 = Zx$ and hence $Sx = 0$. Let $\tilde{\gamma} := (\gamma^*, 0^*)^* \in \mathbb{C}^n$:

$$\sigma_1^2(C - S) \geqslant x^*(C - S)^*(C - S)x$$

$$= x^* C^* C x$$

$$= \tilde{\gamma}^* V^* V \Sigma U^* U \Sigma V^* V \tilde{\gamma}$$

$$= \tilde{\gamma}^* \Sigma^2 \tilde{\gamma} = \sum_{i=1}^{k+1} (\gamma_i \sigma_i)^2$$

$$\geqslant \sigma_{k+1}^2.$$

The next argument provides an upper bound on the change in the largest singular value caused by a change in a matrix. From the reverse triangle inequality for norms, it follows that

$$\sigma_1(A - B) = \|A - B\|_2 \geqslant |\|A\|_2 - \|B\|_2| = |\sigma_1(A) - \sigma_1(B)|.$$

The following generalization provides inequalities for the other singular values. For each matrix G and each index ℓ, $\sigma_1(G - G_\ell) = \sigma_{\ell+1}(G)$. Consequently, for all matrices $G, H \in \mathbb{M}_{m \times n}(\mathbb{C})$, and for all indices k and ℓ, the foregoing result leads to

$$\sigma_{\ell+1}(G) + \sigma_{k+1}(H)$$

$$= \sigma_1(G - G_\ell) + \sigma_1(H - H_k) \geqslant \sigma_1([G - G_\ell] + [H - H_k])$$

$$= \sigma_1([G+H] - [G_\ell + H_k])$$

$$\geqslant \sigma_{\ell+k+1}(G+H)$$

because the rank of $G_\ell + H_k$ cannot exceed $\ell + k$. Equivalently, if $A := G + H$ and $B := H$, then

$$\sigma_{\ell+1}(A-B) \geqslant \sigma_{\ell+k+1}(A) - \sigma_{k+1}(B)$$

Finally, in the particular case where S has rank k, setting $G := C - S$ and $H := S$ gives

$$\sigma_{\ell+1}(C-S) + 0 = \sigma_{\ell+1}(C-S) + \sigma_{k+1}(S) \geqslant \sigma_{\ell+1+k}(C)$$

Finally,

$$\|C - S\|_F^2 = \sum_{i=1}^r \sigma_i^2(C-S) \geqslant \sum_{i=1}^r \sigma_{i+k}^2(C) = (\sigma_{k+1}^2 + \cdots + \sigma_r^2)(C).$$

Equality holds with $S = \sum_{i=1}^k \sigma_i u_i v_i^*$, for which $\|C - S\|_F^2 = \|\sigma_{k+1} u_{k+1} v_{k+1}^* + \cdots + \sigma_r u_r v_r^*\|_F^2 = \sigma_{k+1}^2 + \cdots + \sigma_r^2$. $\qquad\square$

The approximation theorem of Schmidt, Mirsky, and Weyl amounts to identifying a matrix S minimizing a rotationally invariant norm $\|C - S\|$, for instance, the Euclidean norm, or the Frobenius norm, subject to the linear constraints $\sigma_{k+1}(S) = \cdots = \sigma_n(S) = 0$. There also exist other types of constraints, for example, with the vector of singular values $\sigma = (\sigma_1(S), \ldots, \sigma_n(S))$ subject to a linear system of constraints $K\sigma = d$ [37].

6.2. Total least squares

For a linear system $Ax = b$, the problem of *ordinary* least squares consists in determining the shortest vector $\tilde{x}$ that minimizes the Euclidean norm of the discrepancy between b and $\tilde{b} := A\tilde{x}$, possibly subject to constraints. In other words, the *ordinary* least-squares solution $\tilde{x}$ solves *exactly* a related linear system $A\tilde{x} = \tilde{b}$ with $\|\tilde{b} - b\|_2$ minimum. In contrast, the problem of *total* least squares allows for minimal adjustments not only of b but also of A, also possibly subject to constraints. The problem of total least squares admit several mutually equivalent mathematical formulations. Their solutions in terms of singular value decompositions was published in 1980 by Golub and Van Loan [18;19, pp. 576–581]. Van Huffel and Vandewalle's monograph [50] describes further extensions and applications.

6.2.1. Geometric formulations of total least squares

Geometrically, the problem of total least squares amounts to fitting a hyperplane H minimizing the average squared Euclidean distance (measured perpendicularly to the fitted hyperplane) to data points $c_1, \ldots, c_m$ in $\mathbb{C}^{n+1}$. The problem then reduces to finding a point $c_0 \in H$ and a non-zero normal vector $x \perp H$ that minimize the sum D of the squared distances:

$$D(x, c_0; c_1, \ldots, c_m) := \sum_{i=1}^m \frac{|\langle c_i - c_0, x \rangle|^2}{\langle x, x \rangle}.$$

To simplify notation, for every point $c_0 \in \mathbb{C}^{n+1}$, and for all data $c_1, \ldots, c_m$ in $\mathbb{C}^{n+1}$, define a matrix $C_{c_0} \in \mathbb{M}_{m \times (n+1)}(\mathbb{C})$ with ith row $c_i^* - c_0^*$:

$$C_{c_0} := \begin{pmatrix} c_1^* - c_0^* \\ \vdots \\ c_m^* - c_0^* \end{pmatrix}.$$

Consequently,

$$D(x, c_0; c_1, \ldots, c_m) = \frac{\|C_{c_0} x\|_2^2}{\|x\|_2^2}.$$

The following lemma reveals that an optimal hyperplane must pass through the centroid of the data,

$$\bar{c} = \frac{1}{m} \sum_{i=1}^{m} c_i,$$

which can thus serve as the point $c_0 \in H$.

