International Scholarly Research Network
ISRN Probability and Statistics
Volume 2012, Article ID 895178, 25 pages
doi:10.5402/2012/895178

Research Article

Does the Best-Fitting Curve Always Exist?

N. Chernov, Q. Huang, and H. Ma

Department of Mathematics, University of Alabama at Birmingham, Birmingham, AL 35294, USA

Correspondence should be addressed to N. Chernov, chernov@uab.edu

Received 14 June 2012; Accepted 25 July 2012

Academic Editors: C. Fox, J. Hu, J. Jiang, J. López-Fidalgo, and S. Lototsky

Fitting geometric shapes to observed data points (images) is a popular task in computer vision and modern statistics (errors-in-variables regression). We investigate the problem of existence of the best fit using geometric and probabilistic approaches.

1. Introduction

In many areas of human practice, one needs to approximate a set of points $P_1, \ldots, P_n \in \mathbb{R}^d$ (representing experimental data or observations) by a simple geometric shape, such as a line, plane, circular arc, elliptic arc, and spherical surface. This problem is known as fitting a model object (line, circle, sphere, etc.) to observed data points.

The best fit is achieved when the geometric distances from the given points to the model object are minimized, in the least squares sense. Finding the best fit reduces to the minimization of the objective function

$$\mathcal{F}(S) = \sum_{i=1}^{n} [\operatorname{dist}(P_i, S)]^2, \tag{1.1}$$

where $S \subset \mathbb{R}^d$ denotes the model object (line, circle, sphere, etc.).

While other fitting criteria are used as well, it is the minimum of (1.1) that is universally recognized as the most desirable solution of the fitting problem. It has been adopted as a standard in industrial applications [1, 2]. In statistics, the minimum of (1.1) corresponds to the maximum likelihood estimate under usual assumptions known as functional model; see [3, 4]. Minimization of (1.1) is also called orthogonal distance regression (ODR).

Table 1

Objects	Existence in all cases	Existence in typical cases	Uniqueness
Lines	Yes	Yes	No
Circles	No	Yes	No
Ellipses	No	No	No
All conics	No	Yes	No

Geometric fitting in the sense of minimizing (1.1) has a long history. First publications on fitting linear objects (lines, planes, etc.) to given points in 2D and 3D date back to the late nineteenth century [5]. The problem of fitting lines and planes was solved analytically in the 1870s [6], and the statistical properties of the resulting estimates were studies throughout the twenties century [7–9], the most important discoveries being made in the 1970s [3, 10, 11]. These studies gave rise to a new branch of mathematical statistics now known as errors-in-variable (EIV) regression analysis [3, 4].

Since the 1950s, fitting circles and spheres to data points became popular in archaeology, industry, and so forth, [12, 13]. In the 1970s, researchers started fitting ellipses and hyperbolas to data points [14–17].

The interest to the problem of geometric fitting surged in the 1990s when it became an agenda issue for the rapidly growing computer science community. Indeed, fitting simple contours to digitized images is one of the basic tasks in pattern recognition and computer vision. In most cases, those contours are lines, circles, ellipses and other conic sections (called simply *conics*, for brevity) in 2D, and planes, spheres, ellipsoids in 3D. More complicated curves and surfaces are used occasionally too [1, 18], but it is more common to divide complex images into smaller segments and approximate each of those by a line or by a circular arc. This way one can approximate a complex contour by a polygonal line or a circular spline (see, e.g., [19, 20]).

Most publications on the geometric fitting problem are devoted to analytic solutions (which are only possible for lines and planes), or practical algorithms for finding the best-fitting object, or statistical properties of the resulting estimates. Very rarely one addresses fundamental issues such as the existence and uniqueness of the best fit. If these issues do come up, one either assumes that the best fit exists and is unique, or just points out examples to the contrary without deep investigation.

In this paper we address the issue of existence of the best fit in a general and comprehensive manner. The issue of uniqueness will be treated in a separate paper. These issues turn out to be quite nontrivial and lead to unexpected conclusions. As a glimpse of our results, we provide Table 1 summarizing the state of affairs in the problem of fitting 2D objects (here Yes means the best-fitting object exists or is unique in all respective cases; No means the existence/uniqueness fails in some of the respective cases).

We see that the existence and uniqueness of the best-fitting object cannot be just taken for granted. Actually 2/3 of the answers in Table 1 are negative. The uniqueness can never be guaranteed. The existence is guaranteed only for lines. In typical cases (i.e., for typical sets of data points, in probabilistic sense; see precise definition in Section 6), the best-fitting line and circle do exist, but the best-fitting ellipse does not.

The existence and uniqueness of the best fit are not only of theoretical interest but also practically relevant. For example, knowing under what conditions the problem does not have a solution might help us understand why the computer algorithm keeps diverging, or returns

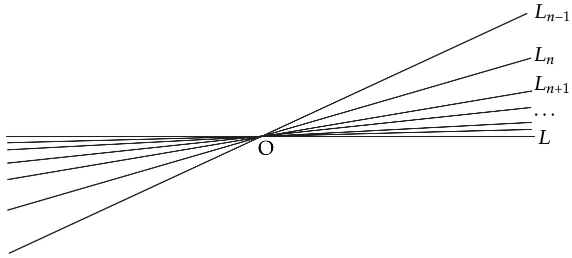


Figure 1: Sequence of lines L_n and the limit line L.

nonsense, or crashes altogether. While the cases where the best fit does not theoretically exist may be exceptional, nearby cases may be practically hard to handle, as the best-fitting object may be extremely difficult to find.

The nonuniqueness has its practical implications too. It means that the best-fitting object may not be stable under slight perturbations of the data points. An example is described by Nievergelt [21]; he presented a set of $n = 4$ points that can be fitted by three different circles equally well. Then by arbitrarily small changes in the coordinates of the points one can make any of these three circles fit the points a bit better than the other two circles, thus the best-fitting circle will change abruptly. See also a similar example in [4, Section 2.2].

