Available online at www.sciencedirect.com
SCIENCE DIRECT®

ELSEVIER
Linear Algebra and its Applications 378 (2004) 1–30

LINEAR ALGEBRA
AND ITS
APPLICATIONS

www.elsevier.com/locate/laa

Fitting conics of specific types to data ☆

Yves Nievergelt

Department of Mathematics, Eastern Washington University, 216 Kingston Hall, Cheney, Washington 99004-2418, USA

Received 30 September 2002; accepted 19 August 2003

Submitted by H. Schneider

Abstract

For each finite set of points in the Euclidean plane, and for each type of conic section—elliptic, hyperbolic, or parabolic—the algorithm presented here determines all the algebraically best fitting conics of the selected type: the best ellipse, the best hyperbola, or the best parabola. The supporting theory expands on Golub, Hoffman, and Stewart's generalization of the Schmidt–Mirsky matrix approximation theorem, on Bookstein's and Pratt's methods to fit conics, and on Gander, Golub, and Strebel's method to fit ellipses. Because neither the set of ellipses nor the set of hyperbolae is closed, the algorithm and its supporting theory must accommodate their boundary, which consists of the parabolic conics. The corresponding optimization problem consists in minimizing a quadratic form with two quadratic constraints, which an orthogonal change of variables transforms into a least-squares problem with one quadratic constraint. Hence analogies with geodetic coordinates identify geometric causes of numerical instability. For each type of conic, the resulting best fitting conic remains invariant under Euclidean transformations. Applications include the theory and use of sundials in archaeology and astronomy.

AMS classification: Primary 65D10; 51M16; Secondary 14H50; 14Q05; 15A18; 15A45; 41A29

Keywords: Curve fitting; Conics; Algebraic; Total least-squares

0. Introduction

The problem of fitting to data a conic of a type specified in advance arises in several applications. For instance, the determination of the intended orientation of a

☆ This work was supported in part by a Faculty Research Grant from Eastern Washington University.
E-mail address: ynievergelt@ewu.edu (Y. Nievergelt).

doi:10.1016/j.laa.2003.08.022

sundial in archeology, and of the direction of due north in astronomy, can proceed by fitting to shadow-plots a conic, the type of which depends on the latitude of the intended location of the sundial [31]. However, there exist data for which there does not exist any best fitting conic of the specified type. For example, horizontal shadow-plots at latitudes between the artic circles must lie on hyperbolae, but through measurement or rounding errors the data can lie closer to an ellipse. Then there can exist an infinite sequence of hyperbolae, which fit the data increasingly well, but fail to converge to any limiting hyperbola. Bookstein had already identified this issue:

> The fitting of a parabola is a limiting case, exactly transitional between ellipse and hyperbola. As the center of an ellipse moves off toward infinity [. . .]—[4, p. 58],

but without offering an algorithm to fit parabolae in general positions. Nevertheless, the algorithm presented here identifies this situation, and produces the best fitting parabola, at the boundary between the sets of hyperbolae and ellipses. To this end, the algorithm minimizes a quadratic form subject to two simultaneous quadratic equality constraints. There already exist several algorithms to fit general conics to data by minimizing quadratic forms subject to one quadratic constraint—and optionally to additional linear constraints—to obtain circles [20], conic splines, or conics with axes parallel to the coordinate axes [4,21], and to fit circles or ellipses [5;6, Section 21.10;12;30]. Yet any algebraic method to fit parabolae in general positions hitherto appears to be lacking.

The proofs and the algorithm presented here rely on the following notation for the singular-value decomposition (SVD) of matrices [14, p. 71]. Let $\mathbb{M}_{m\times n}(\mathbb{R})$ denote the set of all matrices with m rows, n columns, and entries in the real numbers $\mathbb{R}$. With $\ell := \min\{m, n\}$, define $D := \text{diagonal}(d_1, \dots, d_\ell) \in \mathbb{M}_{m\times n}(\mathbb{R})$ to be the matrix such that $D_{i,i} = d_i$ for every $i \in \{1, \dots, \ell\}$, and $D_{i,j} = 0$ otherwise. For example, let $I_{\ell\times\ell} := \text{diagonal}(1, \dots, 1) \in \mathbb{M}_{\ell\times\ell}(\mathbb{R})$ denote the identity matrix. For every matrix $G \in \mathbb{M}_{m\times n}(\mathbb{R})$ with rank r, the singular-value decomposition of $G = U\Sigma V^{\mathrm{T}}$ consists of orthogonal matrices $V \in \mathbb{M}_{n\times n}(\mathbb{R})$ and $U \in \mathbb{M}_{m\times m}(\mathbb{R})$, and of a positive semi-definite diagonal matrix $\Sigma = \text{diagonal}(\sigma_1 \geqslant \cdots \geqslant \sigma_r > 0 = \sigma_{r+1} = \cdots = \sigma_\ell) \in \mathbb{M}_{m\times n}(\mathbb{R})$. Also, the pseudo-inverse [14, p. 243] of the diagonal matrix $\Sigma = \text{diagonal}(\sigma_1, \dots, \sigma_r, 0, \dots, 0)$ is $\Sigma^\dagger := \text{diagonal}(1/\sigma_1, \dots, 1/\sigma_r, 0, \dots, 0) \in \mathbb{M}_{n\times m}(\mathbb{R})$, and the pseudo-inverse of G is defined as $G^\dagger := V\Sigma^\dagger U^{\mathrm{T}}$. Finally, the subordinate Euclidean norm of G is $\|G\|_2 := \sigma_1$, and the square of its Frobenius norm is $\|G\|_{\mathrm{F}}^2 := \sum_{i=1}^m \sum_{j=1}^n G_{i,j}^2 = \sum_{i=1}^r \sigma_i^2$.

The notation just established allows for detailed comparisons and contrasts between the theory and algorithms presented here and closely related current methods [4,12,20,21]. Specifically, each conic $\mathscr{C}$ in the Cartesian plane $\mathbb{R}^2$ can be specified by a symmetric matrix $S \in \mathbb{M}_{3\times 3}(\mathbb{R})$, so that each point $\vec{\mathbf{x}} \in \mathbb{R}^2$ lies on $\mathscr{C}$ if and only if it satisfies the quadratic equation

$$\begin{pmatrix} \vec{\mathbf{x}}^{\mathrm{T}} & 1 \end{pmatrix} S \begin{pmatrix} \vec{\mathbf{x}} \\ 1 \end{pmatrix} = 0. \tag{0.1}$$

The matrix S also admits of a partition in the form

$$S = \begin{pmatrix} A & \vec{\mathbf{b}} \\ \vec{\mathbf{b}}^{\mathrm{T}} & c \end{pmatrix}, \tag{0.2}$$

with a symmetric matrix $A \in \mathbb{M}_{2\times 2}(\mathbb{R})$, a vector $\vec{\mathbf{b}} \in \mathbb{R}^2$, and a scalar $c \in \mathbb{R}$. Equivalently, the entries of S can be arranged in lexicographic (or any other) order in vectors $\vec{\mathbf{w}}$ and $\vec{\mathbf{s}}$ defined by

$$\vec{\mathbf{w}} := \left(a_{1,1}, \sqrt{2}a_{1,2}, a_{2,2}\right)^{\mathrm{T}}, \tag{0.3}$$

$$\vec{\mathbf{s}} := (c, 2\vec{\mathbf{b}}, \vec{\mathbf{w}})^{\mathrm{T}}. \tag{0.4}$$

Following Bookstein [4], and Gander et al. [12], the algebraic objective function F adopted here equals the sum of the squares of the values of the left-hand side of Eq. (0.1) at each data point $\vec{\mathbf{x}}_i = (x_i, y_i)$:

$$F(S) := \sum_{i=1}^{N} \left[\begin{pmatrix} \vec{\mathbf{x}}_i^{\mathrm{T}} & 1 \end{pmatrix} S \begin{pmatrix} \vec{\mathbf{x}}_i \\ 1 \end{pmatrix} \right]^2, \tag{0.5}$$

which measures the extent to which all the data fail to lie on a common conic $\mathscr{C}$. Moreover, with the monomials in the entries of the data arranged in the order of Eq. (0.3), there is a matrix M with monomials from the data,

$$M := \begin{pmatrix} 1; & x_1 & y_1; & x_1^2, & \sqrt{2}x_1y_1, & y_1^2 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \\ 1; & x_N & y_N; & x_N^2, & \sqrt{2}x_Ny_N, & y_N^2 \end{pmatrix}, \tag{0.6}$$

so that the function F becomes the square of a Euclidean norm $\| \quad \|_2^2$:

$$F(\vec{\mathbf{s}}) = \|M\vec{\mathbf{s}}\|_2^2. \tag{0.7}$$

Because scaling equation (0.1) by a non-zero factor does not affect its solution S, the minimization of the objective function F requires a scaling constraint, which must remain invariant under Euclidean transformations for applications that are invariant. For instance, Bookstein [4] imposes the invariant constraint $\|A\|_{\mathrm{F}} = \|\vec{\mathbf{w}}\|_2 = 1$.

By Golub et al.'s theorem [13], minimizing F subject to $\|\vec{\mathbf{w}}\|_2 = 1$ amounts to determining the singular matrix $\widetilde{M}$ closest to M with the same first three columns. To this end, partition $M = [M_{\mathrm{I}}; M_{\mathrm{II}}]$ with

$$M_{\mathrm{I}} := \begin{pmatrix} 1 & x_1, & y_1 \\ \vdots & \vdots & \vdots \\ 1 & x_N, & y_N \end{pmatrix}, \quad M_{\mathrm{II}} := \begin{pmatrix} x_1^2, & \sqrt{2}x_1y_1, & y_1^2 \\ \vdots & \vdots & \vdots \\ x_N^2, & \sqrt{2}x_Ny_N, & y_N^2 \end{pmatrix}. \tag{0.8}$$

Then factor $M = QR$ with $Q \in \mathbb{M}_{N\times N}(\mathbb{R})$ orthogonal and $R \in \mathbb{M}_{N\times 6}(\mathbb{R})$ upper triangular [14, Section 5.2]. The matrix R admits of a corresponding partition

$$R = \begin{pmatrix} R_{1,\mathrm{I}} & R_{1,\mathrm{II}} \\ 0 & R_{2,\mathrm{II}} \end{pmatrix}, \tag{0.9}$$

with $R_{1,\mathrm{I}} \in \mathbb{M}_{3\times 3}(\mathbb{R})$ and $R_{2,\mathrm{II}} \in \mathbb{M}_{(N-3)\times 3}(\mathbb{R})$ both upper triangular.

Yet this factorization requires the determination of a basis for the range of M_{I} [13, p. 320], which can be numerically unstable [14, pp. 245–248]. In principle, the solution $\vec{\mathbf{w}}$ is a right-singular vector for the smallest singular value of $R_{2,\mathrm{II}}$ [13, p. 322, Eq. (3.6)], so that $\vec{\mathbf{w}}$ solves the problem of

$$\begin{aligned} &\text{minimizing} \quad \|R_{2,\mathrm{II}}\vec{\mathbf{w}}\|_2 && (0.10)\\ &\text{subject to} \quad \|\vec{\mathbf{w}}\|_2 = 1. && (0.11)\end{aligned}$$

Then $\vec{\mathbf{v}} := (c, 2\vec{\mathbf{b}}^{\mathrm{T}})^{\mathrm{T}}$ solves $R_{1,\mathrm{I}}\vec{\mathbf{v}} = -R_{1,\mathrm{II}}\vec{\mathbf{w}}$. However, $R_{1,\mathrm{I}}$ can be ill-conditioned or singular if all the data lie near or on a straight line. Also, a theorem of Stewart's [26, p. 515] shows that the sensitivity of the factorization $M = QR$ diverges to infinity as the data near a perfect fit. Indeed, for all matrices $G, G' \in \mathbb{M}_{m\times n}(\mathbb{R})$, let $\Delta G := G' - G$. Then with $\| \quad \|$ denoting the Frobenius or spectral norm, $G^{\dagger}$ the pseudo-inverse of G, and $\kappa(G) := \|G\| \cdot \|G^{\dagger}\|$ the corresponding condition number, Stewart's theorem states that for all matrices $G, G' \in \mathbb{M}_{m\times n}(\mathbb{R})$, if $G = QR$ and $G' = Q'R'$, then

$$\frac{\|\Delta R\|}{\|G\|} \leqslant 2\big(2+\sqrt{2}\big)n\kappa(G)\frac{\|\Delta G\|}{\|G\|}, \tag{0.12}$$

$$\frac{\|\Delta Q\|}{\|Q\|_2} \leqslant \frac{3\kappa(G)\frac{\|\Delta G\|}{\|G\|}}{1 - 2\kappa(G)\frac{\|\Delta G\|}{\|G\|}}. \tag{0.13}$$

With $G := M$, at a perfect fit $M(\vec{\mathbf{v}}^{\mathrm{T}}, \vec{\mathbf{w}}^{\mathrm{T}})^{\mathrm{T}} = M\vec{\mathbf{s}} = \vec{\mathbf{0}}$, and then $\kappa(M) = \infty$. This paradox disappears with the realization that the sensitivity of the fitted conic depends on the factorization $M_{\mathrm{I}} = Q_{\mathrm{I}}R_{1,\mathrm{I}}$ rather than on that of M, as explained in the last paragraph of Section 6.

A different issue arises if the algorithm produces a conic $\mathscr{C}$ of a type different from the one sought. The conic $\mathscr{C}$ corresponds to the equivalence class $[\vec{\mathbf{s}}]$ of $\vec{\mathbf{s}}$ in the projective space $\mathbb{P}(\mathbb{R}^6) = \mathbb{P}^5$ [10, p. 108]. This bijection $\mathscr{C} \mapsto [\vec{\mathbf{s}}]$ from the set of all conics to the projective space $\mathbb{P}^5$ also pulls back to the set of conics any topology from $\mathbb{P}^5$, for instance, the Hausdorff topology induced from $\mathbb{R}^6$, or the Zariski topology, where a set is closed if and only if it is a projective algebraic set [10, p. 132]. In either topology, the set of all parabolic conics is closed, because it is defined by the homogeneous equation $\det(A) = 0$. Hence its complement—where $\det(A) \neq 0$—is open but not closed. In the Hausdorff topology, the set of elliptic conics, where $\det(A) > 0$, and the set of hyperbolic conics, where $\det(A) < 0$, are each open but not closed either. Therefore, a continuous objective function can fail to have a global minimum on either set. However, their closures—where $\det(A) \geqslant 0$ or where $\det(A) \leqslant 0$—are compact in the Hausdorff topology, so that any continous objective function has a global mimimum.

Thus the problem of fitting to data a conic of a specified type can involve minimizing such an objective as in equation (0.10) subject to *two* simultaneous constraints: one on $\|A\|_{\mathrm{F}}$ and the other on $\det(A)$.