Lemma 20. *For every normal vector $x \in \mathbb{C}^{n+1} \setminus \{0\}$, for every point $c_0 \in \mathbb{C}^{n+1}$, and for all data $c_1, \ldots, c_m$ in $\mathbb{C}^{n+1}$,*

$$D(x, c_0; c_1, \ldots, c_m) \geqslant D(x, \bar{c}; c_1, \ldots, c_m),$$

with equality if and only if $\langle x, (r - c_0) \rangle = \langle x, (r - \bar{c}) \rangle$ for every r. Consequently, a hyperplane of total least squares must pass through the centroid $\bar{c}$.

Proof. Consider the vector $w := C_{c_0} x$, so that $w_i = \langle c_i - c_0, x \rangle$ and

$$D(x, c_0; c_1, \ldots, c_m) = \frac{\|w\|_2^2}{\|x\|_2^2}.$$

Also, consider the vector $z := C_{\bar{c}} x$, so that $z_i = \langle c_i - \bar{c}, x \rangle$ and

$$D(x, \bar{c}; c_1, \ldots, c_m) = \frac{\|z\|_2^2}{\|x\|_2^2}.$$

Moreover, define $1 := (1, \ldots, 1) \in \mathbb{C}^m$, and $h := \langle x, (\bar{c} - c_0) \rangle$, so that

$$w = z + h1.$$

Then $z \perp 1$:

$$\langle z, 1 \rangle = 1^* (C_{\bar{c}} x) = (1^* C_{\bar{c}}) x = \left(m \bar{c}^* - \sum_{j=1}^{m} c_j^* \right) x = 0^* x = 0.$$

Finally, the Pythagorean Theorem applied to $z \perp 1$ and $w = z + h1$ gives

$$D(x, c_0; c_1, \ldots, c_m) = \|w\|_2^2 / \|x\|_2^2$$

$$= (\|z\|_2^2 + h^2 \|1\|_2^2) / \|x\|_2^2$$

$$= D(x, \bar{c}; c_1, \ldots, c_m) + h^2 m / \|x\|_2^2$$

$$\geqslant D(x, \bar{c}; c_1, \ldots, c_m),$$

with equality if and only if $0 = h = \langle x, (\bar{c} - c_0) \rangle$, which means that c_0 also lies in the hyperplane passing through $\bar{c}$ perpendicularly to x. $\square$

The following lemma reveals that an optimal normal vector must be a right-singular vector corresponding to the smallest singular value of C_{c_0}.

Lemma 21. *For every point $c_0 \in \mathbb{C}^{n+1}$ and all data $c_1, \ldots, c_m$ in $\mathbb{C}^{n+1}$, let v be a right-singular vector corresponding to the smallest singular value σ of C_{c_0}. Then for every vector $x \in \mathbb{C}^{n+1} \setminus \{0\}$, the following inequality holds:*

$$D(x, c_0; c_1, \ldots, c_m) \geqslant D(v, c_0; c_1, \ldots, c_m),$$

with equality if, but only if, x is also a right-singular vector corresponding to the smallest singular value σ of C_{c_0}. Consequently, a hyperplane of total least squares must be perpendicular to such a singular vector. Moreover,

$$D(v, c_0; c_1, \ldots, c_m) = \sigma^2.$$

Proof. From $D(x, c_0; c_1, \ldots, c_m) = \|C_{c_0} x\|_2^2 / \|x\|_2^2$ it follows that D reaches its minimum at a unit vector $v = x / \|x\|_2$ that minimizes $\|C_{c_0} v\|_2$. The theory of the SVD shows that v coincides with any singular vector v for to the smallest singular value σ of C_{c_0}, with $D(v, c_0; c_1, \ldots, c_m) = \|C_{c_0} v\|_2^2 = \sigma^2$. $\square$

Theorem 22. *For every set of data points $c_1, \ldots, c_m$ in $\mathbb{C}^{n+1}$, each hyperplane of total least-squares passes through the centroid of the data $\bar{c}$ perpendicularly to a right-singular vector v corresponding the smallest singular value σ of the matrix $C_{\bar{c}}$ with ith row $c_i^* - \bar{c}^*$. Moreover, for such a hyperplane, the sum of the squared distances to the data is σ^2:*