Here we develop a general approach to the studies of existence of the best fit. Our motivation primarily comes from popular applications where one fits lines, circles, ellipses, and other conics, but our methods and ideas can be applied to much more general models. Our approach works in any dimension; that is, we can treat data points $P_1, \ldots, P_n \in \mathbb{R}^d$ and model objects $S \subset \mathbb{R}^d$ for any $d \geq 2$, but for the sake of notational simplicity and ease of illustrations we mostly restrict the exposition to the planar case $d = 2$.

2. W-Convergence and the Induced Topology

A crucial notion in our analysis is that of convergence (of model objects).

Motivating Example

Consider a sequence of lines $L_n = \{y = x/n\}$. They all pass through the origin $(0, 0)$ and their slopes $1/n$ decrease as n grows (Figure 1). Naturally, we would consider this sequence as convergent; it converges to the x-axis, that is, to the line $L = \{y = 0\}$, as $n \to \infty$.

However, it is not easy to define convergence so that the above sequence of lines would satisfy it. For example, we may try to use Hausdorff distance to measure how close two objects S_1 and $S_2 \subset \mathbb{R}^2$ are. It is defined by

$$\operatorname{dist}_{\mathrm{H}}(S_1, S_2) = \max\left\{ \sup_{P_1 \in S_1} \operatorname{dist}(P_1, S_2), \sup_{P_2 \in S_2} \operatorname{dist}(P_2, S_1) \right\}. \tag{2.1}$$

If the two sets S_1 and S_2 are closed and the Hausdorff distance between them is zero, that is, $\operatorname{dist}_{\mathrm{H}}(S_1, S_2) = 0$, then they coincide $S_1 = S_2$. If the Hausdorff distance is small, the two sets

S_1 and S_2 nearly coincide with each other. So the Hausdorff distance seems appropriate for our purposes.

But it turns out that the Hausdorff distance between the lines L_n and L in our example is *infinite*, that is, $\mathrm{dist}_{\mathrm{H}}(L_n, L) = \infty$ for every n. Thus the Hausdorff distance cannot be used to characterize convergence of lines as we would like to see it.

Windows

So why do we think that the line L_n converges to the line L, despite an infinite Hausdorff distance between them? It is because we do not really see an infinite line, we only "see" objects in a certain finite area, like in Figure 1. Suppose we see objects in some rectangle

$$R = \{-A \le x \le A, -B \le y \le B\}, \tag{2.2}$$

which for the moment will play the role of our "window" through which we look at the plane. Then we see segments of our lines within R; that is, we see intersections $L_n \cap R$ and $L \cap R$. Now clearly the segment $L_n \cap R$ gets closer to $L \cap R$, as n grows, and in the limit $n \to \infty$ they become identical. This is why we see the lines L_n converging to L. We see this convergence no matter how large the window R is. Note that the Hausdorff distance between $L_n \cap R$ and $L \cap R$ indeed converges to zero: $\mathrm{dist}_{\mathrm{H}}(L_n \cap R, L \cap R) \to 0$ as $n \to \infty$.

Window-Restricted Hausdorff Distance

Taking cue from the above example, we define the Hausdorff distance between sets S_1 and S_2 *within a finite window* R as follows:

$$\mathrm{dist}_{\mathrm{H}}(S_1, S_2; R) = \max\left\{ \sup_{P \in S_1 \cap R} \mathrm{dist}(P, S_2), \sup_{Q \in S_2 \cap R} \mathrm{dist}(Q, S_1) \right\}. \tag{2.3}$$

The formula (2.3) applies whenever both sets, S_1 and S_2, intersect the window R. If only one set, say S_1, intersects R, we modify (2.3) as follows:

$$\mathrm{dist}_{\mathrm{H}}(S_1, S_2; R) = \sup_{P \in S_1 \cap R} \mathrm{dist}(P, S_2). \tag{2.4}$$

A similar modification is used if only S_2 intersects R. If neither set intersects the window R, we simply set $\mathrm{dist}_{\mathrm{H}}(S_1, S_2; R) = 0$ (because we "see" two empty sets, which are not distinguishable).

We now define our notion of convergence for a sequence of sets.

Definition 2.1. Let $S_n \subset \mathbb{R}^2$ be some sets and $S \subset \mathbb{R}^2$ another set. We say that the sequence S_n *converges* to S if for any finite window R we have

$$\mathrm{dist}_{\mathrm{H}}(S_n, S; R) \longrightarrow 0 \quad \text{as } n \longrightarrow \infty. \tag{2.5}$$

In this definition the use of finite windows is crucial; hence we will call the resulting notion *Window convergence*, or *W convergence*, for short. According to this definition, the sequence of lines L_n in the above example indeed converges to the limit line L.

For finite-size (bounded) objects, like circles or ellipses, the W-convergence is equivalent to the convergence with respect to the Hausdorff distance. However for unbounded objects, such as lines and hyperbolas, we have to use our window-restricted Hausdorff distance and formula (2.5).

Our definition of convergence (i.e., W-convergence) is intuitively clear, but some complications may arise if it is used too widely. For example, if the limit object S is not necessarily closed, then the same sequence $\{S_n\}$ may have more than one limit S (though the closure $\overline{S}$ of every limit set S will be the same). Thus, from now on, to avoid pathological situations, we will assume that all our sets $S \subset \mathbb{R}^2$ are closed. All the standard model objects—lines, circles, ellipses, hyperbolas—are closed.

Convergence induces topology in the collection of all closed sets $S \subset \mathbb{R}^2$, which we denote by $\mathbb{X}$. A set $Y \subset \mathbb{X}$ is *closed* if for any sequence of sets $S_n \in Y$ converging to a limit set S the limit set also belongs to Y, that is, $S \in Y$. A set $U \subset \mathbb{X}$ is *open* if its complement $\mathbb{X} \setminus U$ is closed.