To simplify the theory and the computations, Section 1 performs an orthogonal change of variables on the vector of the entries of the symmetric matrix A, so that the quadratic forms $\|A\|_{\mathrm{F}}^2$ and $\det(A)$ are both diagonal. Section 2 then shows how to identify the optimal solution from multiple stationary points, and Section 4 describes the final algorithm in detail. Section 3 charaterizes the solution in terms of a dual minimal perturbation ofthe data. For further perturbation analyses, Section 5 collects results on the sensitivity of singular vectors of projections of matrices. Analyses of the sensitivity to perturbations of the data then follow in Section 6 for general conics, and in Section 7 for parabolae. Finally, Section 8 shows applications and test cases.

1. Transformations of constraints

For each symmetric matrix $A \in \mathbb{M}_{2\times 2}(\mathbb{R})$, the square of the Frobenius norm $\|A\|_{\mathrm{F}}^2$ and the determinant $\det(A)$ are quadratic forms in the entries of A. For the matrices of both quadratic forms to be diagonal, Theorem 1.1 introduces a diagonal matrix D and an orthogonal transformation $\vec{\mathbf{r}} := Z\vec{\mathbf{w}}$ defined by

$$D := \frac{1}{4}\begin{pmatrix} -1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \quad Z := \frac{1}{\sqrt{2}}\begin{pmatrix} -1 & 0 & 1 \\ 0 & \sqrt{2} & 0 \\ 1 & 0 & 1 \end{pmatrix}. \tag{1.1}$$

Theorem 1.1. *For each symmetric matrix $A \in \mathbb{M}_{2\times 2}(\mathbb{R})$ if $\vec{\mathbf{r}} = Z\vec{\mathbf{w}}$, then*

$$\|A\|_{\mathrm{F}}^2 = \vec{\mathbf{r}}^{\mathrm{T}} I \vec{\mathbf{r}}, \tag{1.2}$$

$$\det\begin{pmatrix} a_{1,1} & a_{1,2} \\ a_{1,2} & a_{2,2} \end{pmatrix} = \vec{\mathbf{r}}^{\mathrm{T}} D \vec{\mathbf{r}}. \tag{1.3}$$

Proof. From (0.3) and (1.1) the proof is a straightforward calculation. □

For future reference, $r_3 = (w_1 + w_3)/\sqrt{2} = (a_{1,1} + a_{2,2})/\sqrt{2} = \mathrm{trace}(A)/\sqrt{2}$. Thus Z transforms the simultaneous quadratic constraints

$$\|A\|_{\mathrm{F}}^2 = 1, \tag{1.4}$$

$$\det(A) = 0, \tag{1.5}$$

into the simultaneously diagonal quadratic constraints

$$\vec{\mathbf{r}}^{\mathrm{T}}\vec{\mathbf{r}} = 1, \tag{1.6}$$

$$\vec{\mathbf{r}}^{\mathrm{T}} D \vec{\mathbf{r}} = 0. \tag{1.7}$$

Adding and subtracting Eqs. (1.6) and (1.7) gives

$$\vec{\mathbf{r}}^{\mathrm{T}}\vec{\mathbf{r}} = r_1^2 + r_2^2 + r_3^2 = 1, \tag{1.8}$$

$$\vec{\mathbf{r}}^{\mathrm{T}} D \vec{\mathbf{r}} = -(r_1^2 + r_2^2) + r_3^2 = 0; \tag{1.9}$$

$$\vec{\mathbf{r}}^{\mathrm{T}} \vec{\mathbf{r}} + \vec{\mathbf{r}}^{\mathrm{T}} D \vec{\mathbf{r}} = 2r_3^2 = 1, \tag{1.10}$$

$$\vec{\mathbf{r}}^{\mathrm{T}} \vec{\mathbf{r}} - \vec{\mathbf{r}}^{\mathrm{T}} D \vec{\mathbf{r}} = 2(r_1^2 + r_2^2) = 1. \tag{1.11}$$

The sum (1.10) gives the constant constraint trace$(A) = \sqrt{2} r_3 = \pm 1$. The difference (1.11) leads to the transformation

$$\vec{\mathbf{q}} = \begin{pmatrix} q_1 \\ q_2 \end{pmatrix} := \sqrt{2} \begin{pmatrix} r_1 \\ r_2 \end{pmatrix}, \tag{1.12}$$

so that the constraints (1.4) and (1.5) or (1.6) and (1.7) reduce to one constraint:

$$\|\vec{\mathbf{q}}\|_2 = 1. \tag{1.13}$$

The problem of fitting to data a parabola will amount to minimizing a quadratic form of the type $\|T\vec{\mathbf{w}}\|_2^2$ subject to such constraints, as in Eqs. (0.10) and (0.11), with a matrix $T \in \mathbb{M}_{k\times 3}(\mathbb{R})$ equal to a variant of $R_{2,\mathrm{II}} \in \mathbb{M}_{(m-3)\times 3}(\mathbb{R})$ in Eq. (0.9). For each matrix $T \in \mathbb{M}_{k\times 3}(\mathbb{R})$, the problem of

- minimizing $\|T\vec{\mathbf{w}}\|_2$
- subject to the two constraints (1.4) and (1.5): $\|\vec{\mathbf{w}}\|_2 = 1$ and $\det(A) = 0$

reduces, after the change of variables defined by Z, to the problem of

- minimizing $\|(TZ)\vec{\mathbf{r}}\|_2$
- subject to the two constraints (1.6) and (1.7): $\vec{\mathbf{r}}^{\mathrm{T}}\vec{\mathbf{r}} = 1$ and $\vec{\mathbf{r}}^{\mathrm{T}} D\vec{\mathbf{r}} = 0$.

This problem remains invariant under multiplication by -1, so that the choice $\sqrt{2} r_3 = +1$ with $\sqrt{2}\vec{\mathbf{r}}^{\mathrm{T}} = (\vec{\mathbf{q}}^{\mathrm{T}}, 0) + \vec{\mathbf{e}}_3^{\mathrm{T}}$ lead to the problem of

- minimizing $\|(TZ)_{1-2}\vec{\mathbf{q}} + (TZ)\vec{\mathbf{e}}_3\|_2$
- subject to the single constraint (1.13): $\|\vec{\mathbf{q}}\|_2 = 1$,

where $(TZ)_{1-2} \in \mathbb{M}_{k\times 2}(\mathbb{R})$ consists of columns 1 and 2 of TZ. With $G := (TZ)_{1-2}$ and $\vec{\mathbf{p}} := -G\vec{\mathbf{e}}_3$, Section 2 minimizes $\|G\vec{\mathbf{q}} - \vec{\mathbf{p}}\|_2$ subject to $\|\vec{\mathbf{q}}\|_2 = 1$. The solution $\vec{\mathbf{q}}$ then corresponds to the symmetric matrix A with entries

$$\begin{pmatrix} a_{1,1} \\ \sqrt{2} a_{1,2} \\ a_{2,2} \end{pmatrix} = \vec{\mathbf{w}} = Z\vec{\mathbf{r}} = Z\frac{1}{\sqrt{2}} \begin{pmatrix} q_1 \\ q_2 \\ 1 \end{pmatrix}. \tag{1.14}$$

2. The secular equation

The minimization of $\|G\vec{\mathbf{q}} - \vec{\mathbf{p}}\|_2$ under the quadratic constraint $\|\vec{\mathbf{q}}\|_2 = 1$ can proceed through the method of Lagrange multipliers. The equation that determines the

Lagrange multiplier—called the “secular equation” [14, p. 564]—can have several solutions, corresponding to several stationary points of the objective functions. The problem of determining which solution of the secular equation (which Lagrange multiplier) corresponds to a global minimum of the objective over the feasible set has been studied by Gander [11, Section 8]. To this end, a comparison with an equivalent geometric problem also proves useful. Specifically, let $\mathscr{S}^{n-1}$ denote the unit sphere in $\mathbb{R}^n$. Then its image $G(\mathscr{S}^{n-1})$ by a matrix $G \in \mathbb{M}_{m\times n}(\mathbb{R})$ is an ellipsoid in $\mathbb{R}^m$. Hence the following two problems are mutually equivalent:

Problem 2.1. Find a *unit* vector $\vec{\mathbf{q}} \in \mathscr{S}^{n-1}$ minimizing $\|G\vec{\mathbf{q}} - \vec{\mathbf{p}}\|_2$.

Problem 2.2. On the ellipsoid $G(\mathscr{S}^{n-1})$, find a point $G\vec{\mathbf{q}}$ closest to $\vec{\mathbf{p}}$.

For $n = 2$, Problem 2.2 shows that Problem 2.1 amounts to computing the geodetic distance from a point $\vec{\mathbf{p}} \in \mathbb{R}^m$ to an ellipse $G(\mathscr{S}^1) \subset \mathbb{R}^m$. (This ellipse is the set of feasible parameters; in particular, this ellipse has no relation to the fitted conic $\mathscr{C}$, other than containing a point $G\vec{\mathbf{q}}$ corresponding to the yet unknown parameters $\vec{\mathbf{q}}$ of $\mathscr{C}$.) If the point $\vec{\mathbf{p}}$ lies outside the ellipse, then by convexity there exists exactly one closest point; moreover, for ellipses with small eccentricities, the distance can be computed accurately by iterating a contracting map [17]. However, if the point $\vec{\mathbf{p}}$ lies inside the ellipse, then the distance from $\vec{\mathbf{p}}$ can have local extrema besides the global minimum on the ellipse [2, p. 238, Theorem 17.5.5.6], as shown in Fig. 1.

This geometric formulation also shows that such a problem can have either exactly one solution, or multiple solutions. Such multiple solutions occur if and only if $\vec{\mathbf{p}}$ is perpendicular to a shortest principal axis, and closer to the center of the ellipse than the center of curvature at the opposite vertex, as in Fig. 2. Otherwise, then by symmetry and convexity there exists only one closest point on the ellipse, in the same quadrant where $\vec{\mathbf{p}}$ lies.

With a singular-value decomposition $G = U\Sigma V^{\mathrm{T}} = \sigma_1\vec{\mathbf{u}}_1\vec{\mathbf{v}}_1^{\mathrm{T}} + \sigma_2\vec{\mathbf{u}}_2\vec{\mathbf{v}}_2^{\mathrm{T}}$, the vector $\vec{\mathbf{y}} := U^{\mathrm{T}}\vec{\mathbf{p}}$ is the orthogonal projection of $\vec{\mathbf{p}}$ on the range of G. Moreover, the ellipse $G(\mathscr{S}^1)$ has its major and minor principal axes with lengths σ_1 and σ_2 along the left-singular vectors $\vec{\mathbf{u}}_1$ and $\vec{\mathbf{u}}_2$ of G. Thus the condition that $\vec{\mathbf{p}}$ not be perpendicular to the shorter axis becomes $|\sigma_2 y_2| > 0 = \sigma_2^2 - \sigma_2^2$. Furthermore, the radius of curvature at the vertex on the major principal axis equals σ_2^2/σ_1, so that the

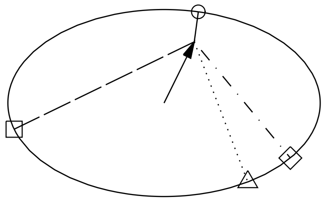

Fig. 1. Global minimum (○) and maximum (□), local minimum (△) and local maximum (◇), for the distance from a point $\vec{\mathbf{p}}$ (tip of ↑) to an ellipse.

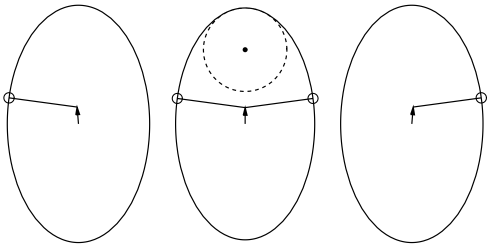

Fig. 2. Multiple global minima or large perturbations of the global minimum (○) can occur for small perturbations of a point $\vec{\mathbf{p}}$ (tip of ↑) farther away from a vertex than the center of curvature (•) at that vertex.

center of curvature lies at $\sigma_1 - \sigma_2^2/\sigma_1 = (\sigma_1^2 - \sigma_2^2)/\sigma_1$ [2, p. 246, Section 17.7.4]. Thus the condition that $\vec{\mathbf{y}}$ lies farther away from the center of the ellipse than the center of curvature does becomes $|\sigma_1 y_1| > \sigma_1^2 - \sigma_2^2$. Corroborating this geometric analysis, Theorem 2.3 confirms that Problem 2.1 has a unique solution—counted with multiplicities—if and only if $|\sigma_\ell y_\ell| > \sigma_\ell^2 - \sigma_2^2$ for at least one $\ell \in \{1, 2\}$. With the notation adopted here, Gander has shown that this solution corresponds to the smallest Lagrange multiplier [11, Section 4]; the proof of Theorem 2.3 refines this estimate by providing a finite interval, computable from the data, which contains this Lagrange multiplier.