Proof. Combine the proofs of Lemmas 20 and 21. $\square$

The matrix C can have a multiple smallest singular value $\sigma = \sigma_{k+1} = \cdots = \sigma_{k+\ell}$, corresponding to a linear subspace $V_\sigma \subseteq \mathbb{C}^{n+1}$ spanned by multiple singular vectors $v_{k+1}, \ldots, v_{k+\ell}$. In this situation, there exists a set $\mathcal{H}$ of hyperplanes of total least squares, with each hyperplane $H \in \mathcal{H}$ perpendicular to a vector $v \in V$ and containing the "axis" $\bar{c} + V^\perp$. In particular, if $\sigma = 0$, then the data lies at the intersection $\cap \mathcal{H}$ of all such hyperplanes, which is an affine subspace $\bar{c} + V^\perp$ of dimension $n + 1 - \ell$. For example, if $n + 1 = 3$ and $\ell = 2$, then $n + 1 - \ell = 1$ and all the data points lie on a common straight line in space.

With $x = v$ and σ computed, the vector

$$\hat{c}_i := c_i - \langle c_i - \bar{c}, v \rangle v$$

is the orthogonal projection of the data c_i onto H. Consequently,

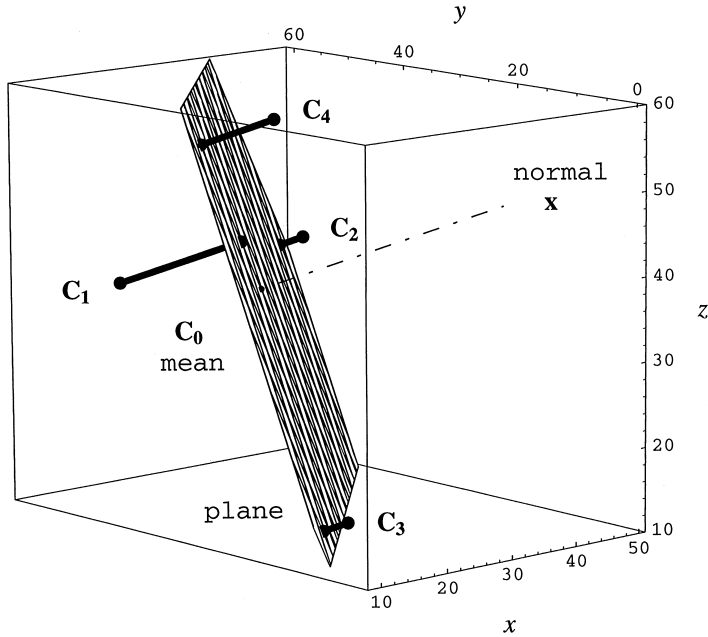$$\sum_{i=1}^{m} \|\hat{c}_i - c_i\|_2^2 = D(v, \bar{c}; c_1, \ldots, c_m) = \sigma^2.$$

Fig. 6. The TLS plane minimizes the Euclidean distance to the data.

Example 23. Consider the four data points in space displayed in Fig. 6:

$$c_1 = (11 \ 45 \ 38),$$

$$c_2 = (47 \ 54 \ 38),$$

$$c_3 = (17 \ 12 \ 14),$$

$$c_4 = (21 \ 29 \ 58).$$

For these data, $\bar{c} = (\frac{1}{4}) \sum_{i=1}^{4} c_i = (24, 35, 37)$, and

$$C_{\bar{c}} = \begin{pmatrix} c_1 - \bar{c} \\ c_2 - \bar{c} \\ c_3 - \bar{c} \\ c_4 - \bar{c} \end{pmatrix} = \begin{pmatrix} -13 & 10 & 1 \\ 23 & 19 & 1 \\ -7 & -23 & -23 \\ -3 & -6 & 21 \end{pmatrix}.$$

The smallest singular value of $C_{\bar{c}}$ is $\sigma = \sigma_3 = 18$, and the corresponding singular vector is $v_3 = (\frac{2}{3}, -\frac{2}{3}, \frac{1}{3})^*$. Thus, the hyperplane H passes through $c_0 = \bar{c} = (24, 35, 37)$ and lies perpendicularly to the vector $x = v_3 = (\frac{2}{3}, -\frac{2}{3}, \frac{1}{3})^*$, so that H satisfies the equation

$$\tfrac{2}{3}(x - 24) - \tfrac{2}{3}(y - 35) + \tfrac{1}{3}(z - 37) = 0.$$

Moreover, $C_{\bar{c}}v = (-15, 3, 3, 9)^*$ contains the signed distances $d(c_i, H)$ from the data points to the hyperplane H, here $\sum_{i=1}^{4} d(c_i, H)^2 = \|C_{\bar{c}}v\|_2^2 = \sigma^2 = 18^2$, which gives the orthogonal projections

$\hat{c}_1, \ldots, \hat{c}_4$ of the data on H:

$$\hat{C} = \begin{pmatrix} \hat{c}_1 \\ \vdots \\ \hat{c}_m \end{pmatrix} = \begin{pmatrix} c_1 \\ \vdots \\ c_m \end{pmatrix} - \begin{pmatrix} -15 \\ 3 \\ 3 \\ 9 \end{pmatrix} \begin{pmatrix} \dfrac{2}{3} & -\dfrac{2}{3} & \dfrac{1}{3} \end{pmatrix} = \begin{pmatrix} 21 & 35 & 43 \\ 45 & 56 & 37 \\ 15 & 14 & 13 \\ 15 & 35 & 55 \end{pmatrix}.$$