The above topology on $\mathbb{X}$ is metrizable. This means we can define a metric on $\mathbb{X}$, that is, a distance between closed sets in $\mathbb{R}^2$, in such a way that the W-convergence $S_n \to S$ means exactly that the distance between S_n and S goes down to zero. Such a distance can be defined in many ways; here is one of them.

Definition 2.2. The *W distance* (or *Window distance*) between two closed sets S_1 and $S_2 \subset \mathbb{R}^2$ is defined as follows:

$$\mathrm{dist}_{\mathrm{W}}(S_1, S_2) = \sum_{k=1}^{\infty} 2^{-k} \mathrm{dist}_{\mathrm{H}}(S_1, S_2; R_k), \tag{2.6}$$

where R_k is a square window of size $2k \times 2k$, that is, $R_k = \{|x| \le k, |y| \le k\}$.

In this formula, we use a growing sequence of nested windows and the Hausdorff distances between S_1 and S_2 within those windows balanced by the factors 2^{-k}. In the formula (2.6), the first nonzero term corresponds to the smallest window R_k that intersects at least one of the two sets, S_1 or S_2. The sum in (2.6) is always finite. Indeed, let us suppose, for simplicity, that both S_1 and S_2 intersect each window R_k. Then (since the distance between any two points in R_k is at most $2\sqrt{2}k$) the above sum is bounded by $2\sqrt{2}\sum_{k=1}^{\infty} k2^{-k} < 6$.

3. Continuity of Objective Function

Recall finding the best-fitting object S for a given set of points $P_1, \ldots, P_n \in \mathbb{R}^2$ consists of minimization of the objective function (1.1).

Let $\mathbb{M}$ denote the collection of model objects. We put no restrictions on $\mathbb{M}$ other than all $S \in \mathbb{M}$ are assumed to be closed sets. The collection $\mathbb{M}$ is then a subset of the space $\mathbb{X}$ of all closed subsets $S \subset \mathbb{R}^2$. The topology and metric on $\mathbb{X}$ defined above now induce a topology and metric on $\mathbb{M}$.

Redundancy Principle

If an object $S' \in \mathbb{M}$ is a subset of another object $S \in \mathbb{M}$, that is, $S' \subset S$, then for any point P we have $\operatorname{dist}(P, S) \leq \operatorname{dist}(P, S')$; thus S' cannot fit any set of data points better than S does. So for the purpose of minimizing $\mathcal{F}$, that is, finding the best-fitting object, we may ignore all model objects that are proper subsets of other model objects (they are *redundant*). This may reduce the collection $\mathbb{M}$ somewhat. Such a reduction is not necessary, it is just a matter of convenience, and we will occasionally apply it below.

Conversely, there is no harm in considering any subset $S' \subset S$ of an object $S \in \mathbb{M}$ as a (smaller) object, too. Indeed, if S' provides a best fit (i.e., minimizes the objective function $\mathcal{F}$), then so does S, because $\mathcal{F}(S) \leq \mathcal{F}(S')$. Hence including S' into the collection $\mathbb{M}$ will not really be an extension of $\mathbb{M}$, its inclusion will not change the best fit.

Theorem 3.1 (continuity of $\mathcal{F}$). *For any given points $P_1, \dots, P_n$ and any collection $\mathbb{M}$ of model objects the function $\mathcal{F}$ defined by* (1.1) *that is continuous on $\mathbb{M}$. This means that if a sequence of objects $S_m \in \mathbb{M}$ converges (i.e., W converges) to another object $S \in \mathbb{M}$, then $\mathcal{F}(S_m) \to \mathcal{F}(S)$.*

Proof. Since $\mathcal{F}(S)$ is the sum of squares of distances $\operatorname{dist}(P_i, S)$ to individual points P_i, see (1.1), it is enough to verify that the distance $\operatorname{dist}(P, S)$ is a continuous function of S for any given point P.

Suppose $P \in \mathbb{R}^2$ and a sequence of closed sets S_m W converging to a closed set S. We denote by $Q \in S$ the point in S closest to P, that is, such that $\operatorname{dist}(P, Q) = \operatorname{dist}(P, S)$; such a point Q exists because S is closed. Denote by D the disk centered on P of radius $1 + \operatorname{dist}(P, Q)$; it contains Q. Let R be a window containing the disk D.

Since R contains Q, it intersects with S, that is, $R \cap S \neq \emptyset$. This guarantees that $\operatorname{dist}_{\mathrm{H}}(S_m, S; R) \to 0$. Thus, there are points $Q_m \in S_m$ such that $Q_m \to Q$. Since $\operatorname{dist}(P, S_m) \leq \operatorname{dist}(P, Q_m)$, we conclude that the upper limit of the sequence $\operatorname{dist}(P, S_m)$ does not exceed $\operatorname{dist}(P, S)$, that is,

$$\limsup \operatorname{dist}(P, S_m) \leq \operatorname{dist}(P, S). \tag{3.1}$$

On the other hand, we will show that the lower limit of the sequence $\operatorname{dist}(P, S_m)$ cannot be smaller than $\operatorname{dist}(P, S)$, that is,

$$\liminf \operatorname{dist}(P, S_m) \geq \operatorname{dist}(P, S). \tag{3.2}$$

The estimates (3.1) and (3.2) together imply that $\operatorname{dist}(P, S_m) \to \operatorname{dist}(P, S)$, as desired, and hence the distance function $\operatorname{dist}(P, S)$ will be continuous on $\mathbb{M}$. It remains to prove (3.2).