Theorem 2.3. *For each $G \in \mathbb{M}_{m\times 2}(\mathbb{R})$ with a singular-value decomposition $G = U\Sigma V^{\mathrm{T}}$, and for each $\vec{\mathbf{p}} \in \mathbb{R}^m$ with $\vec{\mathbf{y}} := U^{\mathrm{T}}\vec{\mathbf{p}}$, Problem 2.1 has exactly one—simple—solution if and only if $|\sigma_\ell y_\ell| > \sigma_\ell^2 - \sigma_2^2$ for some $\ell \in \{1, 2\}$.*

Proof. Problem 2.1 is equivalent to finding $\vec{\mathbf{z}} := V^{\mathrm{T}}\vec{\mathbf{q}} \in \mathbb{R}^2$ minimizing

$$f(\vec{\mathbf{z}}) := \|\Sigma\vec{\mathbf{z}} - \vec{\mathbf{y}}\|_2^2 = \|U^{\mathrm{T}}\{G(V\vec{\mathbf{z}}) - \vec{\mathbf{p}}\}\|_2^2 = \|G\vec{\mathbf{q}} - \vec{\mathbf{p}}\|_2^2, \tag{2.1}$$

subject to the constraint $1 = \|\vec{\mathbf{q}}\|_2^2 = \|V^{\mathrm{T}}\vec{\mathbf{q}}\|_2^2 = \|\vec{\mathbf{z}}\|_2^2$, or, equivalently,

$$g(\vec{\mathbf{z}}) := \vec{\mathbf{z}}^{\mathrm{T}}\vec{\mathbf{z}} - 1 = 0. \tag{2.2}$$

Lagrange's equation gradient$(f) = \lambda$gradient(g) takes the form [14, p. 563]

$$(\Sigma^{\mathrm{T}}\Sigma - \lambda I)\vec{\mathbf{z}} = \Sigma^{\mathrm{T}}\vec{\mathbf{y}}. \tag{2.3}$$

At a minimum $\vec{\mathbf{z}}$, the matrix $\Sigma^{\mathrm{T}}\Sigma - \lambda I$ must be positive semi-definite on the space $\vec{\mathbf{z}}^{\perp}$ tangent to the unit circle at $\vec{\mathbf{z}}$, whence $\lambda \leqslant \sigma_2^2$ [9, p. 166, #11]. If $\lambda = \sigma_2^2$, then $\sigma_2 y_2$ must vanish, because system (2.3) becomes

$$(\sigma_1^2 - \sigma_2^2)z_1 = \sigma_1 y_1, \tag{2.4}$$

$$0z_2 = \sigma_2 y_2. \tag{2.5}$$

To keep track of which product(s) $\sigma_i y_i = 0$, define

$$k := \begin{cases} 2 & \text{if } \sigma_2 y_2 \neq 0, \\ 1 & \text{if } \sigma_2 y_2 = 0 \neq \sigma_1 y_1, \\ 0 & \text{if } \sigma_2 y_2 = 0 = \sigma_1 y_1. \end{cases} \tag{2.6}$$

If $\sigma_1 y_1 \neq 0$ or $\sigma_2 y_2 \neq 0$, and $\lambda \notin \{\sigma_1^2, \sigma_2^2\}$, then (2.3) has one solution $\vec{\mathbf{z}}(\lambda)$:

$$z_i(\lambda) := \frac{\sigma_i y_i}{\sigma_i^2 - \lambda} \tag{2.7}$$

for each $i \in \{1, 2\}$. The constraint $\|\vec{\mathbf{z}}(\lambda)\|_2^2 = 1$ is the "secular" equation

$$\varphi(\lambda) := \sum_{i=1}^{k} \frac{\sigma_i^2 y_i^2}{(\sigma_i^2 - \lambda)^2} = \|\vec{\mathbf{z}}(\lambda)\|_2^2 = 1. \tag{2.8}$$

Eq. (2.8) has exactly one (simple) solution λ_* such that $\lambda_* < \sigma_k^2$, because φ increases on the open interval $]-\infty, \sigma_k^2[$, with $\lim_{\lambda \searrow -\infty} \varphi(\lambda) = 0$ and $\lim_{\lambda \nearrow \sigma_k^2} \varphi(\lambda) = +\infty$, thanks to $\sigma_k^2 y_k^2 > 0$. To narrow down this interval, let

$$\lambda_\flat := \sigma_k^2 - \sqrt{k} \max_{1 \leqslant i \leqslant k} |\sigma_i y_i|, \tag{2.9}$$

$$\lambda_\sharp := \sigma_k^2 - |\sigma_k y_k|. \tag{2.10}$$

If $\lambda < \lambda_\flat$, then $\sigma_i^2 - \lambda > \sigma_i^2 - \lambda_\flat \geqslant \sqrt{k} \max_{1 \leqslant i \leqslant k} |\sigma_i y_i| \geqslant \sqrt{k}|\sigma_i y_i|$, whence

$$\varphi(\lambda) = \sum_{i=1}^{k} \frac{\sigma_i^2 y_i^2}{(\sigma_i^2 - \lambda)^2} < \sum_{i=1}^{k} \frac{\sigma_i^2 y_i^2}{(\sqrt{k}|\sigma_i y_i|)^2} = 1. \tag{2.11}$$

Similarly, if $\lambda_\sharp < \lambda$, then $(\sigma_k^2 - \lambda)^2 < \sigma_k^2 y_k^2$, whence

$$1 < \frac{\sigma_k^2 y_k^2}{(\sigma_k^2 - \lambda)^2} \leqslant \sum_{i=1}^{k} \frac{\sigma_i^2 y_i^2}{(\sigma_i^2 - \lambda)^2} = \varphi(\lambda). \tag{2.12}$$

Thus, the unique solution $\lambda_* < \sigma_k^2$ lies in the finite closed interval $[\lambda_\flat, \lambda_\sharp]$.

Direct implication. Assume that $|\sigma_\ell y_\ell| > \sigma_\ell^2 - \sigma_2^2$ for some $\ell \in \{1, 2\}$; in particular, $|\sigma_\ell y_\ell| > \sigma_\ell^2 - \sigma_2^2 \geqslant \sigma_2^2 - \sigma_2^2 \geqslant 0$. Firstly, $\lambda := \sigma_2^2$ is *not* a Lagrange multiplier minimizing (2.1) subject to (2.2): if $\lambda = \sigma_2^2$, then (2.5) forces $|\sigma_2 y_2| = 0 = \sigma_2^2 - \sigma_2^2$, whence $\ell = 1$ and $|\sigma_1 y_1| > \sigma_1^2 - \sigma_2^2 \geqslant 0$ by hypothesis. Hence (2.4) has a solution if and only if $\sigma_1^2 - \sigma_2^2 \neq 0$, but then $|z_1| = |\sigma_1 y_1|/(\sigma_1^2 - \sigma_2^2) > 1$, so that (2.3) has *no* solutions with $\|\vec{\mathbf{z}}\|_2 = 1$.

Secondly, $\lambda_\sharp < \sigma_2^2$: if $k = 2$, then $\sigma_2 y_2 \neq 0$, whence $\lambda_\sharp = \sigma_2^2 - |\sigma_2 y_2| < \sigma_2^2$; if $k = 1$, then $\sigma_2 y_2 = 0$, but the hypothesis guarantees that $|\sigma_1 y_1| > \sigma_1^2 - \sigma_2^2$, so that $\lambda_\sharp = \sigma_1^1 - |\sigma_1 y_1| < \sigma_k^1 - (\sigma_1^2 - \sigma_2^2) = \sigma_2^2$. Consequently, $\lambda_* \leqslant \lambda_\sharp < \sigma_2^2$. Therefore, λ_* corresponds to the unique global constrained minimum of f.

Converse implication. Assume that $|\sigma_i y_i| \leqslant \sigma_i^2 - \sigma_2^2$ for every $i \in \{1, 2\}$. Hence $0 \leqslant |\sigma_2 y_2| \leqslant \sigma_2^2 - \sigma_2^2 = 0$, and the secular equation reduces to

$$\varphi(\lambda) = \sum_{i=1}^{k} \frac{\sigma_i^2 y_i^2}{(\sigma_i^2 - \lambda)^2} = \frac{\sigma_1^2 y_1^2}{(\sigma_1^2 - \lambda)^2} = 1. \tag{2.13}$$

Case 1. If $\sigma_1 y_1 \neq 0$, then (2.13) has two solutions: $\lambda_2 := \sigma_1^2 + |\sigma_1 y_1| > \sigma_1^2$ and $\lambda_1 := \sigma_1^2 - |\sigma_1 y_1| \geqslant \sigma_1^2 - (\sigma_1^2 - \sigma_2^2) = \sigma_2^2$. Hence $\sigma_2^2 \leqslant \lambda_1 < \lambda_2$, so that $\Sigma^{\mathrm{T}}\Sigma - \lambda_j I$ fails to be positive definite for each solution λ_j. Yet derived from Eq. (2.8), the secular equation (2.13) holds only for $\lambda \notin \{\sigma_1^2, \sigma_2^2\}$. Thus the minimum occurs for some $\lambda \in \{\sigma_1^2, \sigma_2^2\}$, whence $\lambda = \sigma_2^2$.

From $\sigma_1 y_1 \neq 0$, it also follows that $0 < |\sigma_1 y_1| \leqslant \sigma_1^2 - \sigma_2^2$, whence (2.4) has exactly one solution, given by Eq. (2.7):

$$z_1(\lambda) = \frac{\sigma_1 y_1}{\sigma_1^2 - \lambda} = \frac{\sigma_1 y_1}{\sigma_1^2 - \sigma_2^2}, \tag{2.14}$$

with $|z_1(\lambda)| \leqslant 1$ because $|\sigma_1 y_1| \leqslant \sigma_1^2 - \sigma_2^2$ under the new hypotheses. From $\sigma_2 y_2 = 0$ and $\lambda = \sigma_2^2$, it also follows that Eq. (2.5) become $0 z_2 = 0$, which admits infinitely many solutions z_2. From $\|\vec{\mathbf{z}}(\lambda)\|_2 = 1$, it follows that $z_2 = \pm\sqrt{1 - |z_1(\lambda)|^2}$ gives two solutions, counted with multiplicities.

Case 2. If $\sigma_i y_i = 0$ for each $i \in \{1, 2\}$, which occurs if and only if $\vec{\mathbf{p}}$ is perpendicular to the range of G, then in its common form (2.8) [3, p. 207, Eq. (5.3.21)], [14, p. 564] the secular equation has *no* solutions. Nevertheless, system (2.3) admits non-zero solutions if and only if it becomes singular, which occurs if and only if $\lambda \in \{\sigma_1^2, \sigma_2^2\}$, in particular, $\lambda := \sigma_2^2$ at a minimum. If $\sigma_1 > \sigma_2$, then the solution $\lambda = \sigma_2^2$ corresponds to $\vec{\mathbf{z}} = \pm\vec{\mathbf{e}}_2$ and hence $\vec{\mathbf{q}} = V\vec{\mathbf{z}} = \pm\vec{\mathbf{v}}_2$, which minimizes f subject to g, because $\|G\vec{\mathbf{q}} - \vec{\mathbf{p}}\|_2^2 = \|G\vec{\mathbf{q}}\|_2^2 + \|\vec{\mathbf{p}}\|_2^2 = \sigma_2^2 + \|\vec{\mathbf{p}}\|_2^2$ with $\vec{\mathbf{p}}$ perpendicular to the range of G. If $\sigma_1 = \sigma_2$, then f is constant on $\mathscr{S}^1$, and every $\vec{\mathbf{z}} \in \mathscr{S}^1$ gives the minimum $\sigma_2^2 + \|\vec{\mathbf{p}}\|_2^2$. □

The literature proposes several numerical methods to solve Problem 2.1. For applications seeking the shortest least-squares solution minimzing $\|G\vec{\mathbf{q}} - \vec{\mathbf{p}}\|_2$, estimations of a suitable Lagrange multiplier—without exactly solving the secular equation—followed by orthogonal transformations to minimize the objective can prove computationally efficient [3, Section 5.3;11, Section 4;18, p. 190]. However, for applications calling for a least-squares solution with a specified length, such as $\|\vec{\mathbf{q}}\|_2 = 1$ here, or for greater numerical accuracy rather than efficiency, the literature recommends computing the singular-value decomposition of G [3, p. 208;14, p. 564], and then solving the secular equation through Newton's method [14, p. 564;18, p. 193].

To this end, Algorithm 2.4 solves the secular equation (2.8). In general, with $\sigma_2 y_2 \neq 0$ and $|\sigma_1 y_1| > \sigma_1^2 - \sigma_2^2$, the function φ is strictly convex, so that Newton's method converges monotonically and quadratically to λ_* from the upper bound

$\lambda_0 := \lambda_\sharp$. In all the other situations, either algebra gives $\lambda_* = \sigma_2^2 - |\sigma_2 y_2|$ or $\lambda_* = \sigma_1^2 - |\sigma_1 y_1|$, or the condition for a minimum gives $\lambda_* = \sigma_2^2$.

Algorithm 2.4 (*Solution of the secular equation*).

```
PROCEDURE secular($\sigma_1, \sigma_2, y_1, y_2$).
DATA: $0 \leqslant \sigma_2 \leqslant \sigma_1 \in \mathbb{R}$, $y_1, y_2 \in \mathbb{R}$.
RESULT: Lagrange multiplier $\lambda_* \in \mathbb{R}$ with $\lambda_* \leqslant \sigma_2^2$.
START
If $\sigma_2 y_2 = 0$ and $|\sigma_1 y_1| \leqslant \sigma_1^2 - \sigma_2^2$, then $\lambda := \sigma_2^2$;
else if $\sigma_2 y_2 = 0$ and $|\sigma_1 y_1| > \sigma_1^2 - \sigma_2^2$, then $\lambda := \sigma_1^2 - |\sigma_1 y_1|$;
else if $\sigma_2 y_2 \neq 0$ and $|\sigma_1 y_1| \leqslant \sigma_1^2 - \sigma_2^2$, then $\lambda := \sigma_2^2 - |\sigma_2 y_2|$;
else if $\sigma_2 y_2 \neq 0$ and $|\sigma_1 y_1| > \sigma_1^2 - \sigma_2^2$, then
    $\lambda_\flat := \sigma_2^2 - \sqrt{2}\max\{|\sigma_1 y_1|, |\sigma_2 y_2|\}$,
    $\lambda_\sharp := \sigma_2(\sigma_2 - |y_2|)$,
    $\varphi(\lambda) := \left(\frac{\sigma_1 y_1}{\sigma_1^2-\lambda}\right)^2 + \left(\frac{\sigma_2 y_2}{\sigma_i^2-\lambda}\right)^2$,
    solve $\varphi(\lambda_*) = 0$ on $[\lambda_\flat, \lambda_\sharp]$, e.g., with Newton's method;
end if;
return secular($\sigma_1, \sigma_2, y_1, y_2$) := $\lambda_*$.
STOP.
```

Following [14, p. 564], Algorithm 2.5 relies on Theorem 2.3 and Algorithm 2.4 to minimize $\|G\vec{\mathbf{q}} - \vec{\mathbf{p}}\|_2$ subject to $\|\vec{\mathbf{q}}\|_2 = 1$.

Algorithm 2.5 (*Minimizing* $\|G\vec{\mathbf{q}} - \vec{\mathbf{p}}\|_2$ *subject to* $\|\vec{\mathbf{q}}\|_2 = 1$).

```
PROCEDURE geodetic($m, G, \vec{\mathbf{p}}$).
DATA: $m \geqslant 2$, $G \in \mathbb{M}_{m\times 2}(\mathbb{R})$, $\vec{\mathbf{p}} \in \mathbb{R}^m$.
RESULT: $\vec{\mathbf{q}} \in \mathbb{R}^2$ minimizes $\|G\vec{\mathbf{q}} - \vec{\mathbf{p}}\|_2$ with $\|\vec{\mathbf{q}}\|_2 = 1$.
START
Compute the singular-value decomposition
    $G = U\Sigma V^{\mathrm{T}} = \sigma_1\vec{\mathbf{u}}_1\vec{\mathbf{v}}_1^{\mathrm{T}} + \sigma_2\vec{\mathbf{u}}_2\vec{\mathbf{v}}_2^{\mathrm{T}}$;
compute the orthogonal projection
    $\vec{\mathbf{y}} := (y_1, y_2) := (\vec{\mathbf{u}}_1^{\mathrm{T}}\vec{\mathbf{p}}, \vec{\mathbf{u}}_2^{\mathrm{T}}\vec{\mathbf{p}})$;
compute the Lagrange multiplier, e.g., by Algorithm 2.4
    $\lambda_* :=$ secular($\sigma_1, \sigma_2, y_1, y_2$);
if $\lambda_* = \sigma_2^2$, then
    $\vec{\mathbf{z}} := \vec{\mathbf{e}}_2 := (0, 1)^{\mathrm{T}}$;
else
    $z_1 := \sigma_1 y_1/(\sigma_1^2 - \lambda_*)$,
    $z_2 := \sigma_2 y_2/(\sigma_2^2 - \lambda_*)$,
end if;
$\vec{\mathbf{q}} := V\vec{\mathbf{z}} = z_1\vec{\mathbf{v}}_1 + z_2\vec{\mathbf{v}}_2$;
return geodetic$(m, G, \vec{\mathbf{p}}) := \vec{\mathbf{q}}$.
STOP.
```

3. Equivalent perturbations of the data

Linear least-squares problems can have two equivalent "dual" formulations.