6.2.2. Algebraic formulations of total least squares

For a linear system $Ax = b$, the problem of *ordinary* least squares consists in determining a vector $\tilde{b}$ that minimizes $\|\tilde{b} - b\|_2$ subject to the constraint that the system $A\tilde{x} = \tilde{b}$ have a solution. More generally, the algebraic formulation of the problem of *total* least squares consists in determining a *matrix* $\hat{A}$ and a vector $\hat{b}$ that minimize the Frobenius norm

$$\|[\hat{A}; \hat{b}] - [A; b]\|_F$$

subject to the condition that the system

$$\hat{A}\hat{x} = \hat{b}$$

have a solution. Thus, with

$$C := [A; b],$$

$$\hat{C} := [\hat{A}; \hat{b}],$$

the problem reduces to determining a matrix $\hat{C}$ that minimizes $\|\hat{C} - C\|_F$ subject to the condition that Kernel $(\hat{C})$ contains a vector of the form $(\hat{x}^*, -1)^*$.

If the matrix $C = [A; b]$ has a singular-value decomposition

$$C = \sum_{i=1}^{n+1} \sigma_i u_i v_i^*$$

and the $n + 1$ columns of C are linearly independent, then $\hat{C} = [\hat{A}, \hat{b}]$ must have rank at most n, and the Schmidt–Mirsky approximation theorem states that the closest such matrix is

$$\hat{C} = \sum_{i=1}^{n} \sigma_i u_i v_i^* = [A; b] - \sigma_{n+1} u_{n+1} v_{n+1}^*.$$

Then v_{n+1} spans Kernel $(\hat{C})$, and two cases arise. If $(v_{n+1})_{n+1} \neq 0$, then the problem admits the solution

$$\begin{pmatrix} \hat{x} \\ -1 \end{pmatrix} = \frac{-1}{(v_{n+1})_{n+1}} v_{n+1}.$$

If $(v_{n+1})_{n+1} = 0$, then the problem has no solution.

Some practical problems lead to problems of total least squares subject to the condition that the first k columns $A_1 := (A(; 1), \ldots, A(; k))$ of the matrix $A \in \mathbb{M}_{m \times n}(\mathbb{C})$ be kept exact (unadjusted) while the last $n - k$ columns $A_2 := (A(; k + 1), \ldots, A(; n))$ are subject to adjustments. Golub, Hoffman, and Stewart published the following method of solution [17]. If A_1 has rank $\ell \leqslant r$, if Q is the

orthogonal projection on the column space of A_1 and $Q^\perp$ is the orthogonal projection on its orthogonal complement, then the matrix $[A_1; \hat{A}_2]$ of rank r that minimizes the Frobenius norm

$$\|[A_1; \hat{A}_2] - [A_1; A_2]\|_F$$

is defined in terms of a singular-value decomposition of $Q^\perp A_2$,

$$Q^\perp A_2 = \sum_{i=1}^{n-k} \tau_i w_i z_i^*,$$

by the formula

$$\hat{A}_2 := Q A_2 + \sum_{i=1}^{r-\ell} \tau_i w_i z_i^*.$$

Example 24. To the data in Example 1, Laplace fitted an affine model for the length Δs of and arc of 1 grad along the meridian,

$$c_0 + c_1 [\sin(\lambda)]^2 = \Delta s.$$

The coefficient 1 of c_0 is exact. Consequently, the first column $A_1 := 1$ remains fixed. The orthogonal projection on the space spanned by $1 \in \mathbb{C}^m$ has matrix $Q := 1/m \, 11^*$. Thus, $Q^\perp = I - Q$ subtracts from each column the mean of that column. Here, the matrix $C := Q^\perp A_2$ corresponds to a linear system for c_1:

$$
\begin{pmatrix}
0.00\,000 - 0.43\,925 \\
0.30\,156 - 0.43\,925 \\
0.39\,946 - 0.43\,925 \\
0.46\,541 - 0.43\,925 \\
0.52\,093 - 0.43\,925 \\
0.54\,850 - 0.43\,925 \\
0.83\,887 - 0.43\,925
\end{pmatrix}
(c_1) =
\begin{pmatrix}
25\,538.85 - 25\,659.93 \\
25\,666.65 - 25\,659.93 \\
25\,599.60 - 25\,659.93 \\
25\,640.55 - 25\,659.93 \\
25\,658.28 - 25\,659.93 \\
25\,683.30 - 25\,659.93 \\
25\,832.25 - 25\,659.93
\end{pmatrix}.
$$