To prove (3.2) we assume that it is false. Then there is a subsequence S_{m_k} in our sequence of sets S_m such that

$$\lim_{k \to \infty} \operatorname{dist}(P, S_{m_k}) = \liminf \operatorname{dist}(P, S_m) < \operatorname{dist}(P, S). \tag{3.3}$$

Denote by $Q_m \in S_m$ the point in S_m closest to P, that is, such that $\operatorname{dist}(P, Q_m) = \operatorname{dist}(P, S_m)$. Then we have

$$\lim_{k \to \infty} \operatorname{dist}(P, Q_{m_k}) = \lim_{k \to \infty} \operatorname{dist}(P, S_{m_k}) < \operatorname{dist}(P, S) = \operatorname{dist}(P, Q). \tag{3.4}$$

Since the points Q_{m_k} are closer to P than the point Q is, we have $Q_{m_k} \in D \subset R$. Recall that $\mathrm{dist}_{\mathrm{H}}(S_m, S; R) \to 0$, hence

$$\mathrm{dist}(Q_{m_k}, S) \longrightarrow 0 \quad \text{as } k \longrightarrow \infty. \tag{3.5}$$

Denote by $H_{m_k} \in S$ the point in S closest to Q_{m_k}, that is, such that $\mathrm{dist}(Q_{m_k}, H_{m_k}) = \mathrm{dist}(Q_{m_k}, S)$. Now we have, by triangle inequality,

$$\begin{aligned}\mathrm{dist}(P, S) \leq \mathrm{dist}(P, H_{m_k}) &\leq \mathrm{dist}(P, Q_{m_k}) + \mathrm{dist}(Q_{m_k}, H_{m_k}) \\ &= \mathrm{dist}(P, Q_{m_k}) + \mathrm{dist}(Q_{m_k}, S).\end{aligned} \tag{3.6}$$

The limit of the first term on the right hand side of (3.6) is $< \mathrm{dist}(P, S)$ by (3.4), and the limit of the second term is zero by (3.5). This implies $\mathrm{dist}(P, S) < \mathrm{dist}(P, S)$, which is absurd. The contradiction proves (3.2). And the proof of (3.2) completes the proof of the theorem. □

4. Existence of the Best Fit

The best-fitting object $S_{\mathrm{best}} \in \mathbb{M}$ corresponds to the (global) minimum of the objective function $\mathcal{F}$, that is, $S_{\mathrm{best}} = \arg\min_{S\in\mathbb{M}} \mathcal{F}(S)$. The function $\mathcal{F}$ defined by (1.1) cannot be negative; thus, it always has an infimum

$$\mathcal{F}_0 = \inf_{S\in\mathbb{M}} \mathcal{F}(S). \tag{4.1}$$

This means that one cannot find objects $S \in \mathbb{M}$ such that $\mathcal{F}(S) < \mathcal{F}_0$, but one can find objects $S \in \mathbb{M}$ such that $\mathcal{F}(S)$ is arbitrarily close to $\mathcal{F}_0$. More precisely, there is a sequence of objects S_m such that $\mathcal{F}(S_m) > \mathcal{F}_0$ for each m and $\mathcal{F}(S_m) \to \mathcal{F}_0$ as $m \to \infty$.

In practical terms, one usually runs a computer algorithm that executes a certain iterative procedure. It produces a sequence of objects S_m (here m denotes the iteration number) such that $\mathcal{F}(S_m) < \mathcal{F}(S_{m-1})$; that is, the quality of approximations improves at every step. If the procedure is successful, the value $\mathcal{F}(S_m)$ converges to the minimal possible value, $\mathcal{F}_0$, and the sequence of objects S_m converges (i.e., W-converges) to some limit object S_0. Then the continuity of the objective function $\mathcal{F}$ (proven earlier) guarantees that $\mathcal{F}(S_0) = \mathcal{F}_0$; that is, S_0 indeed provides the global minimum of the objective function, so it is the best-fitting object $S_0 = S_{\mathrm{best}}$.

A problem arises if the limit object S_0 does *not* belong to the given collection $\mathbb{M}$; hence it is not admissible. Then we end up with a sequence of objects S_m, each of which fits (approximates) the given points better than the previous one, but not as good as the next one. None of them would qualify as the best fit, and in fact the best fit would not exist, so the fitting problem would have no solution.

Illustrative Example

Suppose that our model objects are circles, and our given points are $P_1 = (-1, 0)$, $P_2 = (0, 0)$ and $P_3 = (1, 0)$. Then the sequence of circles S_m defined by $x^2 + (y - m)^2 = m^2$ will fit the points

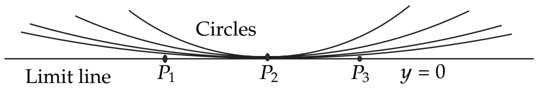


Figure 2: Sequence of circles and the limit line.

progressively better (tighter) as m grows, so that $\mathcal{F}(S_m) \to 0$ as $m \to \infty$ (Figure 2). On the other hand, no circle can pass through three collinear points; hence no circle S satisfies $\mathcal{F}(S) = 0$. Thus the circle-fitting problem has no solution.

The above sequence of circles S_m converges (in our terms, W-converges) to the x axis, which is a line, so it is natural to declare that line "the best fit." This may not be satisfactory in some practical applications, where one really needs to produce an estimate of the circle's center and radius. But if we want to present the best-fitting object here, it is clearly and undeniably the line $y = 0$.

Thus, in order to guarantee the existence of the best-fitting object in all cases, we need to include in our collection $\mathbb{M}$ all objects that can be obtained as limits of sequences of objects from $\mathbb{M}$. Such "limit objects" are called *limit points* of $\mathbb{M}$, in the language of topology.

Definition 4.1. A collection $\mathbb{M}$ which already contains all its "limit points" is said to be *closed*. If the collection $\mathbb{M}$ is not closed, then an extended collection $\overline{\mathbb{M}}$ that includes all the limit points of $\mathbb{M}$ is called the closure of $\mathbb{M}$.