For example, if $G \in \mathbb{M}_{m\times n}(\mathbb{R})$ has rank n and a singular-value decomposition $G = \sum_{i=1}^{n} \sigma_i \vec{\mathbf{u}}_i \vec{\mathbf{v}}_i^{\mathrm{T}}$, then the problem of finding a unit vector $\vec{\mathbf{v}}$ minimizing $\|G\vec{\mathbf{v}}\|_2$ admits the solution $\vec{\mathbf{v}} := \vec{\mathbf{v}}_n$. Dually, the problem of finding a singular matrix $S \in \mathbb{M}_{m\times n}(\mathbb{R})$ minimizing $\|G - S\|_{\mathrm{F}}$ admits the solution $S := G - \sigma_n \vec{\mathbf{u}}_n \vec{\mathbf{v}}_n^{\mathrm{T}}$ by a theorem of Schmidt's [22] and Mirsky's [19].

More generally, the problem of finding a unit vector $\vec{\mathbf{v}}$ perpendicular to the first k columns of G and minimizing $\|G\vec{\mathbf{v}}\|_2$ is dual to the problem of finding a singular matrix $S \in \mathbb{M}_{m\times n}(\mathbb{R})$ with the same first k columns of G and minimizing $\|G - S\|_{\mathrm{F}}$, by Golub et al.'s theorem [13].

In either of the examples just cited, if the entries of the matrix G arise from data, and if the vector $\vec{\mathbf{v}}$ consists of parameters fitted to the data, then the dual formulation corresponds to the smallest perturbation S of the matrix G that fits the parameters in $\vec{\mathbf{v}}$ exactly, because $S\vec{\mathbf{v}} = \vec{\mathbf{0}}$.

To find a dual to Problem 2.1, denote by P_G the orthogonal projection of $\mathbb{R}^m$ on the range $\mathscr{R}(G)$ of a matrix $G \in \mathbb{M}_{m\times n}(\mathbb{R})$, and let $P_G^{\perp} := I - P_G$.

Problem 3.1. Find matrices $J, H \in \mathbb{M}_{m\times n}(\mathbb{R})$ and a *unit* vector $\vec{\mathbf{q}} \in \mathbb{R}^n$

- minimizing the Frobenius norms $\|J\|_{\mathrm{F}}$ and $\|H\|_{\mathrm{F}}$
- subject to $(G + J)\vec{\mathbf{q}} = P_G(\vec{\mathbf{p}})$ and $(G + J + H)\vec{\mathbf{q}} = \vec{\mathbf{p}}$.

The following theorem shows that the optimal solution $\vec{\mathbf{q}}_* \in \mathscr{S}^{n-1}$ minimizing $\|G\vec{\mathbf{q}} - \vec{\mathbf{p}}\|_2$ corresponds to a minimal perturbation of the matrix G.

Theorem 3.2. *Problems* 2.1 *and* 3.1 *are equivalent*.

Proof. With $\vec{\mathbf{z}}_* := V^{\mathrm{T}}\vec{\mathbf{q}}_*$, define $\vec{\mathbf{t}}_* := (U^{\mathrm{T}}P_G)\vec{\mathbf{p}} - \Sigma\vec{\mathbf{z}}_*$ and $\Gamma := \vec{\mathbf{t}}_*\vec{\mathbf{z}}_*^{\mathrm{T}}$. Thus $\Gamma\vec{\mathbf{z}}_* = \vec{\mathbf{t}}_*$, whence $(\Sigma + \Gamma)\vec{\mathbf{z}}_* = (U^{\mathrm{T}}P_G)\vec{\mathbf{p}}$, and then $\|\Gamma\|_2 = \|\vec{\mathbf{t}}_*\|_2 = \|(U^{\mathrm{T}}P_g)\vec{\mathbf{p}} - \Sigma\vec{\mathbf{z}}_*\|_2$. For every matrix Υ such that $(\Sigma + \Upsilon)\vec{\mathbf{z}}_* = (U^{\mathrm{T}}P_G)\vec{\mathbf{p}}$,

$$\|\Upsilon\|_{\mathrm{F}} \geqslant \|\Upsilon\|_{2,2} \geqslant \|\Upsilon\|_{2,2} \cdot \|\vec{\mathbf{z}}_*\|_2 \geqslant \|\Upsilon\vec{\mathbf{z}}_*\|_2 = \|\vec{\mathbf{t}}_*\|_2. \tag{3.1}$$

Thus Γ and $J := U\Gamma V^{\mathrm{T}}$ are the matrices with the smallest Euclidean and Frobenius norms with $(\Sigma + \Gamma)\vec{\mathbf{z}}_* = (U^{\mathrm{T}}P_G)\vec{\mathbf{p}}$ and $(G + J)\vec{\mathbf{q}}_* = P_G(\vec{\mathbf{p}})$.

Similarly, if $\vec{\mathbf{h}} := P_G^{\perp}\vec{\mathbf{p}}$, and if $H := \vec{\mathbf{h}}\vec{\mathbf{q}}_*^{\mathrm{T}}$, then $\|H\|_{\mathrm{F}} = \|H\|_{2,2} = \|\vec{\mathbf{h}}\|_2$ are the smallest norms such that $(G + J + H)\vec{\mathbf{q}}_* = \vec{\mathbf{p}}$.

Moreover, since $U\vec{\mathbf{t}}_* \in \mathscr{R}(G)$ and $\vec{\mathbf{h}} \perp \mathscr{R}(G)$, it follows that for both norms $\|J + H\|^2 = \|J\|^2 + \|H\|^2$. Because $\mathscr{R}(J) \perp \mathscr{R}(H)$, and for every matrix $B \in \mathbb{M}_{m\times n}(\mathbb{R})$ and every unit vector $\vec{\mathbf{q}} \in \mathscr{S}^{n-1}$ such that $(G + B)\vec{\mathbf{q}} = \vec{\mathbf{p}}$,

$$\|B\|_{\mathrm{F}}^2 \geqslant \|B\|_2^2 \geqslant \|B\vec{\mathbf{q}}\|_2^2 = \|(P_G\vec{\mathbf{p}} - G\vec{\mathbf{q}}) + P_G^{\perp}\vec{\mathbf{p}}\|_2^2 \tag{3.2}$$

$$= \|P_G\vec{\mathbf{p}} - G\vec{\mathbf{q}}\|_2^2 + \|P_G^{\perp}\vec{\mathbf{p}}\|_2^2 \geqslant \|J\|_2^2 + \|H\|_2^2 = \|J + H\|_2^2 \tag{3.3}$$

$$= \|J\|_2^2 + \|H\|_2^2 = \|J\|_{\mathrm{F}}^2 + \|H\|_{\mathrm{F}}^2 = \|J + H\|_{\mathrm{F}}^2. \tag{3.4}$$

Thus $B := J + H$ has the smallest norm with $(G + B)\vec{\mathbf{q}} = \vec{\mathbf{p}}$ for $\vec{\mathbf{q}} \in \mathscr{S}^{n-1}$. □

4. Fitting conics of specific types to data

The following algorithm is a variant of one by Gander, Golub, and Strebel's [12, p. 564], with an option to fit parabolae in general positions. To guarantee the invariance under translations, the algorithm first subtracts the mean

$$\bar{\mathbf{x}} := \frac{1}{N}\sum_{i=1}^{N}\vec{\mathbf{x}}_i \tag{4.1}$$

from each data point $\vec{\mathbf{x}}_i$, thus producing centered data defined by

$$(\check{x}_i, \check{y}_i)^{\mathrm{T}} = \check{\mathbf{x}}_i := \vec{\mathbf{x}}_i - \bar{\mathbf{x}}. \tag{4.2}$$

Then the algorithm minimizes Gander et al.'s objective (0.10),

$$F(\vec{\mathbf{w}}) := \|R_{2,\mathrm{II}}\vec{\mathbf{w}}\|_2, \tag{4.3}$$

with Bookstein's constraint [12, p. 564], thus minimizing F subject to

$$\|\vec{\mathbf{w}}\|_2^2 = 1. \tag{4.4}$$

However, because of the change of parameters Z related to the constraint $\det(A) = 0$ in Section 1, the algorithm presented here performs this change of variables at the outset, producing (instead of M) a matrix of monomials

$$\check{M} := \begin{pmatrix} 1 & \check{x}_1 & \check{y}_1 & \check{y}_1^2 - \check{x}_1^2 & 2\check{x}_1\check{y}_1 & \check{y}_1^2 + \check{x}_1^2 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & \check{x}_N & \check{y}_N & \check{y}_N^2 - \check{x}_N^2 & 2\check{x}_N\check{y}_N & \check{y}_N^2 + \check{x}_N^2 \end{pmatrix}. \tag{4.5}$$

Pratt uses the same matrix $\check{M}$, but with a different constraint [21, p. 150]. By Golub et al.'s theorem [13], minimizing F subject to $\|\vec{\mathbf{w}}\|_2^2 = 1$ amounts to determining the singular matrix $\widetilde{M}$ closest to $\check{M}$ with the first three columns kept constant. To this end, let $\check{M}_j$ be the jth column of $\check{M}$, and let $\check{M}_{k-\ell}$ consist of columns k through ℓ

of $\check{M}$. The constraint (4.4) thus corresponds to the partition $\check{M} = [\check{M}_1; \check{M}_{2-3}; \check{M}_{4-6}]$ with

$$\check{M}_1 := \vec{\mathbf{1}} := \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}, \quad \check{M}_{2-3} := \begin{pmatrix} \check{x}_1 & \check{y}_1 \\ \vdots & \vdots \\ \check{x}_N & \check{y}_N \end{pmatrix}, \tag{4.6}$$

$$\check{M}_{4-6} := \begin{pmatrix} \check{y}_1^2 - \check{x}_1^2 & 2\check{x}_1\check{y}_1 & \check{y}_1^2 + \check{x}_1^2 \\ \vdots & \vdots & \vdots \\ \check{y}_N^2 - \check{x}_N^2 & 2\check{x}_N\check{y}_N & \check{y}_N^2 + \check{x}_N^2 \end{pmatrix}. \tag{4.7}$$

With centered data, $\sum_{i=1}^{m} \check{x}_i = 0 = \sum_{i=1}^{m} \check{y}_i$, so that $\check{M}_1 = \vec{\mathbf{1}} \perp \check{M}_{2-3}$. Consequently, $\operatorname{rank}(\check{M}_{1-3}) = 1 + \operatorname{rank}(\check{M}_{2-3})$. Moreover, $\check{M}_{2-3}$ is also the matrix used to fit a straight line to data by total least-squares [7]. Hence computing $\operatorname{rank}(\check{M}_{1-3})$ amounts to determining whether $1/\kappa_{2,2}(\check{M}_{2-3})$ is numerically negligible. If so, then the process must stop, because $\operatorname{rank}(\check{M}) \leqslant 3$, and then $R_{2,\mathrm{II}} = 0$ identically. Indeed, the rank remains invariant under changes of coordinates, and if the line lies on the first coordinate axis, then $\check{M}_{3-5} = 0$.

If $\check{M}_{2-3}$ is not numerically singular, then $\check{M}$ can still be singular, indeed $\check{M}$ is (nearly) singular if the data lie (nearly) on a common conic, which can needlessly complicate the perturbation analysis of its QR factorization. However, the success of Gander et al.'s strategy [12, p. 564] does not require a complete factorization $\check{M} = \check{Q}\check{R}$. Indeed, Golub et al.'s theorem [13, p. 319] requires only the subtraction from $\check{M}_{4-6}$ of its orthogonal projection on $\check{M}_{1-3}$. To this end, the algorithm presented here need factor only $\check{M}_{1-3}$, in the form

$$\check{M}_{1-3} = \check{Q}_3\check{R}_3 = \check{Q}_3 \begin{pmatrix} \check{R}_{1,1} \\ 0 \end{pmatrix}, \tag{4.8}$$

with $\check{Q}_3 \in \mathbb{M}_{N\times N}(\mathbb{R})$ orthogonal and $\check{R}_{1,1} \in \mathbb{M}_{3\times 3}(\mathbb{R})$ upper triangular [14, Section 5.2]. The algorithm then applies $\check{Q}_3^{\mathrm{T}}$ to all of $\check{M}$ to produce

$$\check{Q}_3^{\mathrm{T}}\check{M} = \check{R} = \begin{pmatrix} \check{R}_{1,1} & \check{R}_{1,2} \\ 0 & \check{R}_{2,2} \end{pmatrix}, \tag{4.9}$$

where $\check{R}_{2,2} \in \mathbb{M}_{(N-3)\times 3}(\mathbb{R})$ need *not* be upper triangular. The solution $\vec{\mathbf{w}}$ is then a right-singular vector for the smallest singular value of $\check{R}_{2,2}$, and then $\vec{\mathbf{v}} := (c, 2\vec{\mathbf{b}}^{\mathrm{T}})^{\mathrm{T}}$ is the solution of the upper triangular system $\check{R}_{1,1}\vec{\mathbf{v}} = -\check{R}_{1,2}\vec{\mathbf{w}}$.

If the solution $(c, 2\vec{\mathbf{b}}, \vec{\mathbf{w}})$ does not correspond to a conic of the specified type, then the constraint $\det(A) = 0$ must be activated, and the algorithm proceeds as outlined in Section 2. The resulting algorithm proceeds as follows.