The singular-value decomposition of C (computed with the command SVD on the HP 48GX [21, pp. 14–22]) shows two singular values,

$$221.279 = \sigma_1 > \sigma_2 = 0.266\,719,$$

the smallest of which corresponds to the right singular vector

$$
\begin{pmatrix} \hat{x} \\ -1 \end{pmatrix} = \frac{-1}{(v_2)_2} v_2 = \frac{-1}{0.002\,562} \begin{pmatrix} -0.999\,997 \\ 0.002\,562 \end{pmatrix} = \begin{pmatrix} 390.356 \\ -1 \end{pmatrix}.
$$

Adding $Q A_2$ amounts to adding back the means, which yields

$$\hat{c}_1 = 390.356,$$

$$\hat{c}_0 = 25\,659.93 - 390.356 * 0.43\,925 = 25\,488.46.$$

These values lead to the estimate of the squared eccentricity $\hat{e}^2 = (\frac{2}{3})\hat{c}_1/\hat{c}_0 = 0.010\,210$, which is farther from the current estimate $e^2 = 0.006\,694\,379\,990\,13$ than the value obtained by ordinary least squares $\tilde{e}^2 = 0.006\,339$. The same values then lead to the estimate of the equatorial radius $\hat{a} = \hat{c}_0/(1-\hat{e}^2) = 99\,352$ m, which is closer to the current estimate $a = 6\,378\,137.000\,00$ m than the value obtained by ordinary least squares $\tilde{a} = 100\,170.25$ m, but still very inaccurate. The corresponding results for weighted total least squares, with $C = W[A;b]$, are $\hat{e}^2 = 0.008\,153$ and $\hat{a} = 100\,307$ m. Unweighted least squares, ordinary or total, give yet worse results.

Because the estimate of the eccentricity closest to the current estimate comes from ordinary least squares, rather than total least squares, such results suggest than most of the errors lie in the measurement of lengths along the meridian, rather than in the geodetic latitudes of the locations. Bowditch's comments corroborate such suggestions [31, Book II, Section 41].

For a system $AX = B$ with r right-hand sides, the problem of total least squares consists in determining matrices $\hat{A} \in \mathbb{M}_{m\times n}(\mathbb{C})$, $\hat{B} \in \mathbb{M}_{m\times r}(\mathbb{C})$, and $\hat{X} \in \mathbb{M}_{n\times r}(\mathbb{C})$, minimizing $\|[A;B] - [\hat{A};\hat{B}]\|_F$ subject to the constraint $\hat{A}\hat{X} = \hat{B}$. Equivalently, with $I \in \mathbb{M}_{r\times r}(\mathbb{C})$, there must exist a solution $[\hat{X}^*; -I]^*$ to the system $[\hat{A};\hat{B}][\hat{X}^*; -I]^* = O$. In particular, because the last r rows of $[\hat{X}^*; -I]^*$ are linearly independent, it follows that the rank of $S := [\hat{A};\hat{B}]$ cannot exceed $(n+r) - r = n$. Therefore, with

$$C := [A;B] = \sum_{i=1}^{n+r} \sigma_i u_i v_i^*$$

the matrix S must be a matrix of rank at most n that minimizes $\|C - S\|_F$. By the approximation theorem of Schmidt, Mirsky, and Weyl, it follows that

$$S = [\hat{A};\hat{B}] = \sum_{i=1}^{n} \sigma_i u_i v_i^*.$$

The problem then admits a solution in the form $[\hat{X}^*; -I]^*$ if and only if in the matrix $(v_{n+1}, \ldots, v_{n+r})$ the last $n+r$ rows are linearly independent and hence form in invertible matrix $V_{n+1;n+r} \in \mathbb{M}_{r\times r}(\mathbb{C})$, so that

$$\begin{pmatrix} X \\ -I \end{pmatrix} = V_{n+1;n+r}^{-1}(v_{n+1}, \ldots, v_{n+r}).$$

6.3. Relations between the algebraic and geometric formulations

The matrix $\hat{C} = [\hat{A};\hat{b}]$ in the algebraic formulation corresponds to the matrix C_0 in the geometric formulation. In other words, the algebraic formulation corresponds to the problem of fitting to the rows of $C = [A;b]$ a hyperplane constrained to pass through the origin $0 \in \mathbb{C}^{n+1}$ instead of through the centroid $\bar{c}$. Indeed, the system $\hat{A}\hat{x} = \hat{b}$ in the form

$$[\hat{A};\hat{b}]\begin{pmatrix} \hat{x} \\ -1 \end{pmatrix} = 0$$

states that every row of $[\hat{A}; \hat{b}]$ lies on the hyperplane H passing through the origin perpendicularly to $(\hat{x}^*; -1)^*$. Moreover, the condition that $\hat{C}$ minimizes the Frobenius norm, or, equivalently, its square,

$$\|\hat{C} - C\|_F^2 = \sum_{i=1}^{m} \sum_{j=1}^{n+1} (c_{i,j} - \hat{c}_{i,j})^2 = \sum_{i=1}^{m} \|c_i - \hat{c}_i\|_2^2,$$

shows that $\hat{C}$ minimizes the sum of the squared distances from the rows of C in $\mathbb{C}^{n+1}$ to the rows of $\hat{C}$ on H. Therefore, the ith row $\hat{C}(i;)$ of $\hat{C}$ is the orthogonal projection of the ith row $C(i;)$ of C, for otherwise these orthogonal projections would lie on the same hyperplane H and would give a smaller total least squares, or squared Frobenius norm. However, Lemma 21 shows that v_{n+1} and hence $(\hat{x}^*; -1)^*$ is the normal direction of total least squares for all the hyperplanes passing through the origin.