For example, the collection $\mathbb{M}_L$ of all lines in $\mathbb{R}^2$ is closed, as a sequence of lines can only converge to a line. The collection $\mathbb{M}_C$ of all circles in $\mathbb{R}^2$ is *not* closed, as a sequence of circles may converge to a line. The closure of the collection of circles $\mathbb{M}_C$ includes all circles and all lines, that is,

$$\overline{\mathbb{M}}_C = \mathbb{M}_C \cup \mathbb{M}_L. \tag{4.2}$$

(Strictly speaking, a sequence of circles may also converge to a single point, see Section 5, so singletons need to be included too.)

We see that the collection $\mathbb{M}$ of model objects must be *closed* if we want the best-fitting object to exist in all cases. If $\mathbb{M}$ is not closed, we have to extend it by adding all its limit points and thus make it closed.

Theorem 4.2 (existence of the best fit). *Suppose that the given collection $\mathbb{M}$ of model objects is closed. Then for any given points $P_1, \ldots, P_n$ there exists the best-fitting object $S_{best} \in \mathbb{M}$, that is, the objective function $\mathcal{F}$ attains its global minimum on $\mathbb{M}$.*

Proof. The key ingredients of our proof will be the *continuity* of the objective function $\mathcal{F}$ and the *compactness* of a restricted domain of that function. We recall that a continuous real-valued function on a compact set always takes its maximum value and its minimum value on that set. In metric spaces (like our $\mathbb{M}$), a subset $\mathbb{M}_0 \subset \mathbb{M}$ is compact if every sequence of its elements $S_m \in \mathbb{M}_0$ has a subsequence S_{m_k} that converges to another element $S \in \mathbb{M}_0$, that is, $S_{m_k} \to S$ as $k \to \infty$.

We recall that our objective function $\mathcal{F}$ is continuous, and its domain $\mathbb{M}$ is now assumed to be closed. If it *was* compact, the above general fact would guarantee that $\mathcal{F}$ had a global minimum, as desired. But in most practical settings $\mathbb{M}$ is *not* compact. Indeed, if it *was* compact, the function $\mathcal{F}$ would not only have a minimum, but also a maximum. And this is usually impossible; one can usually find model objects arbitrarily far from the given points, thus making the distances from the points to the object arbitrarily large. Thus we need to find a smaller (restricted) subcollection $\mathbb{M}_0 \subset \mathbb{M}$ which will be compact, and then we will apply the above general fact.

Let $D_r = \{x^2 + y^2 \le r^2\}$ denote the disk of radius r centered on the origin $(0, 0)$. We can find a large enough disk D_r that satisfies two conditions: (i) D_r covers all the given points $P_1, \dots, P_n$ and (ii) D_r intersects at least one object $S_0 \in \mathbb{M}$, that is, $D \cap S_0 \neq \emptyset$. The distances from the given points to S_0 cannot exceed the diameter of D_r, which is $2r$, hence $\mathcal{F}(S_0) \le (2r)^2 n$.

Now we define our subcollection $\mathbb{M}_0 \subset \mathbb{M}$; it consists of all model objects $S \in \mathbb{M}$ that intersect the larger disk D_{3r} of radius $3r$. Objects that lie entirely outside D_{3r} are not included in $\mathbb{M}_0$. Note that the subcollection $\mathbb{M}_0$ contains at least one object (S_0 mentioned above); hence $\mathbb{M}_0$ is not empty.

Recall that all our given points $P_1, \dots, P_n$ lie in D_r. They are separated from the region outside the larger disk D_{3r} by a "no man's land"—the ring $D_{3r} \setminus D_r$, which is $2r$ wide. Thus the distances from the given points to any object S which was *not* included in $\mathbb{M}_0$ are greater than $2r$; hence for such objects we have $\mathcal{F}(S) > (2r)^2 n$. So objects not included in $\mathbb{M}_0$ cannot fit our points better than S_0 does. Thus they can be ignored in the process of minimization of $\mathcal{F}$. More precisely, if we find the best-fitting object S_{best} within the subcollection $\mathbb{M}_0$, then for any other object $S \in \mathbb{M} \setminus \mathbb{M}_0$ we will have

$$\mathcal{F}(S) > (2r)^2 n \ge \mathcal{F}(S_0) \ge \mathcal{F}(S_{\text{best}}), \tag{4.3}$$

which shows that S_{best} will be also the best-fitting object within the entire collection $\mathbb{M}$.

It remains to verify that the subcollection $\mathbb{M}_0$ is compact; that is, every sequence of objects $S_m \in \mathbb{M}_0$ has a subsequence converging to another object $S^* \in \mathbb{M}_0$. We will use the following standard fact. In a compact metric space any sequence of compact subsets has a convergent subsequence with respect to the Hausdorff metric.

Now let $j = [3r] + 1$ be the smallest integer greater than $3r$. Recall that all the objects in $\mathbb{M}_0$ are required to intersect D_{3r}; thus, they all intersect D_j as well. The sets $S_m \cap D_j$ are compact. By the above general fact, there is a subsequence $S_k^{(j)}$ in the sequence S_m and a compact subset $S_j^* \subset D_j$ such that $\text{dist}_{\text{H}}(S_k^{(j)} \cap D_j, S_j^*) \to 0$ as $k \to \infty$. Next, from the subsequence $S_k^{(j)}$ we extract a subsubsequence, call it $S_k^{(j+1)}$ that converges in the larger disk D_{j+1}; that is, such that $\text{dist}_{\text{H}}(S_k^{(j+1)} \cap D_{j+1}, S_{j+1}^*) \to 0$ as $k \to \infty$ for some compact subset $S_{j+1}^* \subset D_{j+1}$. Since $\text{dist}_{\text{H}}(S_k^{(j+1)} \cap D_j, S_j^*) \to 0$, we see that $S_{j+1}^* \cap D_j = S_j^*$; that is, the limit sets S_j^* and S_{j+1}^* "agree" within D_j.