Algorithm 4.1 (*To fit conics by algebraic total least-squares*).

```
PROCEDURE conic($N, \vec{\mathbf{x}}_1, \ldots, \vec{\mathbf{x}}_N$).
DATA: A positive integer $N \in \mathbb{N}^*$, and a sequence $(\vec{\mathbf{x}}_1, \ldots, \vec{\mathbf{x}}_N) \in (\mathbb{R}^2)^N$.
RESULTS: coefficients of the equation of the fitted conic $(c, 2\vec{\mathbf{b}}, \vec{\mathbf{w}}, \bar{\mathbf{x}}) \in \mathbb{R}^9$.
START
Compute the mean $\bar{\mathbf{x}} := \frac{1}{N} \sum_{i=1}^{N} \vec{\mathbf{x}}_i$;
for each $i \in \{1, \ldots, N\}$, center the data by
  $(\check{x}_i, \check{y}_i)^{\mathrm{T}} := \check{\mathbf{x}}_i := \vec{\mathbf{x}}_i - \bar{\mathbf{x}}$;
end for;
form the matrix $\check{M} = [\check{M}_1; \check{M}_{2-3}; \check{M}_{4-6}]$ as in equations (4.5), (4.6), (4.7);
compute the singular-value decomposition of $\check{M}_{2-3} \in \mathbb{M}_{N\times 2}(\mathbb{R})$;
if $1/\kappa_{2,2}(\check{M}_{2-3}) = 0$, then
  fit the TLS line, with equation $(\vec{\mathbf{x}} - \bar{\mathbf{x}})^{\mathrm{T}} \vec{\mathbf{v}}_2 = 0$:
  $\vec{\mathbf{b}} := \vec{\mathbf{v}}_2$, $c := 0$, $\vec{\mathbf{w}} := \vec{\mathbf{0}}$;
else, if $\sigma_2(\check{M}_{2-3})/\sigma_1(\check{M}_{2-3}) > 0$, then
  factor $\check{M}_{1-3} = \check{Q}_3 \check{R}_3$ as in equation (4.8);
  compute $\check{R} := \check{Q}_3^{\mathrm{T}} \check{M}$ as in equation (4.9);
  if the desired conic is elliptic or hyperbolic, then
    compute the SVD of $\check{R}_{2,2} = \check{R}(4-N; 3-6) \in \mathbb{M}_{(N-3)\times 3}(\mathbb{R})$ in equation (4.9);
    let $\vec{\mathbf{q}} \in \mathbb{R}^3$ be a right-singular vector for the smallest singular value of $\check{R}_{2,2}$;
    identify the type of conic through trace$(A) = q_3$ and det$(A) = (1/2)\vec{\mathbf{q}}^{\mathrm{T}} Z \vec{\mathbf{q}}$;
  end if;
  if the type just obtained does not match the specified type,
  or if the desired conic is parabolic, then
    $G := (\check{R}_{2,2})_{1-2}$,
    $\vec{\mathbf{p}} := -\check{R}_{2,2} \vec{\mathbf{e}}_3$,
    $(q_1, q_2)^{\mathrm{T}} :=$ geodetic($N - 3, G, \vec{\mathbf{p}}$) with Algorithm 2.5,
    $\vec{\mathbf{q}} := (q_1, q_2, 1)^{\mathrm{T}}$,
  end if;
  solve the upper-triangular system $\check{R}_{1,1}(c, 2\vec{\mathbf{b}}) = -\check{R}_{1,2}\vec{\mathbf{q}}$;
  change coordinates to $\vec{\mathbf{w}} := (1/\sqrt{2}) Z \vec{\mathbf{q}}$;
end if;
return conic($N, \vec{\mathbf{x}}_1, \ldots, \vec{\mathbf{x}}_N$) := $(c, 2\vec{\mathbf{b}}, \vec{\mathbf{w}}, \bar{\mathbf{x}})$.
STOP.
```

5. Perturbations of projected submatrices

Perturbing the data from an initial sequence $\vec{\mathbf{x}}_1, \ldots, \vec{\mathbf{x}}_N$ to a new sequence $\vec{\mathbf{x}}'_1, \ldots, \vec{\mathbf{x}}'_N$ also perturbs the matrix $\check{M}$ to a new matrix $\check{M}'$. Each such perturbation of the data decomposes into the sum of a translation of the mean $\Delta\overline{\mathbf{x}} := \overline{\mathbf{x}}' - \overline{\mathbf{x}}$ and a difference matrix $\Delta\check{M}_{2-3}$, which contains the perturbation of the centered data. Because the singular values and the *right*-singular vectors of $\check{M}_{2-3}$ form the principal semi-axes of the distribution of the data [6, p. 278], their perturbations correspond to a rotation and two dilations of the data. These perturbations then propagate from the parameters $\vec{\mathbf{v}}$, $\vec{\mathbf{w}}$ to $\vec{\mathbf{v}}'$, $\vec{\mathbf{w}}'$ through the orthogonal projection $P\check{M}_{4-6}$ of $\check{M}_{4-6}$ on $\mathscr{R}(\check{M}_{2-3})$. The first theorem pertains to the norm of the composition of two projections.

Theorem 5.1. *For all orthogonal projections* P, P', *if* $\|P - P'\|_2 < 1$, *then*

$$\|P^{\perp}P'\|_2 = \|P - P'\|_2 = \|P'P^{\perp}\|_2. \tag{5.1}$$

Proof. A theorem of Kato's [16, Chapter I, Theorem 6.34, pp. 56–58] asserts that if $\|P - P'\|_2 < 1$, then $\|P - P'\|_2 = \|(I - P')P\|_2 = \|(I - P)P'\|_2 = \|P^{\perp}P'\|_2$. The same inequalities hold for the complementary projections $P^{\perp} = I - P$ and $P'^{\perp} = I - P'$, because $\|(I - P) - (I - P')\|_2 = \|P - P'\|_2 < 1$, whence $\|(I - P) - (I - P')\|_2 = \|P(I - P')\|_2 = \|P'(I - P)\|_2 = \|P'P^{\perp}\|_2$. □

For matrices G, G' of rank r, the second theorem shows that the factor

$$\tilde{\kappa}(G) := \frac{\sigma_1(G)}{\sigma_{r-1}(G) - \sigma_r(G)} \tag{5.2}$$

plays the role of a "condition number" [3, p. 38] for the perturbation $\Delta\vec{\mathbf{v}}_r := \vec{\mathbf{v}}_r(G') - \vec{\mathbf{v}}_r(G)$ of the last right-singular vector $\vec{\mathbf{v}}_r(G)$.

Theorem 5.2. *For all matrices* $G, G' \in \mathbb{M}_{m\times n}(\mathbb{R})$ *of rank* $r \leqslant \min\{m, n\}$,

$$\limsup_{\|\Delta G\|_2/\|G\|_2 \to 0} \frac{\|\Delta\vec{\mathbf{v}}_r\|_2/\|\vec{\mathbf{v}}_r\|_2}{\|\Delta G\|_2/\|G\|_2} \leqslant \frac{\sigma_1(G)}{\sigma_{r-1}(G) - \sigma_r(G)}. \tag{5.3}$$

Proof. If θ denotes the angle between $\vec{\mathbf{v}}_r$ and $\vec{\mathbf{v}}_r'$, then Wedin's *generalized* $\sin(\theta)$ *theorem* [27, p. 262;32, p. 102] applied to $G := G_1 + G_0 := \sum_{i=1}^{r-1} \sigma_i \vec{\mathbf{u}}_i \vec{\mathbf{v}}_i^{\mathrm{T}} + \sigma_r \vec{\mathbf{u}}_r \vec{\mathbf{v}}_r^{\mathrm{T}}$ and Kato's theorem (in the form of Theorem 5.1) give

$$\begin{aligned} \left\| P_{G_1'}^{\perp} - P_{G_1}^{\perp} \right\|_2 &= \left\| P_{G_1}^{\perp} P_{G_1'} \right\|_2 = |\sin(\theta)| \\ &\leqslant \frac{\sigma_1(G)}{\sigma_{r-1}(G) - \sigma_r(G)} \cdot \frac{\|\Delta G\|_2}{\|G\|_2}. \end{aligned} \tag{5.4}$$

Thus, if $\|\Delta G\|_2/\|G\|_2$ tends to 0, then so does $|\sin(\theta)|$, and $|2\sin(\theta/2)/\sin(\theta)|$ tends to 1. Moreover, $\|\Delta\vec{\mathbf{v}}_r\|_2/\|\vec{\mathbf{v}}_r\|_2 = 2\sin(\theta/2)$ because $\|\vec{\mathbf{v}}_r\|_2 = 1$. Combining these equalities with inequality (5.4) yields inequality (5.3). □

Theorems 5.1 and 5.2 together provide estimates of the perturbations of right-singular vectors of the composition of a projection and another matrix. In the present context, however, the perturbations of the data appear in $\check{M}_{2-3}$ while the solution is a right-singular vector of $P^{\perp}\check{M}_{4-6}$. To make the transition from $\check{M}_{2-3}$ to $\check{M}_{4-6}$, the following theorems relate perturbations of the data, recorded in $\Delta\check{M}_{2-3}$, to perturbations of the variances and covariance, recorded in $\Delta\check{M}_{4-6}$. The proofs repeatedly use the following consequences of the Cauchy-Schwartz inequality [14, p. 54]. For every vector $\vec{\mathbf{z}} \in \mathbb{R}^N$,

$$\|\vec{\mathbf{z}}\|_1^2 := \left[\sum_{i=1}^{N} |z_i|\right]^2 = \left[\sum_{i=1}^{N} 1 \cdot |z_i|\right]^2 \leqslant \sum_{i=1}^{N} 1^2 \cdot \sum_{i=1}^{N} |z_i|^2 = N \cdot \|\vec{\mathbf{z}}\|_2^2. \tag{5.5}$$

In particular, with $N = 2$, inequality (5.5) gives

$$(|p| + |q|)^2 \leqslant 2(p^2 + q^2). \tag{5.6}$$

The next theorem establishes bounds for the norms of the submatrices

$$\check{M}_{4-5} := \begin{pmatrix} \check{y}_1^2 - \check{x}_1^2 & 2\check{x}_1\check{y}_1 \\ \vdots & \vdots \\ \check{y}_N^2 - \check{x}_N^2 & 2\check{x}_N\check{y}_N \end{pmatrix}, \quad \check{M}_6 := \begin{pmatrix} \check{y}_1^2 + \check{x}_1^2 \\ \vdots \\ \check{y}_N^2 + \check{x}_N^2 \end{pmatrix} \tag{5.7}$$

in terms of the norm of $\check{M}_{2-3}$.

Theorem 5.3. *With the notation defined by Eqs.* (4.6), (4.7), (5.7),

$$\frac{\|\check{M}_{2-3}\|_{\mathrm{F}}^2}{\sqrt{N}} \leqslant \|\check{M}_{4-6}\|_{\mathrm{F}} = \sqrt{2}\|\check{M}_{4-5}\|_{\mathrm{F}} = \sqrt{2}\|\check{M}_6\|_{\mathrm{F}} \leqslant 2\|\check{M}_{2-3}\|_{\mathrm{F}}^2. \tag{5.8}$$

Proof. The proof uses the equalities $\|\check{M}_{4-6}\|_{\mathrm{F}}^2 = \|\check{M}_{4-5}\|_{\mathrm{F}}^2 + \|\check{M}_6\|_{\mathrm{F}}^2$ and

$$\|\check{M}_{4-5}\|_{\mathrm{F}}^2 = \sum_{i=1}^{N} \left[(\check{y}_i^2 - \check{x}_i^2)^2 + 4\check{y}_i^2\check{x}_i^2\right] = \sum_{i=1}^{N} \left[(\check{y}_i^2 + \check{x}_i^2)^2\right] = \|\check{M}_6\|_{\mathrm{F}}^2. \tag{5.9}$$

The upper bound in inequality (5.8) then follows from inequality (5.6) and

$$\begin{align} \|\check{M}_6\|_{\mathrm{F}}^2 &= \sum_{i=1}^{N} (\check{y}_i^2 + \check{x}_i^2)^2 \tag{5.10} \\ &\leqslant 2\sum_{i=1}^{N} (\check{y}_i^4 + \check{x}_i^4) \leqslant 2\sum_{i=1}^{N} (\check{y}_i^2 + \check{x}_i^2)^2 \tag{5.11} \\ &\leqslant 2\left[\sum_{i=1}^{N} (\check{y}_i^2 + \check{x}_i^2)\right]^2 \tag{5.12} \\ &= 2\|\check{M}_{2-3}\|_{\mathrm{F}}^4. \tag{5.13} \end{align}$$

The lower bound in inequality (5.8) then follows from inequality (5.5) and inequality (5.6) with $z_i := \check{y}_i^2 + \check{x}_i^2 = (\check{M}_{2-3})_{i,1}^2 + (\check{M}_{2-3})_{i,2}^2$, which yield

$$\|\check{M}_{2-3}\|_{\mathrm{F}}^4 = \left[\sum_{i=1}^{N}(\check{y}_i^2 + \check{x}_i^2)\right]^2 \leqslant N \cdot \sum_{i=1}^{N}\left[(\check{y}_i^2 + \check{x}_i^2)^2\right] \tag{5.14}$$

$$\leqslant N \cdot \sum_{i=1}^{N}\left[(\check{y}_i^2 - \check{x}_i^2)^2 + 4\check{y}_i^2\check{x}_i^2 + (\check{y}_i^2 + \check{x}_i^2)^2\right] \tag{5.15}$$

$$= N \cdot \|\check{M}_{4-6}\|_{\mathrm{F}}^2. \qquad \square \tag{5.16}$$

The following theorem establishes upper bounds for the perturbations.

Theorem 5.4. *With the notation defined by Eqs.* (4.6), (4.7), (5.7),

$$\limsup_{\|\Delta\check{M}_{2-3}\|_2/\|\check{M}_{2-3}\|_2 \to 0} \frac{\|\Delta\check{M}_{4-6}\|_2/\|\check{M}_{4-6}\|_2}{\|\Delta\check{M}_{2-3}\|_2/\|\check{M}_{2-3}\|_2} \leqslant 12N, \tag{5.17}$$

$$\limsup_{\|\Delta\check{M}_{2-3}\|_2/\|\check{M}_{2-3}\|_2 \to 0} \frac{\|\Delta\check{M}_{6}\|_2/\|\check{M}_{6}\|_2}{\|\Delta\check{M}_{2-3}\|_2/\|\check{M}_{2-3}\|_2} \leqslant 2\sqrt{2}N, \tag{5.18}$$

$$\limsup_{\|\Delta\check{M}_{2-3}\|_2/\|\check{M}_{2-3}\|_2 \to 0} \frac{\|\Delta\check{M}_{4-5}\|_2/\|\check{M}_{4-5}\|_2}{\|\Delta\check{M}_{2-3}\|_2/\|\check{M}_{2-3}\|_2} \leqslant 4\sqrt{2}N. \tag{5.19}$$

Proof. The proof repeatedly uses the following inequalities, (5.20), (5.21). For every matrix $G \in \mathbb{M}_{m\times n}(\mathbb{R})$ with rank r [14, p. 57],

$$\|G\|_2^2 = [\sigma_1(G)]^2 \leqslant \sum_{i=1}^{r}[\sigma_i(G)]^2 = \|G\|_{\mathrm{F}}^2 \leqslant r[\sigma_1(G)]^2 = r\|G\|_2^2. \tag{5.20}$$

Moreover, for every entry $G_{k,\ell}$ of every matrix G [28, p. 256, #(f)],

$$|G_{k,\ell}|^2 \leqslant \sum_{i=1}^{m}|G_{i,\ell}|^2 = \vec{\mathbf{e}}_\ell^{\mathrm{T}} G^{\mathrm{T}} G \vec{\mathbf{e}}_\ell \leqslant [\sigma_1(G)]^2 = \|G\|_2^2. \tag{5.21}$$