7. Nonlinear least squares

7.1. Nonlinear least squares in astronomy and geodesy

The old problem of estimating the shape of the earth can be formulated as the total least-squares problem of fitting an ellipsoid by minimizing the sum of the squared distances to data points. The problem of reliably computing the distance from a point to an ellipse already causes difficulties, because it amounts to solving a quartic equation, and there does not seem to be any practical forward error bounds for the solutions by the quartic formulae. Nevertheless, for small eccentricities ($e^2 < 2 - \sqrt{2}$), there exists a provably reliable algorithm to compute the distance with a contracting map [28,38].

Similarly, the old problem of estimating the shape of the orbit of a celestial body can be formulated as the total least-squares problem of fitting a plane and a conic section in it by minimizing the sum of the squared distances to data points [41].

For both problems, the particular formulation depends on the type of data, for instance, azimuth and elevation only, or azimuth, elevation, and range (measured by radar, for instance) [41, Chapter 10; 51, pp. 302–305]. Despite the practicality of such problems, however, there does not yet seem to exist any theorem that guarantees the global convergence of any algorithm toward the globally optimum surface or orbit [41, p. 180].

7.2. Fitting circles by total least squares

Although the problem of fitting circles and spheres to observations can be traced to the first millenium B.C., the problem of designing an *algorithm* to calculate the center and the radius of a circle or a sphere fitted to a finite set of points can be traced to the 1970's A.D., through computer scientists' developments of algorithms [8] for medical devices [3] and typography [40], electrical engineers' adjustments of microwave calibrations [27], and particle physicists' writings on fitting circular trajectories to a large number of automated measurements of positions of electrically charged particles within uniform magnetic fields [9]. One method — called an *algebraic fit* — to

fit a circle or a sphere to data points $c_1, \ldots, c_m \in \mathbb{R}^n$ consists in computing the center $x \in \mathbb{R}^n$ and the radius $r \in \mathbb{R}$ that minimize the function $f : \mathbb{R}^{n+1} \to \mathbb{R}$ defined by

$$f(x, r; c_1, \ldots, c_m) := \sum_{j=1}^{m} (\|x - c_j\|_2^2 - r^2)^2 = \sum_{j=1}^{m} (x^*x - r^2 - 2x^*c_j + c_j^*c_j)^2.$$

For each center $x \in \mathbb{R}^n$ the radius $r(x)$ that minimizes f is such that

$$[r(x)]^2 = \frac{1}{m} \sum_{j=1}^{m} \|x - c_j\|_2^2.$$

Substituting $r(x)$ for r in f leads to a linear system $Ax = b$ for the center x, where A is $8m$ times the covariance matrix of the data, and the vector b is defined by $b_i := \sum_{j=1}^{m} \{(c_j)_i - \bar{c}_i\} \|c_i\|_2^2$ [35]. With an approximate arithmetic, however, the computation of the entries of A and b can introduce errors.

An alternate method by Coope [7] performs the charge of coordinates

$$z := 2x, \quad z_{n+1} := r^2 - x^*x,$$

which leads to the following ordinary linear least-squares system $Cz = d$:

$$\begin{pmatrix} c_1^* & 1 \\ & \vdots & \\ c_m^* & 1 \end{pmatrix} \begin{pmatrix} z \\ z_{n+1} \end{pmatrix} = \begin{pmatrix} \|c_1\|_2^2 \\ \vdots \\ \|c_m\|_2^2 \end{pmatrix}.$$

Thus with Coope's method forming the matrix C does not involve any computations and hence does not introduce any computational approximation.

However, Gander, Golub, and Strebel have demonstrated with examples that curves fitted by minimizing such algebraic objectives as f can lie farther from the data than curves fitted by total least squares of the Euclidean distances, called *geometric fits* [14]. Therefore, several authors have suggested using algebraic fits as initial estimates to start iterative methods for the computation of geometric fits [1, p. 357; 7, 14]. Several problems remain unsolved yet.

Example 25. Consider the following data in the plane:

$$c_2 := (0, 2),$$

$$c_4 := (0, 0),$$

$$c_3 := (-\sqrt{3}, -1), \quad c_1 := (-\sqrt{3}, -1)$$

with the sum of the squared distances to the circle with center x and radius r:

$$g(x, r; c_1, \ldots, c_m) := \sum_{j=1}^{m} (\|x - c_j\|_2 - r)^2.$$

Firstly, calculus yields three minima for g, corresponding to circles with centers x_k opposite to c_k, radius $r := \frac{7}{4}$, and $g(x_k, r; c_1, \ldots, c_m) = 2$ for each k:

$$x_1 := e^{4\pi i/3} x_2, \qquad x_3 := e^{2\pi i/3} x_2,$$

$$x_2 := (0, -\tfrac{3}{4}).$$

Secondly, perturbations of any of the data can turn any of the local minima into a global minimum. Finally, the algebraic fit cannot serve as an initial estimate for Newton's methods to converge to any geometric fit. Indeed, the algebraically fitted circle has its center at the origin, which coincides with a data point, where the objective function is not differentiable.