Then we continue this procedure inductively for the progressively larger disks $D_{j+2}, D_{j+3}, \dots$. In the end we use standard Cantor's diagonal argument to construct a single subsequence S_{m_k} such that for every $i \ge j$ we have $\text{dist}_{\text{H}}(S_{m_k} \cap D_i, S_i^*) \to 0$ as $k \to \infty$, and the limiting subsets $S_i^* \subset D_i$ "agree" in the sense $S_{i+1}^* \cap D_i = S_i^*$ for every $i \ge j$. Then it follows that the sequence of objects S_{m_k} converges (i.e., W converges) to the closed set $S^* = \cup_{i \ge j} S_i^*$. The limit set S^* must belong to our collection $\mathbb{M}$ because that collection was assumed to be closed. Lastly, S^* intersects the disk D_{3r}, so it also belongs to the subcollection $\mathbb{M}_0$. Our proof is now complete. $\square$

5. Some Popular Models

Lines and Planes

The most basic (and oldest) fitting application is fitting lines in 2D. The model collection $\mathbb{M}_L$ consists of all lines $L \subset \mathbb{R}^2$. It is easy to see that a sequence of lines L_m can only converge to another line; hence, the collection $\mathbb{M}_L$ of lines is *closed*. This guarantees that the fitting problem always has a solution.

Similarly, the collection of lines in 3D is closed. The collection of planes in 3D is closed, too. Thus the problems of fitting lines and planes always have a solution. The same is true for the problem of fitting k-dimensional planes in a d-dimensional space $\mathbb{R}^d$ for any $1 \le k < d$.

Circles and Spheres

A less-trivial task is fitting circles in 2D. Now the model collection $\mathbb{M}_C$ consists of all circles $C \subset \mathbb{R}^2$. A sequence of circles C_m can converge to an object of three possible types: a circle $C_0 \subset \mathbb{R}^2$, a line $L_0 \subset \mathbb{R}^2$, or a single point (singleton) $P_0 \in \mathbb{R}^2$. Singletons can be regarded as degenerate circles (whose radius is zero), or they can be ignored based on the *redundancy principle* introduced in Section 3. But the lines are objects of a different type; thus, the collection $\mathbb{M}_C$ of circles is *not closed*.

As a result, the circle fitting problem does *not* always have a solution; a simple example was given in Section 4. In order to guarantee the existence of the best-fitting object, one needs to extend the given collection $\mathbb{M}_C$ by adding all its "limit points," in this case-lines. The extended collection will include all circles and all lines, that is,

$$\overline{\mathbb{M}}_C = \mathbb{M}_C \cup \mathbb{M}_L. \tag{5.1}$$

Now the (extended) circle fitting problem always has a solution; for any data points $P_1, \ldots, P_n$ there is a best-fitting object $S_{\text{best}} \in \overline{\mathbb{M}}_C$ that minimizes the sum of squares of the distances to $P_1, \ldots, P_n$. One has to keep in mind, though, that the best-fitting object may be a line, rather than a circle.

The problem of fitting circles to data points has been around since the 1950's [13]. Occasional nonexistence of the best-fitting circle has been first noticed empirically; see, for example, Berman and Culpin [22]. The first theoretical analysis of the non-existence phenomenon was done by Nievergelt [21] who traced it to the noncompactness of the underlying model space and concluded that lines needed to be included in the model space to guarantee the existence of the best fit. He actually studied this phenomenon in many dimensions, so he dealt with hyperspheres and hyperplanes. He called hyperplanes "generalized hyperspheres." Similar analysis and conclusions, for circles and lines, were published later, independently, by Chernov and Lesort [23] and by Zelniker and Clarkson [24]; see also a recent book [4].

Ellipses

A popular task in computer vision is fitting ellipses. Now the model collection $\mathbb{M}_E$ consists of all ellipses $E \subset \mathbb{R}^2$. A sequence of ellipses E_m may converge to an object of many possible types: an ellipse, a parabola, a line, a pair of parallel lines, a segment of a line, a half-line (a ray); see Figure 3.

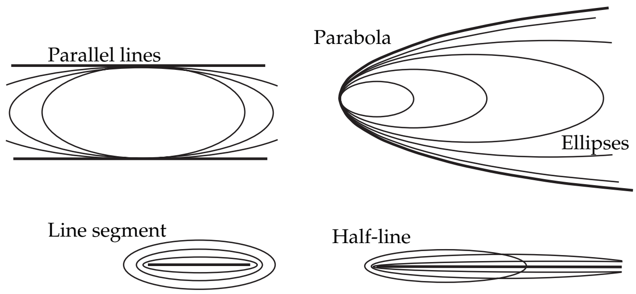


Figure 3: Sequences of ellipses converging to various limit objects.

Line segments can be regarded as degenerate ellipses whose minor axis is zero. Alternatively, line segments and half-lines can be ignored based on the redundancy principle (Section 3). Thus we have lines, pairs of parallel lines, and parabolas as limit objects of different types.

We see that the collection $\mathbb{M}_E$ of ellipses is *not closed*. As a result, the ellipse-fitting problem does *not* always have a solution. For example, this happens when given points that are placed on a line or on a parabola. (Actually, there are many more examples; see Section 6.)

In order to guarantee the existence of the best-fitting object, one needs to extend the collection $\mathbb{M}_E$ of ellipses by adding all its "limit points," in this case—*lines*, *pairs of parallel lines*, and *parabolas*. We denote that extended collection by $\overline{\mathbb{M}}_E$ that

$$\overline{\mathbb{M}}_E = \mathbb{M}_E \cup \mathbb{M}_L \cup \mathbb{M}_{\|} \cup \mathbb{M}_{\cup}, \tag{5.2}$$

where $\mathbb{M}_{\|}$ denotes the collection of pairs of parallel lines, and $\mathbb{M}_{\cup}$ the collection of parabolas (the symbol $\cup$ resembles a parabola).