From $z^2 - (z')^2 = (z - z')(z + z')$ and inequality (5.6) follows the inequality

$$\|\Delta\check{M}_6\|_{\mathrm{F}}^2 = \sum_{i=1}^{N}\left[(\Delta\check{y}_i)(\check{y}_i + \check{y}_i') + (\Delta\check{x}_i)(\check{x}_i + \check{x}_i')\right]^2 \tag{5.22}$$

$$\leqslant 2\sum_{i=1}^{N}\left[(\Delta\check{y}_i)^2(\check{y}_i + \check{y}_i')^2 + (\Delta\check{x}_i)^2(\check{x}_i + \check{x}_i')^2\right]. \tag{5.23}$$

Because $\lim_{\Delta\check{M}_{2-3}\to 0}\check{x}'_i = \check{x}_i$ and $\lim_{\Delta\check{M}_{2-3}\to 0}\check{y}'_i = \check{y}_i$, (5.21) and (5.23) give

$$\limsup_{\frac{\|\Delta\check{M}_{2-3}\|_2}{\|\check{M}_{2-3}\|_2}\to 0} \frac{\|\Delta\check{M}_6\|_{\mathrm{F}}^2}{\|\Delta\check{M}_{2-3}\|_{\mathrm{F}}^2} \leqslant 8\max_{k,\ell}|(\check{M}_{2-3})_{k,\ell}|^2 \leqslant 8\|\check{M}_{2-3}\|_2^2. \tag{5.24}$$

Similarly, from $xy - x'y' = (x - x')y + (y - y')x'$, the formulae for the entries

$$\begin{aligned}(\Delta\check{M}_{4-6})_{i,1}^2 + (\Delta\check{M}_{4-6})_{i,2}^2 = {} & \left[(\Delta\check{y}_i)(\check{y}_i + \check{y}'_i) - (\Delta\check{x}_i)(\check{x}_i + \check{x}'_i)\right]^2 \\ & + 4\left[(\Delta\check{y}_i)\check{x}_i + (\Delta\check{x}_i)\check{y}'_i\right]^2,\end{aligned} \tag{5.25}$$

with $\lim_{\Delta\check{M}_{2-3}\to 0}\check{x}'_i = \check{x}_i$ and $\lim_{\Delta\check{M}_{2-3}\to 0}\check{y}'_i = \check{y}_i$, and inequality (5.21), give

$$\limsup_{\frac{\|\Delta\check{M}_{2-3}\|_2}{\|\check{M}_{2-3}\|_2}\to 0} \frac{\|\Delta\check{M}_{4-5}\|_{\mathrm{F}}^2}{\|\Delta\check{M}_{2-3}\|_{\mathrm{F}}^2} \leqslant 8\max_{k,\ell}|(\check{M}_{2-3})_{k,\ell}|^2 \leqslant 8\|\check{M}_{2-3}\|_2^2. \tag{5.26}$$

Combining (5.20), (5.8), (5.24), with $r = 1$ for $\check{M}_6$ leads to inequality (5.18). Likewise, combining (5.20), (5.8), (5.26), with $r = 2$ for $\check{M}_{4-5}$ leads to inequality (5.19). Moreover, combining inequalities (5.26) and (5.28) with

$$\|\Delta\check{M}_{4-6}\|_{\mathrm{F}}^2 = \|\Delta\check{M}_{4-5}\|_{\mathrm{F}}^2 + \|\Delta\check{M}_6\|_{\mathrm{F}}^2 \tag{5.27}$$

gives

$$\limsup_{\frac{\|\Delta\check{M}_{2-3}\|_2}{\|\check{M}_{2-3}\|_2}\to 0} \frac{\|\Delta\check{M}_{4-6}\|_{\mathrm{F}}^2}{\|\Delta\check{M}_{2-3}\|_{\mathrm{F}}^2} \leqslant 16\max_{k,\ell}|(\check{M}_{2-3})_{k,\ell}|^2 \leqslant 16\|\check{M}_{2-3}\|_2^2. \tag{5.28}$$

Hence inequalities (5.20), (5.16), (5.28) with $r = 3$ for $\check{M}_{4-6}$ yield (5.17). □

Combining the foregoing results, the final theorem relates perturbations of the data, in $\Delta\check{M}_{2-3}$, to perturbations of the projected marix $P^{\perp}\check{M}_{4-6}$.

Theorem 5.5. *For the orthogonal projection* $P := P_{\check{M}_{2-3}}$ *on* $\mathscr{R}(\check{M}_{2-3})$,

$$\limsup_{\frac{\|\Delta\check{M}_{2-3}\|_2}{\|\check{M}_{2-3}\|_2}\to 0} \frac{\|P'^{\perp}\check{M}'_{4-6} - P^{\perp}\check{M}_{4-6}\|_2/\|\check{M}_{4-6}\|_2}{\|\Delta\check{M}_{2-3}\|_2/\|\check{M}_{2-3}\|_2} \leqslant \kappa_{2,2}(\check{M}_{2-3}) + 12N. \tag{5.29}$$

Similar bounds hold for $\check{M}_{4-5}$ *and* $\check{M}_6$, *with* $\sqrt{8}N$ *and* $\sqrt{32}N$ *instead of* $12N$.

Proof. Apply Wedin's *generalized* $\sin(\theta)$ *theorem* [27, p. 262;32, p. 102] to $G := G_1 + G_0 := \check{M}_{2-3} + 0$ in the form of Theorem 5.2, with $\|P^{\perp}\|_2 = 1$:

$$\frac{\|P^{\perp} - P'^{\perp}\|_2}{\|P^{\perp}\|_2} \leqslant \kappa_{2,2}(\check{M}_{2-3})\frac{\|\Delta\check{M}_{2-3}\|_2}{\|\check{M}_{2-3}\|_2}. \tag{5.30}$$

The triangle inequality and the submultiplicativity of norms [16, p. 26] gives

$$\begin{aligned}\|P'^{\perp}\check{M}'_{4-6} - P^{\perp}\check{M}_{4-6}\|_2 \leqslant{}& \|P'^{\perp} - P^{\perp}\|_2 \cdot \|\check{M}_{4-6}\|_2\\ &+ \|P'^{\perp}\|_2 \cdot \|\check{M}'_{4-6} - \check{M}_{4-6}\|_2.\end{aligned} \tag{5.31}$$

Combining inequalities (5.30) and (5.31) gives

$$\frac{\|P'^{\perp}\check{M}'_{4-6} - P^{\perp}\check{M}_{4-6}\|_2}{\|\check{M}_{4-6}\|_2} \leqslant \kappa_{2,2}(\check{M}_{2-3})\frac{\|\Delta\check{M}_{2-3}\|_2}{\|\check{M}_{2-3}\|_2} + \frac{\|\Delta\check{M}_{4-6}\|_2}{\|\check{M}_{4-6}\|_2}. \tag{5.32}$$

Hence inequality (5.17) in Theorem 5.4 yields inequality (5.29). Similar bounds for $\check{M}_{4-5}$ and $\check{M}_6$ follow from Theorem 5.4 in the same manner. □

6. Perturbation analysis for fitted conics

By Golub et al.'s theorem [13], $\vec{\mathbf{w}}$ is a *right*-singular vector for the smallest singular value of $\check{R}_{2,2}$, or of any matrix representing—relative to any bases—the orthogonal projection of $\check{M}_{4-6}$ on the orthogonal complement of the range of $\check{M}_{1-3} = [\check{M}_1; \check{M}_{2-3}]$. However, perturbations of the data do not affect the constant first column $\check{M}_1 = \vec{\mathbf{1}}$. To separate $\check{M}_1$ from the submatrix $\check{M}_{2-3}$ subject to perturbations, let $H \in \mathbb{M}_{N\times N}(\mathbb{R})$ denote the Householder reflection [14, p. 196] that maps $\vec{\mathbf{1}} \in \mathbb{R}^N$ to $-\sqrt{N}\vec{\mathbf{e}}_1 \in \mathbb{R}^N$. Because H is orthogonal, and because $\check{M}_1 = \vec{\mathbf{1}}$ and $\check{M}_{2-3}$ are mutually orthogonal, so are $H(\vec{\mathbf{1}})$ and $H(\check{M}_{2-3})$, so that $H(\check{M}_{1-3})$ has the form

$$H(\check{M}_{1-3}) = \left[\begin{pmatrix}-\sqrt{N}\\ \vec{\mathbf{0}}\end{pmatrix} \quad H(\check{M}_{2-3})\right]. \tag{6.1}$$

Moreover, for each vector $\vec{\mathbf{h}} \in \mathbb{R}^N$ with mean $\overline{h} := (1/N)\sum_{i=1}^{N} h_i$, because $\vec{\mathbf{1}} \perp \vec{\mathbf{h}} - \overline{h}\vec{\mathbf{1}}$ it follows that $-\sqrt{N}\vec{\mathbf{e}}_1 = H(\vec{\mathbf{1}}) \perp H(\vec{\mathbf{h}} - \overline{h}\vec{\mathbf{1}}) = (0, *, \ldots, *)^{\mathrm{T}}$:

$$H(\vec{\mathbf{h}}) = \begin{pmatrix}-\sqrt{N}\overline{h}\\ \vec{\mathbf{0}}\end{pmatrix} + H(\vec{\mathbf{h}} - \overline{h}\vec{\mathbf{1}}). \tag{6.2}$$

In particular, $P_{\vec{\mathbf{1}}}^{\perp}\check{M}_{4-6}$ forms rows 2 through N of $H(\check{M}_{4-6})$:

$$H(\check{M}) = \left[\begin{pmatrix}-\sqrt{N}\\ \vec{\mathbf{0}}\end{pmatrix} \quad H(\check{M}_{2-3}) \quad \begin{pmatrix}P_{\vec{\mathbf{e}}_1}H(\check{M}_{4-6})\\ P_{\vec{\mathbf{e}}_1}^{\perp}H(\check{M}_{4-6})\end{pmatrix}\right]. \tag{6.3}$$

In (6.1)–(6.3), the first row of $H(\check{M}_{2-3})$ and $H(\vec{\mathbf{h}} - \overline{h}\vec{\mathbf{1}})$ is zero. Because the reflection H does not depend on the data, subtracting from $\check{M}_{4-6}$ its projection on $\mathscr{R}(\check{M}_{2-3})$ amounts to subtracting from $P_{\vec{\mathbf{e}}_1}^{\perp}H(\check{M}_{4-6})$ its projection on $\mathscr{R}(H(\check{M}_{2-3}))$. Consequently, perturbation analyses for $\vec{\mathbf{w}}$ may restrict themselves to perturbations of the projection $P^{\perp}$ on $\mathscr{R}(\check{M}_{2-3})^{\perp}$.

Theorem 6.1. *With the notation defined by Eqs.* (4.6), (4.7), (5.2),

$$\limsup_{\|\Delta \check{M}_{2-3}\|_2/\|\check{M}_{2-3}\|_2 \to 0} \frac{\|\Delta \vec{\mathbf{w}}\|_2/\|\vec{\mathbf{w}}\|_2}{\|\Delta \check{M}_{2-3}\|_2/\|\check{M}_{2-3}\|_2}$$

$$\leqslant [\kappa_{2,2}(\check{M}_{2-3}) + 12N]\tilde{\kappa}(\check{R}_{2,2})\frac{\sigma_1(\check{M}_{4-6})}{\sigma_1(\check{R}_{2,2})}. \tag{6.4}$$

Proof. Because $\vec{\mathbf{w}} = \vec{\mathbf{v}}_3(\check{R}_{2,2}) = \vec{\mathbf{v}}_3(P^{\perp}\check{M}_{4-6})$ Theorem 5.2 gives

$$\limsup_{\|\Delta(P^{\perp}\check{M}_{4-6})\|_2/\|P^{\perp}\check{M}_{4-6}\|_2 \to 0} \frac{\|\Delta \vec{\mathbf{w}}\|_2/\|\vec{\mathbf{w}}\|_2}{\|\Delta(P^{\perp}\check{M}_{4-6})\|_2/\|P^{\perp}\check{M}_{4-6}\|_2}$$

$$\leqslant \frac{\sigma_1(P^{\perp}\check{M}_{4-6})}{\sigma_{r-1}(P^{\perp}\check{M}_{4-6}) - \sigma_r(P^{\perp}\check{M}_{4-6})}. \tag{6.5}$$

Also, Theorem 5.5 gives

$$\limsup_{\|\Delta \check{M}_{2-3}\|_2/\|\check{M}_{2-3}\|_2 \to 0} \frac{\|\Delta(P^{\perp}\check{M}_{4-6})\|_2/\|P^{\perp}\check{M}_{4-6}\|_2}{\|\Delta \check{M}_{2-3}\|_2/\|\check{M}_{2-3}\|_2}$$

$$\leqslant [\kappa_{2,2}(\check{M}_{2-3}) + 12N] \cdot \frac{\|\check{M}_{4-6}\|_2}{\|P^{\perp}\check{M}_{4-6}\|_2}. \tag{6.6}$$

The result (6.4) follows from inequalities (6.5) and (6.6). □

The perturbations $\Delta\vec{\mathbf{w}} := \vec{\mathbf{w}}' - \vec{\mathbf{w}}$ and $\Delta\check{R}_{1,2} := \check{R}'_{1,2} - \check{R}_{1,2}$ then induce a perturbation $\Delta\vec{\mathbf{y}} := -\check{R}'_{1,2}\vec{\mathbf{w}}' + \check{R}_{1,2}\vec{\mathbf{w}}$ in the linear system $\check{R}_{1,1}\vec{\mathbf{v}} = -\check{R}_{1,2}\vec{\mathbf{w}}$, or, equivalently, $Q^{\mathrm{T}}H\check{M}_{1-3}\vec{\mathbf{v}} = -Q^{\mathrm{T}}H\check{M}_{4-6}\vec{\mathbf{w}}$. While $\|\vec{\mathbf{w}}\|_2 = 1$ by design, $\vec{\mathbf{v}}$ may be zero. Therefore, the following theorem uses the size of the data, $\|\check{M}_{2-3}\|_2$ and $\|\check{M}_{4-6}\|_2$, as measures of scale for the affine term $\vec{\mathbf{v}}$. The first theorem gives estimates for the perturbations of the right-hand side $\check{M}_{4-6}\vec{\mathbf{w}}$.