7.3. Open problems in nonlinear least squares

Problems of fitting *affine* manifolds by minimizing a weighted sum of squared distances to data are extensively documented. Indeed, linear algebra yields affine parametrizations of their solutions and provides several methods of solution through orthogonal factorizations, for which there exist proven upper bounds on errors from the data or from approximate computations [1,19,22,32,50,51,53].

In contrast, problems of nonlinear least squares, for instance, problems of fitting nonaffine manifolds as simple as circles, remain mostly unsolved. There exist a substantial documentation of algorithms that converge globally (from every initial point) to a *local* minimum [12,29,42]. However, some of their shortcuts can succumb to rounding errors [36], and there does not yet seem to exist any theorem guaranteeing the convergence to a *global* minimum to fit curves as simple as conic sections and surfaces as simple as spheres.

References

[1] Å. Björk, Numerical Methods for Least Squares Problems, Society for Industrial and Applied Mathematics, Philadelphia, PA, 1996.

[2] A.C. Aiken, On least squares and linear combinations of observations, Proc. Roy. Soc. Edinburgh, Section A 55 (1934) 42–47.

[3] F.L. Bookstein, Fitting conic sections to scattered data, Comput. Graphics Image Process. 9 (1979) 56–71.

[4] L.A. Brown, The Story of Maps, Dover, New York, NY, 1979.

[5] S.-s. Chern, Curves and surfaces in Euclidean space, in: S.-s. Chern (Ed.), Global Differential Geometry, MAA Studies in Mathematics, Vol. 27, Mathematical Association of America, Washington, D.C, 1989, pp. 99–139.

[6] V. Chvátal, Linear Programming, W.H. Freeman, New York, NY, 1983.

[7] I.D. Coope, Circle fitting by linear and nonlinear least squares, J. Optim. Theory Appl. 76 (2) (1993) 381–388.

[8] M.G. Cox, H.M. Jones, An algorithm for least-squares circle fitting to data with specified uncertainty ellipses, IMA J. Numer. Anal. 9 (3) (1989) 285–298.

[9] J.F. Crawford, A non-iterative method for fitting circular arcs to measured points, Nucl. Instr. and Meth. 211 (2) (1983) 223–225.

[10] G.B. Dantzig, Inductive proof of the simplex method, IBM J. Res. Dev. 4 (1960) 505–506.

[11] G.B. Dantzig, Linear Programming and Extensions, Princeton University Press, Princeton, NJ, 1963.

[12] J.E. Dennis, Jr., R.B. Schnabel, Numerical Methods for Unconstrained Optimization and Nonlinear Equations, Classics In Applied Mathematics, Vol. 16, Society for Industrial and Applied Mathematics, Philadelphia, PA, 1996.

[13] J.L.E. Dreyer, A History of Astronomy from Thales to Kepler, 2nd Edition, Dover, New York, NY, 1953.

[14] W. Gander, G.H. Golub, R. Strebel, Least-squares fitting of circles and ellipses, BIT 34 (1994) 558–578.