Now the (extended) ellipse-fitting problem always has a solution, for any data points $P_1, \ldots, P_n$ there will be a best-fitting object $S_{\text{best}} \in \overline{\mathbb{M}}_E$ that minimizes the sum of squares of the distances to $P_1, \ldots, P_n$. One has to keep in mind, though, that the best-fitting object may be a line, or a pair of parallel lines, or a parabola, rather than an ellipse.

The ellipse fitting problem has been around since the 1970s. The need to deal with limiting cases was first notice by Bookstein [16] who wrote "The fitting of a parabola is a limiting case, exactly transitional between ellipse and hyperbola. As the center of ellipse moves off toward infinity while its major axis and the curvature of one end are held constant...." A first theoretical analysis on the non-existence of the best ellipse was done by Nievergelt in [25] who traced it to the noncompactness of the underlying model space and concluded that parabolas needed to be included in the model space to guarantee the existence of the best fit.

Conics

We now turn to a more general task of fitting quadratic curves (also known as conic sections, or simply *conics*). Now the model collection $\mathbb{M}_Q$ consists of all nondegenerate quadratic curves, by which we mean ellipses, parabolas, and hyperbolas. A sequence of conics may converge to an object of many possible types: a conic, a line, a line segment, a ray, two opposite rays, and a pair of parallel lines, and a pair of intersecting lines; see Figures 3 and 4.

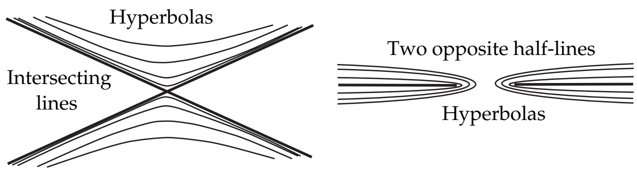


Figure 4: Sequences of hyperbolas converging to various limit objects.

Every line segment, ray, and two opposite rays are a part (subset) of a full line, so these objects can be ignored based on the redundancy principle (Section 3). But pairs of parallel lines and pairs of intersecting lines constitute limit objects of new types. We see that the collection $\mathbb{M}_Q$ of quadratic curves (conics) is *not closed*. As a result, the conic fitting problem does *not* always have a solution.

In order to guarantee the existence of the best-fitting object, one needs to extend the collection $\mathbb{M}_Q$ of quadratic curves by adding all its "limit points"—*pairs of lines* (which may be intersecting, parallel, or coincident)

$$\overline{\mathbb{M}}_Q = \mathbb{M}_Q \cup \mathbb{M}_{LL}, \tag{5.3}$$

where $\mathbb{M}_{LL}$ denotes the collection of pairs of lines.

Now the (extended) conic fitting problem always has a solution; for any data points $P_1, \dots, P_n$ there is a best-fitting object $S_{\text{best}} \in \overline{\mathbb{M}}_Q$ that minimizes the sum of squares of the distances to $P_1, \dots, P_n$. One has to keep in mind, though, that the best-fitting object may be a pair of lines, rather than a genuine conic.

6. Sufficient and Deficient Models

We see that if the collection $\mathbb{M}$ of adopted model objects is not closed, then in order to guarantee the existence of the best fit we have to close it up, that is, add to $\mathbb{M}$ all objects that are obtained as limits of existing objects. The new, extended collection $\overline{\mathbb{M}}$ will be closed and will always provide the best fit. Here we discuss practical implications of extending $\mathbb{M}$.

Definition 6.1. The original objects $S \in \mathbb{M}$ are called *primary objects* and the objects that have been added, that is, $S \in \overline{\mathbb{M}} \setminus \mathbb{M}$ are called *secondary objects*.

The above extension of $\mathbb{M}$ by secondary objects, though necessary, may not be completely harmless. Suppose again that one needs to fit circles to observed points. Since the collection of circles $\mathbb{M}_C$ is not closed, one has to add lines to it. That is, one really has to operate with the extended collection $\overline{\mathbb{M}}_C = \mathbb{M}_C \cup \mathbb{M}_L$ in which circles are primary objects and lines are secondary objects. The best-fitting object to a given set of points will then belong to $\overline{\mathbb{M}}_C$; hence occasionally it will be a line, rather than a circle.

In practical applications, though, a line may not be totally acceptable as the best-fitting object. For example, if one needs to produce an estimate of the circle's radius and center, a line will be of little help; it has no radius or center. Thus, when a secondary object happens to be the best fit to the given data points, various unwanted complications may arise.

Next we investigate how frequently such unwanted events occur.

In most applications points are obtained (measured or observed) with some random noise. The noise has a probability distribution, which in statistical studies is usually assumed to be normal [3, 4]. For our purposes it is enough to just assume that given points $P_1, \dots, P_n$ have an absolutely continuous distribution:

Assumption 6.2. The points $P_1, \dots, P_n$ are obtained randomly with a probability distribution that has a joint probability density $\rho(x_1, y_1, \dots, x_n, y_n) > 0$.

Let us first see how frequently the best-fitting circle fails to exist under this assumption. We begin with the simplest case of $n = 3$ points. If they are not collinear, then they can be interpolated by a circle. If they are collinear (and distinct), there is no interpolating circle, so the best fit is achieved by a line. The three points (x_i, y_i), $1 \le i \le 3$ are collinear if and only if

$$(x_1 - x_2)(y_1 - y_3) = (x_1 - x_3)(y_1 - y_2). \tag{6.1}$$

This equation defines a hypersurface in $\mathbb{R}^6$. Under the above assumption, the probability of that hypersurface is zero; that is, the collinearity occurs with probability zero. In simple practical terms, it is "impossible," it "never happens." If we generate three data points by using a computer random number generator, we will practically never see collinear points. All our practical experience tells us; the best-fitting circle *always* exists. Perhaps for this reason the circle-fitting problem is usually studied without including lines in the collection of model objects.