Theorem 6.2. *With the notation defined by Eqs.* (4.6), (4.7), (5.2),

$$\limsup_{\|\Delta \check{M}_{2-3}\|_2/\|\check{M}_{2-3}\|_2 \to 0} \frac{\|\Delta(\check{M}_{4-6}\vec{\mathbf{w}})\|_2/\|\check{M}_{4-6}\|_2}{\|\Delta \check{M}_{2-3}\|_2/\|\check{M}_{2-3}\|_2} \tag{6.7}$$

$$\leqslant [\kappa_{2,2}(\check{M}_{2-3}) + 12N]\tilde{\kappa}(\check{R}_{2,2})\frac{\sigma_1(\check{M}_{4-6})}{\sigma_1(\check{R}_{2,2})} + 12N. \tag{6.8}$$

Proof. With $\|\vec{\mathbf{w}}\|_2 = 1 = \|\vec{\mathbf{w}}'\|_2$, the proof uses the following relations:

$$\Delta(\check{M}_{4-6}\vec{\mathbf{w}}) = \check{M}_{4-6}(\vec{\mathbf{w}}' - \vec{\mathbf{w}}) + (\check{M}'_{4-6} - \check{M}_{4-6})\vec{\mathbf{w}}', \tag{6.9}$$

$$\frac{\|\Delta(\check{M}_{4-6}\vec{\mathbf{w}})\|_2}{\|\check{M}_{4-6}\|_2} \leqslant \frac{\|\Delta\vec{\mathbf{w}}\|_2}{\|\vec{\mathbf{w}}\|_2} + \frac{\|\Delta\check{M}_{4-6}\|_2}{\|\check{M}_{4-6}\|_2}. \tag{6.10}$$

Hence, Theorems 5.5 and 6.1 give inequality (6.8). □

The second theorem bounds the perturbation of the constant term.

Theorem 6.3. *With the notation defined by Eqs.* (4.6), (4.7), (5.2),

$$\limsup_{\|\Delta\check{M}_{2-3}\|_2/\|\check{M}_{2-3}\|_2\to 0} \frac{|\Delta c|/\|\check{M}_{4-6}\|_2}{\|\Delta\check{M}_{2-3}\|_2/\|\check{M}_{2-3}\|_2} \tag{6.11}$$

$$\leqslant [\kappa_{2,2}(\check{M}_{2-3}) + 12N]\frac{\tilde{\kappa}(\check{R}_{2,2})}{\sqrt{N}}\frac{\sigma_1(\check{M}_{4-6})}{\sigma_1(\check{R}_{2,2})} + 12\sqrt{N}. \tag{6.12}$$

Proof. From the block-matrix form of $H\check{M}_{1-3}$ in Eq. (6.1), the equation for the constant c, which adjusts the scale of the fitted conic, becomes

$$-\sqrt{N}c = (H\check{M}_{1-3}\vec{\mathbf{v}})_1 = -(H\check{M}_{4-6}\vec{\mathbf{w}})_1. \tag{6.13}$$

Inequality (6.8) then gives inequality (6.12). □

The third theorem bounds the perturbation of the linear term.

Theorem 6.4. *With the notation defined by Eqs.* (4.6), (4.7), (5.2),

$$\limsup_{\|\Delta\check{M}_{2-3}\|_2/\|\check{M}_{2-3}\|_2\to 0} \frac{\|\Delta\vec{\mathbf{b}}\|_2/\|\check{M}_{2-3}\|_2}{\|\Delta\check{M}_{2-3}\|_2/\|\check{M}_{2-3}\|_2}$$

$$\leqslant 2N\kappa_{2,2}(\check{M}_{2-3})\Bigg[\kappa_{2,2}(\check{M}_{2-3}) + \frac{1}{4}\left\{\Big[\kappa_{2,2}(\check{M}_{2-3}) + 12N\Big]\tilde{\kappa}(\check{R}_{2,2})\frac{\sigma_1(\check{M}_{4-6})}{\sigma_1(\check{R}_{2,2})} + 12N\right\}\Bigg]. \tag{6.14}$$

Proof. The solution $\vec{\mathbf{v}}$ is also the solution of the ordinary least-squares system $H\check{M}_{1-3}\vec{\mathbf{v}} = -H\check{M}_{4-6}\vec{\mathbf{w}}$. Because $\vec{\mathbf{b}}$ involves only the lower right 2 × 2 block $H\check{M}_{2-3}$ of $H\check{M}_{1-3}$, and because $\check{M}_{2-3}$ has full rank, another theorem of Wedin's [33, p. 224, Theorem 5.1] gives

$$\|\Delta\vec{\mathbf{b}}\|_2 \leqslant \frac{\kappa(\check{M}_{2-3})}{1-\kappa(\check{M}_{2-3})\frac{\|\Delta\check{M}_{2-3}\|_2}{\|\check{M}_{2-3}\|_2}}\left[\frac{\|\Delta\check{M}_{2-3}\|_2}{\|\check{M}_{2-3}\|_2}\|\vec{\mathbf{b}}\|_2 + \frac{\|\Delta(\check{M}_{4-6}\vec{\mathbf{w}})\|_2}{\|\check{M}_{2-3}\|_2}\right], \tag{6.15}$$

where $\|\check{M}_{2-3}\|_2^2 \geqslant \|\check{M}_{2-3}\|_{\mathrm{F}}/N \geqslant (1/2)\|\check{M}_{4-6}\|_{\mathrm{F}}/N \geqslant (1/2)\|\check{M}_{4-6}\|_2/N$ by Theorem 5.3, and where $\|\vec{\mathbf{b}}\|_2 = \|\check{M}_{2-3}^{\dagger}\check{M}_{4-6}\vec{\mathbf{w}}\|_2 \leqslant \kappa_{2,2}(\check{M}_{2-3})2N\|\check{M}_{2-3}\|_2$. Hence, inequality (6.8) yields inequality (6.14). □

Geometric interpretation for fitted conics. The perturbation bounds in Theorem 6.1 for $\vec{\mathbf{w}}$, determining the type of the fitted conic, Theorem 6.3 for c, determining its size, and Theorem 6.4 for $\vec{\mathbf{b}}$, determining its center, all contain the two condition numbers $\kappa_{2,2}(\check{M}_{2-3})$ and $\tilde{\kappa}(\check{R}_{2,2})$.

The factor $\tilde{\kappa}(\check{R}_{2,2}) = \tilde{\kappa}(P^{\perp}\check{M}_{4-6})$ diverges to infinity if and only if $\sigma_2(\check{R}_{2,2})$ tends to $\sigma_3(\check{R}_{2,2})$. Then $\vec{\mathbf{v}}_2(\check{R}_{2,2})$ approaches $\vec{\mathbf{w}} = \vec{\mathbf{v}}_3(\check{R}_{2,2})$, and at the limit any $\vec{\mathbf{w}} \in \operatorname{span}\{\vec{\mathbf{v}}_2, \vec{\mathbf{v}}_3\}$ defines the parameters of an equally well fitting conic with the same residual $\sigma_2(\check{R}_{2,2}) = \sigma_3(\check{R}_{2,2})$. Meanwhile the data near every conic and also the intersection of the pencil spanned by $\vec{\mathbf{v}}_2(\check{R}_{2,2})$ and $\vec{\mathbf{v}}_3(\check{R}_{2,2})$.

Because $1/\kappa_{2,2}(\check{M}_{2-3})$ measures the degree of colinearity of the data [14, p. 247], the leading factors $\kappa_{2,2}(\check{M}_{2-3})$ reveals that the sensitivity of the fitted conic also increases as does the colinearity of the data. Thus, instead of a factor $\kappa(\check{M})$ that would occur with the complete factorization $\check{M} = \check{Q}\check{R}$, the factor $\kappa_{2,2}(\check{M}_{2-3})$ corroborates a remark by Pratt, who had already observed the instability of conics fitted to nearly colinear data [21, p. 149].

7. Perturbation analysis for fitted parabolae

With $\psi(\lambda) := [1 - \varphi(\lambda)]\prod_{i=1}^{2}(\sigma_i^2 - \lambda)^2$, the secular equation (2.8) becomes

$$0 = \psi(\lambda) := \prod_{i=1}^{2}(\sigma_i^2 - \lambda)^2 - \sum_{i=1}^{2}\sigma_i^2 y_i^2 \prod_{j \neq i}(\sigma_j^2 - \lambda)^2 \tag{7.1}$$

$$=: \lambda^4 + \psi_3\lambda^3 + \psi_2\lambda^2 + \psi_1\lambda + \psi_0. \tag{7.2}$$

For the constrained problem for parabolic conics, perturbations of the data change the singular values of the matrix $G := P^{\perp}\check{M}_{4-5} = (\check{R}_{2,2})_{1-2}$, which affects the coefficients $\psi_0, \ldots, \psi_3$ of the polynomial secular equation (7.2) hence also its optimal solution λ_*, which thence perturbs the Lagrangian equation (2.3) and the solution $\vec{\mathbf{w}}_* = (1/\sqrt{2})Z\vec{\mathbf{q}}_* = (1/\sqrt{2})ZV\vec{\mathbf{z}}_*$.

The sensitivity of an isolated optimal solution λ_* of the monic polynomial secular equation (7.2) to perturbations of the coefficient ψ_k of λ^k is inversely proportional to the derivative at that solution [28, pp. 303–306;34, p. 11, Eq. (5.1)], as follows from the implicit-function theorem [9, p. 148]:

$$\frac{\partial\lambda_*}{\partial\psi_k} = \frac{-\lambda_*^k}{\psi'(\lambda_*)}, \tag{7.3}$$

which tends to infinity as the two smallest solutions $\lambda_* < \sigma_k^2 < \lambda_3$ coalesce. As in theorem 2.3, $\psi'(\lambda_*) = 0 = \psi(\lambda_*)$ if and only if $\lambda_* = \sigma_2^2$, in which case $\sigma_2^2 - \lambda_*$ is also a vanishing singular value of the matrix $L := \Sigma^{\mathrm{T}}\Sigma - \lambda_* I$ in system (2.3), and then the sensitivity of $\vec{\mathbf{w}}$ to perturbations is infinite.

If $\lambda_* < \sigma_2^2$, then L is invertible and $\vec{\mathbf{z}}$ solves system (2.3), $L\vec{\mathbf{z}} = \vec{\mathbf{r}} := G^{\mathrm{T}}\vec{\mathbf{p}}$. If also $(L + \Delta L)(\vec{\mathbf{z}} + \Delta\vec{\mathbf{z}}) = \vec{\mathbf{r}} + \Delta\vec{\mathbf{r}}$, then [1, pp. 606–607, Theorem A.13′]

$$\frac{\|\Delta\vec{\mathbf{z}}\|}{\|\vec{\mathbf{z}}\|} \leqslant \frac{\kappa(L)}{1 - \kappa(L)\|\Delta L\|/\|L\|}\left(\frac{\|\Delta L\|}{\|L\|} + \frac{\|\Delta\vec{\mathbf{r}}\|}{\|\vec{\mathbf{r}}\|}\right), \tag{7.4}$$

where $L = G^{\mathrm{T}}G - \lambda_* I$ and $G = P^{\perp}\check{M}_{4-5} = (\check{R}_{2,2})_{1-2}$, so that

$$\kappa_{2,2}(L) = \frac{\sigma_1^2(P^{\perp}\check{M}_{4-5}) - \lambda_*}{\sigma_2^2(P^{\perp}\check{M}_{4-5}) - \lambda_*}. \tag{7.5}$$

Yet $\vec{\mathbf{r}}$ can be zero, which can cause further instabilities in the solution $\vec{\mathbf{z}}_*$.

Geometric interpretation for fitted parabolae. As in Theorem 2.3, the extreme sensitivities caused by a double root of the secular equation arise only if $0 = \sigma_2 y_2 = (U^{\mathrm{T}}P^{\perp}\check{M}_6)_2$, which is a vanishing condition for the moments, or a measure of symmetry, of the distributon of the data. For instance, if the second principal moments equal each other, and if all the third principal moments vanish, then the condition $0 = \sigma_2 y_2 = (U^{\mathrm{T}}P^{\perp}\check{M}_6)_2$ still requires the vanishing of the fourth moments.

Indeed, after a rotation of the coordinates to the right-singular vectors of $\check{M}_{2-3}$, which are also the principal inertial axes of the data, the mixed moments vanish [6, p. 278], so that $\sum_{i=1}^{2} \check{x}_i\check{y}_i = 0$, which also means that the columns of $\check{M}_{2-3}$ are mutually perpendicular, so that $\check{R}_{1,1}$ is diagonal.

If the distribution is also centrally symmetric about its mean, so that $\check{\mathbf{x}}_i$ and $-\check{\mathbf{x}}_i$ are both in the data, then all the odd-degree moments vanish, in particular, the dot products $\check{M}_2 \bullet \check{M}_5 = \sum_{i=1}^{2} \check{x}_i^2\check{y}_i = 0 = \sum_{i=1}^{2} \check{x}_i\check{y}_i^2 = \check{M}_3 \bullet \check{M}_5$, and similarly $\check{M}_2 \bullet \check{M}_6 = 0 = \check{M}_3 \bullet \check{M}_6$. Thus $\check{M}_5 \perp \check{M}_{1-3}$ and $\check{M}_6 \perp \check{M}_{1-3}$. If also the principal second moments are equal, then $\sum_{i=1}^{2} \check{x}_i^2 = \sum_{i=1}^{2} \check{y}_i^2$, and hence also $\check{M}_4 \perp \check{M}_{1-3}$. In such a case, $\check{M}_6 \perp \check{M}_{4-5}$. if and only if

$$\check{M}_4 \bullet \check{M}_6 = 0 = \sum_{i=1}^{2} \check{x}_i\check{y}_i(\check{x}_i^2 + \check{y}_i^2), \tag{7.6}$$

$$\check{M}_5 \bullet \check{M}_6 = 0 = \sum_{i=1}^{2} (\check{x}_i^2 - \check{y}_i^2)(\check{x}_i^2 + \check{y}_i^2) \tag{7.7}$$

$$= \sum_{i=1}^{2} (\check{x}_i^4 - \check{y}_i^4) \tag{7.8}$$

Thus the fourth principal moments must also equal each other.

After the computation of the quadratic coefficients $\vec{\mathbf{w}}$, the same linear system $\check{R}_{1,1}\vec{\mathbf{v}} = -\check{R}_{1,2}\vec{\mathbf{w}}$ as in Section 6 yields the affine coefficients $\vec{\mathbf{v}}$, which are thus again unstable if the data are nearly colinear.

8. Applications

8.1. Fitting a hyperbola

Bookstein [4], Gander et al. [12, p. 564], Pratt [21], Späth [24,25], and Van Loan [29, pp. 302–305] offer algorithms, which, in principle, can fit hyperbolae with axes that are not necessarily parallel to the coordinate axes. However, they do not provide any such example or application. The example provided here illustrates Algorithm 4.1 to fit hyperbolae.

The problem of fitting a conic to data arises in the study of sundials, for instance, in the measurement of due north, or in the determination of the latitude and orientation of ancient sundials, as explained by James Evans:

> Local noon is the time of day when the shadow of a vertical gnomon is shortest. [...] Local noon need not occur at twelve o'clock [because of the inclination of the ecliptic]. The direction in which the shortest shadow points is called north. [...] As it is difficult to tell exactly when the shadow is shortest, let us consider an alternative procedure. Sketch a smooth curve through the points of the shadow-plot. [...]—[8, pp. 27–28].

Evans does not specify the type of curve to be fitted, but the shadow of the tip of a gnomon traces a conic on a planar sundial [15, pp. 84–86]. On horizontal sundials located between the southern and northern artic circles, the shadow traces a branch of a hyperbola [31, p. 137], with asymptotes pointing toward the sunrise and sunset. Along each artic circle, the shadow can also trace a parabola on the summer soltice, and within either artic region, the shadow can also trace an ellipse during midnight suns.