[15] N. Gastinel, Matrices du Second Degré et Normes Générales en Analyse Numérique Linéaire, Thèse d'Etat, Université de Grenoble, 1960.
[16] C.F. Gauss, Theoria Combinationis Observationum Erroribus Minimis Obnoxiae, Society for Industrial and Applied Mathematics, Philadelphia, PA, 1995. (Theory of the Combination of Observations Least Subject to Errors, original with translation by Gilbert W. Stewart).
[17] G.H. Golub, A. Hoffman, G.W. Stewart, A generalization of the Eckart-Young-Mirsky matrix approximation theorem, Linear Algebra Appl. 88/89 (1987) 317–327.
[18] G.H. Golub, C.F.V. Loan, An analysis of the total least squares problem, SIAM J. Numer. Anal. 17 (6) (1980) 883–893.
[19] G.H. Golub, C.F.V. Loan, Matrix Computations, 2nd Edition, Johns Hopkins University Press, Baltimore, MD, 1989.
[20] N. Grossman, The Sheer Joy of Celestial Mechanics, Birkhäuser, Boston, Cambridge, MA, 1996.
[21] Hewlett-Packard Co., Corvallis Division, 1000 NE Circle Blvd., Corvallis, OR 97330, USA. HP 48G Series User's Guide, 7th edition, March 1994. (Part Number 00048-90126.).
[22] N.J. Higham, Accuracy and Stability of Numerical Algorithms, Society for Industrial and Applied Mathematics, Philadelphia, PA, 1996.
[23] B. Hofmann-Wellenhof, H. Lichtenegger, J. Collins, Global Positioning System: Theory and Practice, 4th revised Edition, Springer-Verlag, Wien, 1997.
[24] R.V. Hogg, A.T. Craig, Introduction to Mathematical Statistics, 4th Edition, Macmillan, New York, NY, 1978.
[25] W.M. Kahan, Numerical linear algebra, Canad. Math. Bull. 9 (6) (1966) 757–801.
[26] H. Karttunen, P. Kröger, H. Oja, M. Poutanen, K.J. Donner (Eds.), Fundamental Astronomy, 2nd enlarged Edition, Springer, New York, 1994.
[27] I. Kàsa, A circle fitting procedure and its error analysis, IEEE Trans. Instr. Meas. 25 (1) (1976) 8–14.
[28] S.P. Keeler, Y. Nievergelt, Computing geodetic coordinates, SIAM Rev. 40 (2) (1998) 300–309.
[29] C.T. Kelley, Iterative Methods for Linear and Nonlinear Equations, Frontiers in Applied Mathematics, Vol. 16, Society for Industrial and Applied Mathematics, Philadelphia, PA, 1995.
[30] D. Kincaid, W. Cheney, Numerical Analysis: Mathematics of Scientific Computing, Brooks/Cole, Pacific Grove, CA, 1996.
[31] P.S.d. Laplace, Traité de Mécanique Céleste, Durprat, Paris, 1798–1825. (Translated with commentary by Nathaniel Bowditch, Boston, MA, 1829; Chelsea, Bronx, NY, 1966).
[32] C.L. Lawson, R.J. Hanson, Solving Least Squares Problems, Classics In Applied Mathematics, Vol. 15, Society for Industrial and Applied Mathematics, Philadelphia, PA, 1995.
[33] A.M. Legendre, Nouvelle méthodes pour la déterminiation des orbites des comètes, Courcier, Paris, 1805.
[34] L. Mirsky, Symmetric gauge functions and unitarily invariant norms, Quart. J. Math. Oxford, Ser. 2 11 (41) (1960) 50–59.
[35] Y. Nievergelt, Computing circles and spheres of arithmetic least squares, Comput. Phys. Comm. 81 (3) (1994) 343–350.
[36] Y. Nievergelt, The condition of Steffensen's acceleration in several variables, J. Comput. Appl. Math. 58 (1995) 291–305.
[37] Y. Nievergelt, Schmidt-Mirsky matrix approximation with linearly constrained singular values, Linear Algebra Appl. 261 (1997) 207–219.
[38] Y. Nievergelt, S.P. Keeler, Computing geodetic coordinates in space, J. Spacecraft Rockets 37 (2000) 293–296.
[39] B. Pourciau, Reading the master: Newton and the birth of celestial mechanics, Amer. Math. Mon. 104 (1) (1997) 1–19.
[40] V. Pratt, Direct least-squares fitting of algebraic surfaces, ACM Comput. Graphics 21 (4) (1987) 145–152.
[41] J.E. Prussing, B.A. Conway, Orbital Mechanics, Oxford University Press, New York, NY, 1993.
[42] W.C. Rheinboldt, Methods for Solving Systems of Nonlinear Equations, CBMS-NSF Regional Conference Series in Applied Mathematics, Vol. 14, Society for Industrial and Applied Mathematics, Philadelphia, PA, 1974.
[43] E. Schmidt, Zur Theorie der linearen und nichtlinearen Integralgleichungen. 1. Teil: Entwicklung willkürlicher Funktionen nach Systemen vorgeschriebener, Math. Ann. 63 (1907) 433–476.
[44] J.P. Snyder, Map Projections — A Working Manual (U.S. Geological Survey Professional Paper 1395). United States Government Printing Office, Washington, D.C. 20402, 1987.
[45] G.W. Stewart, On the early history of the singular value decomposition, SIAM Rev. 35 (4) (1993) 551–566.

[46] J. Stoer, R. Bulirsch, Introduction to Numerical Analysis, 2nd Edition, Springer, New York, NY, 1993.

[47] G. Strang, Introduction to Applied Mathematics, Wellesley-Cambridge Press, Wellesley, MA, 1986.

[48] A.E. Taylor, Introduction to Functional Analysis, Wiley, New York, NY, 1958.

[49] B.L. van der Waerden, Geometry and Algebra in Ancient Civilizations, Springer, New York, NY, 1983.

[50] S. Van Huffel, J. Vandewalle, The Total Least Squares Problem: Computational Aspects and Analysis, Society for Industrial and Applied Mathematics, Philadelphia, PA, 1991.

[51] C.F. Van Loan, Introduction to Scientific Computing, Prentice-Hall, Upper Saddle River, NJ, 1997.

[52] H. Weyl, Das asymptotische Vertielungsgesetz der Eigenwert linearer partieller Differentialgleichungen (mit einer Anwendung auf der Theorie der Hohlraumstrahlung), Math. Ann. 71 (1912) 441–479.

[53] J.H. Wilkinson, The Algebraic Eigenvalue Problem, Oxford Science Publication, Monographs on Numerical Analysis, Oxford University Press, Oxford, UK, paperback edition, 1988.