Example 8.1. Fig. 3 reproduces the base of a sundial with shadows plotted by Evans in Seattle, about at latitude 47°50′ north [8, p. 27, Fig. 1.6]. If the shadow-plot still lies as it did during the plotting operations, then due north lies along the direction from the gnomon to the shortest shadow, which has not been plotted. Therefore, finding due north on this shadow-plot amounts to locating the vertex of a hyperbola passing through, or fitted to, the plotted points.

Evans' data do not include any coordinates, which again justifies the requirement that the fitting algorithm remain invariant under rotations, symmetries, and

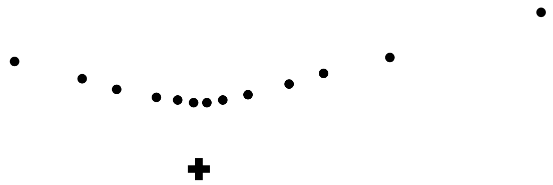

Fig. 3. Shadows (•) of the tip of a gnomon (+), from 8:23 AM (left) to 4:53 PM (right), adapted from Evans [8, p. 27, Fig. 1.6].

Table 1
Coordinates measured off [8, p. 27, Fig. 1.6], in [mm] (±0.25)

x_i	-34.75	-22	-15.5	-8.0	-4	-1.0	1.5	4.5	9.25	17	23.5	36	64.5
y_i	20.25	17	15.0	13.5	13	12.5	12.5	13.0	14.00	16	18.0	21	29.5

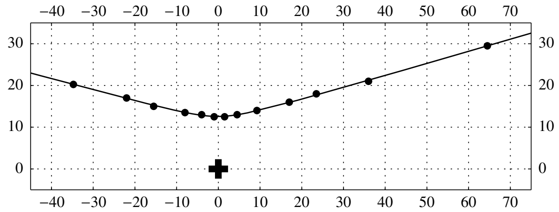

Fig. 4. Hyperbola fitted to shadows coordinates measured off a grid.

translations of coordinates. Thus Table 1 lists the coordinates of the shadows measured off a Cartesian grid superimposed on Evans's data. Fig. 4 shows this grid, and the best fitting conic section—a hyperbola—from Algorithm 4.1. To a few significant digits, the fitted hyperbola has its center at (0.126, 10.6), and the principal semi-axes have lengths 7.09 and 1.99; they point in the directions (0.999955, 0.00950672) and (−0.00950672, 0.999955), which indicates that the sundial was oriented nearly along the north–south and east–west directions.

For these data, the condition numbers are (rounded to a few digits)

$$\frac{1}{\kappa_{2,2}(\check{M}_{2-3})} = \frac{\sigma_2(\check{M}_{2-3})}{\sigma_1(\check{M}_{2-3})} = \frac{13.2809}{90.1900} = 0.147254, \tag{8.1}$$

$$\kappa_{2,2}(\check{M}_{2-3}) = \frac{\sigma_1(\check{M}_{2-3})}{\sigma_2(\check{M}_{2-3})} = \frac{90.1900}{13.2809} = 6.79097, \tag{8.2}$$

$$\tilde{\kappa}(\check{R}_{2,2}) = \frac{\sigma_1(\check{R}_{2,2})}{\sigma_2(\check{R}_{2,2}) - \sigma_3(\check{R}_{2,2})} = \frac{1087.26}{217.622 - 7.27432} = 5.17. \tag{8.3}$$

The reciprocal $1/\kappa_{2,2}(\check{M}_{2-3}) = 0.15$ indicates that $\check{M}_{2-3}$ is not numerically singular [14, p. 247], and hence that the data are not colinear. The condition numbers $\kappa_{2,2}(\check{M}_{2-3}) = 6.8$ and $\tilde{\kappa}(\check{R}_{2,2}) = 5.2$ indicate a moderate sensitivity of the fitted hyperbola to perturbations of the data.

8.2. Fitting a parabola

The next example compares Algorithm 4.1 for parabolae with other algorithms and test-data from the literature.

Table 2
Test data *on* a parabola, from Späth [23, p. 268]

x_i	-6.6	-2.8	-0.2	0.4	1.2	1.4
y_i	8.8	5.4	3.6	7.8	3.4	4.8

Table 3
Test data *near* a parabola, from Späth [23, p. 268]

x_i	-7	-3	0	0	1	1
y_i	9	5	4	8	3	5

Example 8.2. Späth reports several difficulties in fitting parabolae by parametric least squared orthogonal distances: at each of about 100 iterations, his algorithm requires solving and testing all three solutions of a separate cubic equation for each data point, different starting values can lead to different local minima, and convergence to a global minimum is not guaranteed [23].

In contrast, for each of Späth's data sets, Algorithm 4.1 produces only one parabola. For data *on* a parabola, in Table 2, Algorithm 4.1 reproduces that parabola, and for data *near* a parabola, in Table 3, the same algorithm produces a parabola which appears to fit the data closely, as shown in Fig. 5. (The publication [23, p. 268, Example 3] does not provide the angle of rotation of the parabola fitted there, which precludes its inclusion in Fig. 5.)

To the same data *near* a parabola, in Table 3, Bookstein's constraint $\|\vec{\mathbf{w}}\|_2 = 1$ alone—*without* the constraint $\det(A) = 0$—does not yield a parabola, but instead produces the *ellipse* shown in Fig. 6 as the best fitting conic. The conic corresponding to the next smaller singular value is the *hyperbola* shown in Fig. 6. Therefore, to insist on fitting a *parabola*, the constraint $\det(A) = 0$ has to be activated. In other words, the algorithms just presented (2.4, 2.5, 4.1) provide a means that was hitherto not available for fitting parabolae to data algebraically.

For the data *on* the parabola, in Table 2, the condition numbers are

$$\frac{1}{\kappa_{2,2}(\check{M}_{2-3})} = \frac{\sigma_2(\check{M}_{2-3})}{\sigma_1(\check{M}_{2-3})} = \frac{3.25264}{7.85835} = 0.413909, \tag{8.4}$$

$$\kappa_{2,2}(\check{M}_{2-3}) = \frac{\sigma_1(\check{M}_{2-3})}{\sigma_2(\check{M}_{2-3})} = \frac{7.85835}{3.25264} = 2.415991, \tag{8.5}$$

which suggest a moderate sensitivity of the fitted parabola to perturbations. For the polynomial secular equation, computations produce

$$\lambda_* = -0.102179 \cdot 10^{-5}, \quad |\psi(\lambda_*)| < 10^{-12}, \quad \psi'(\lambda_*) = -1382084.5, \tag{8.6}$$

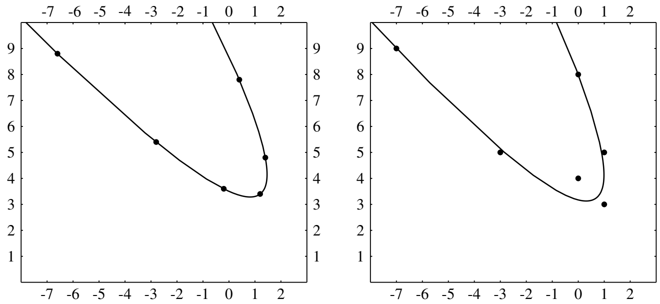

Fig. 5. Parabolae fitted by Algorithm 4.1 to Späth's data [23, p. 268].

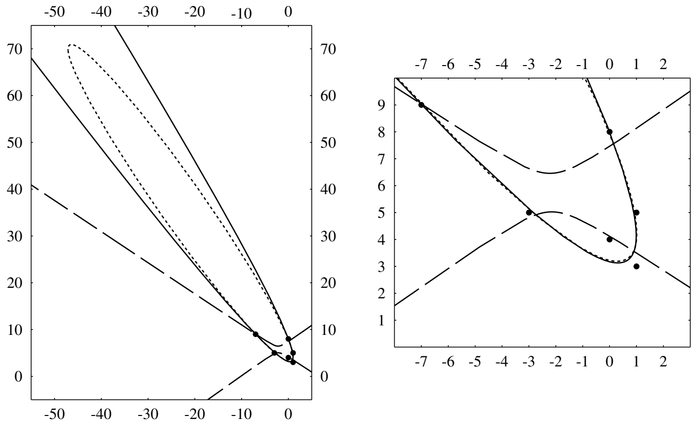

Fig. 6. Ellipse $(\cdots)$ and hyperbola $(--)$ fitted by Bookstein's method, and parabola $(-)$ fitted by Algorithm 4.1, to Späth's data [23, Example 3].

so that $\partial\lambda/\partial\psi_0 = -1/\psi'(\lambda_*) = 7.23545 \times 10^{-7}$ and, for the system (2.3)

$$\kappa_{2,2}(L) = \frac{\sigma_1^2 - \lambda_*}{\sigma_2^2 - \lambda_*} = \frac{16.7748^2 + 1.02179 \times 10^{-6}}{4.35304^2 + 1.02179 \times 10^{-6}} = 14.9, \tag{8.7}$$

also indicating a moderate sensitivity of the fitted parabola to perturbations.

For the data *near* the parabola, in Table 3, the parabola fitted by Algorithm 4.1 admits the parametrization

$$\vec{\mathbf{p}}(t) = \begin{pmatrix} -4/3 \\ 17/3 \end{pmatrix} + \begin{pmatrix} .826 & .564 \\ -.564 & .826 \end{pmatrix} \begin{pmatrix} t \\ [(t-.278)^2 - 3.31]/1.05 \end{pmatrix}. \tag{8.8}$$

The condition numbers are

$$\frac{1}{\kappa_{2,2}(\check{M}_{2-3})} = \frac{\sigma_2(\check{M}_{2-3})}{\sigma_1(\check{M}_{2-3})} = \frac{3.36528}{8.08341} = 0.4163198, \tag{8.9}$$

$$\kappa_{2,2}(\check{M}_{2-3}) = \frac{\sigma_1(\check{M}_{2-3})}{\sigma_2(\check{M}_{2-3})} = \frac{8.08341}{3.36528} = 2.402004. \tag{8.10}$$

For the polynomial secular equation, computations produce

$$\lambda_* = 2.77003, \quad |\psi(\lambda_*)| < 10^{-12}, \quad \psi'(\lambda_*) = -1906306.1, \tag{8.11}$$

so that $\partial\lambda/\partial\psi_0 = -1/\psi'(\lambda_*) = 5.24575 \times 10^{-5}$ and, for the system (2.3)

$$\kappa_{2,2}(L) = \frac{\sigma_1^2 - \lambda_*}{\sigma_2^2 - \lambda_*} = \frac{18.1856^2 - 2.77003}{4.72749^2 - 2.77003} = 16.75, \tag{8.12}$$

also indicating a moderate sensitivity of the fitted parabola to perturbations.

Acknowledgments

The computations were done with `FORTRAN` routines from `www.netlib.org`: `DGEQR2.f` for QR factorizations, `svd.f` for singular-value decompositions. I thank the referees for several suggestions that greatly improved and clarified the manuscript.

References

[1] O. Axelsson, Iterative Solution Methods, Cambridge University Press, Cambridge, UK, 1994.

[2] M. Berger, Geometry II, Springer-Verlag, Berlin, 1987.

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

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

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

[6] H. Cramér, Mathematical Methods of Statistics, Princeton Landmarks in Mathematics, Princeton University Press, Princeton, NJ, 1999.

[7] P. de Groen, An introduction to total least squares, Nieuw Arch. Wisk. (July) (1996) 237–253.

[8] J. Evans, The History & Practice of Ancient Astronomy, Oxford University Press, New York, NY, 1998.

[9] W.H. Fleming, Functions of Several Variables, second ed., Springer-Verlag, New York, NY, 1987.

[10] W. Fulton, Algebraic Curves, Benjamin/Cummings, Reading, MA, 1969.

[11] W. Gander, Least squares with a quadratic constraint, Numer. Math. 36 (3) (1981) 291–307.

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

[13] 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.

[14] G.H. Golub, C.F. Van Loan, Matrix Computations, second ed., Johns Hopkins University Press, Baltimore, MD, 1989.

[15] G.A. Jennings, Modern Geometry with Applications, Springer-Verlag, New York, 1994.

[16] T. Kato, Perturbation Theory for Linear Operators, second ed., Springer-Verlag, Berlin, 1984.

[17] S.P. Keeler, Y. Nievergelt, Computing geodetic coordinates, SIAM Rev. 40 (2) (1998) 300–309.

[18] 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.

[19] L. Mirsky, Symmetric gauge functions and unitarily invariant norms, Quart. J. Math., Oxford 11 (1960) 50–59.

[20] Y. Nievergelt, Hyperplanes and hyperspheres fitted seamlessly by algebraic constrained total least-squares, Linear Algebra Appl. 331 (1) (2001) 43–59.

[21] V. Pratt, Direct least-squares fitting of algebraic surfaces, ACM Comput. Graphics 21 (4) (1987) 145–152.

[22] E. Schmidt, Zur Theorie der linearen und nichtlinearen Integralgleichungen. 1. Teil: Entwicklung willkürlicher Funktionen nach Systemen vorgeschriebener, Math. Annalen 63 (1907) 433–476.

[23] H. Späth, Orthogonal least squares fitting with parabolas, in: G. Alefeld, J. Herzberger (Eds.), Numerical Methods and Error Bounds: Proceedings of the IMACS-GAMM International Symposium on Numerical Methods and Error Bounds held in Oldenburg, Germany, 9–12 July 1995, Akademie Verlag GmbH, Berlin, 1996, pp. 261–269.

[24] H. Späth, Least-squares fitting of ellipses and hyperbolas, Comput. Statist. 12 (3) (1997) 329–341.

[25] H. Späth, Orthogonal least squares fitting by conic sections, in: S. Van Huffel (Ed.), Recent Advances in Total Least Squares and Errors-In-Variables Models, Society for Industrial and Applied Mathematics, Philadelphia, PA, 1997, pp. 259–264.

[26] G.W. Stewart, Perturbation bounds for the QR factorization of a matrix, SIAM J. Numer. Anal. 14 (3) (1977) 509–518.

[27] G.W. Stewart, J.G. Sun, Matrix Perturbation Theory, Academic Press, London, UK, 1990.

[28] J. Stoer, R. Bulirsch, Introduction to Numerical Analysis, second ed., Springer-Verlag, New York, NY, 1993.

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

[30] J.M. Varah, Least squares data fitting with implicit functions, BIT 36 (1996) 842–854.

[31] A.E. Waugh, Sundials: Their Theory and Construction, Dover, New York, NY, 1973.

[32] P.-Å. Wedin, Perturbation bounds in connection with singular value decomposition, BIT 12 (1) (1972) 99–111.

[33] P.-Å. Wedin, Perturbation theory for pseudo-inverses, BIT 13 (2) (1973) 217–232.