In the case of $n > 3$ data points, the best-fitting circle fails to exist when the points are collinear, but there are other instances too. For example, let $n = 4$ points be at $(0, 1)$, $(0, -1)$, $(A, 0)$, and $(-B, 0)$ for some large $A, B \gg 1$. Then it is not hard to check, by direct inspection, that the best-fitting circle fails to exist, and the best fit is achieved by the line $y = 0$.

It is not easy to describe all sets of points for which the best-fitting circle fails to exist, but such a description exists—it is given in [4] (see Theorem 8 on page 68 there). We skip technical details, only state the final conclusion; for every $n > 3$, the best-fitting circle exists unless the $(2n)$-vector $(x_1, y_1, \dots, x_n, y_n)$ belongs to a certain algebraic submanifold in $\mathbb{R}^{2n}$. So, under our assumption, the best-fitting circle exists with probability one.

In other words, no matter how large or small the data set is, the best-fitting circle will exist with probability one, so practically we never have to resort to secondary objects (lines); that is, we never have to worry about existence. The model of circles is adequate and sufficient; it does not really require an extension.

Definition 6.3. We say that a collection $\mathbb{M}$ of model objects is *sufficient* (for fitting purposes) if the best-fitting object S_{best} exists with probability one under the above assumption. Otherwise the model collection will be called *deficient*.

We note that under our assumption an event occurs with probability zero if and only if it occurs on a subset $A \subset \mathbb{R}^{2n}$ of Lebesgue measure zero, that is, $\text{Leb}(A) = 0$. Therefore, our notion of sufficiency does not depend on the choice of the probability density ρ.

As we have just seen, the collection of circles is sufficient for fitting purposes. Next we examine the model collection $\mathbb{M}_E$ of ellipses $E \subset \mathbb{R}^2$. As we know, this collection is not closed; its closure must include all lines, pairs of parallel lines, and parabolas.

Ellipses

Let us see how frequently the best-fitting ellipse fails to exist. It is particularly easy to deal with $n = 5$ data points. For any set of distinct 5 points in a general linear position (which means that no three points are collinear), there exists a unique quadratic curve (conic) that passes through all of them (i.e., interpolates them). That conic may be an ellipse, or a parabola, or a hyperbola. If 5 points are not in generic linear position (i.e., at least three of them are collinear), then they can be interpolated by a degenerate conic (a pair of lines).

If the interpolating conic is an ellipse, then obviously that ellipse is the best fit (the objective function takes it absolute minimum value, zero). What if the interpolating conic is a parabola or a pair of parallel lines? Then it is a secondary object in the extended model (5.2), and we have an unwanted event; a secondary object provides the best fit. We remark, however, that this event occurs with probability zero, so it is not a real concern. But there are other ways in which unwanted events occur, see below.

Suppose that the interpolating conic is a hyperbola or a pair of intersecting lines. Now the situation is less clear, as such an interpolating conic would not belong to the extended collection $\overline{\mathbb{M}}_E$; see (5.2). Is it possible then that the best-fitting object $S_{\text{best}} \in \overline{\mathbb{M}}_E$ is an ellipse? The answer is "no", and our argument is based on the following theorem that is interesting itself:

Theorem 6.4 (no local minima for $n = 5$). *For any set of five points the objective function $\mathcal{F}$ has no local minima on conics. In other words, for any non interpolating conic S there exist other conics S', arbitrarily close to S, which fit our data points better than S does, that is, $\mathcal{F}(S') < \mathcal{F}(S)$.*

The proof is given in the appendix, and here we answer the previous question. Suppose $n = 5$ points are interpolated by a hyperbola or a pair of intersecting lines. If there *were* a best-fitting ellipse $E_{\text{best}} \in \overline{\mathbb{M}}_E$, then no other ellipse could provide a better fit; that is, for any other ellipse $E \in \mathbb{M}_E$ we would have $\mathcal{F}(E) \geq \mathcal{F}(E_{\text{best}})$, which contradicts the above theorem. Thus the best-fitting object $S_{\text{best}} \in \overline{\mathbb{M}}_E$ is not an ellipse but a secondary object (a parabola, or a line, or a pair of parallel lines), that is, an unwanted event occurs.

One can easily estimate the probability of the above unwanted event numerically. For a given probability distribution one can generate random samples of $n = 5$ points and for each sample find the interpolating conic (by elementary geometry). Every time that conic happens to be other than an ellipse, an unwanted event occurs; that is, the best fit from within the collection (5.2) is provided by a secondary object.

We have run the above experiment with a standard normal distribution, where each coordinate x_i and y_i (for $i = 1, \ldots, 5$) was generated independently according to a standard normal law $\mathcal{N}(0, 1)$. We found that the random points were interpolated by an ellipse with probability 22% and by a hyperbola with probability 78%. (Other conics, including parabolas, never turned up; in fact, it is not hard to prove that they occur with probability zero.) This is a striking result; hyperbolas actually dominate over ellipses! Thus the best-fitting ellipse *fails* to exist with a probability as high as 78%.

We note that this percentage will remain the same if we generate points (x_i, y_i) independently according to any other 2D normal distribution. Indeed, any 2D normal distribution can be transformed to a standard 2D normal distribution by a linear map of the plane $\mathbb{R}^2$. Now under linear transformations conics are transformed to conics, and their types are preserved (i.e., ellipses are transformed to ellipses, hyperbolas to hyperbolas, etc.). Therefore the type of the interpolating conic cannot change.

Table 2: Percentage (%) of randomly generated samples of n points for which the best-fitting ellipse exists.

Distribution	$n = 5$	$n = 6$	$n = 8$	$n = 10$	$n = 20$	$n = 50$
Normal $\mathbf{V}_1$	22	23	23	24	27	27
Normal $\mathbf{V}_2$	22	22	24	26	27	33
Square $[0,1] \times [0,1]$	28	32	40	46	52	60
Rectangle $[0,2] \times [0,1]$	28	29	28	30	31	32

In another experiment we sampled points from the unit square $[0,1] \times [0,1]$ with a uniform distribution; that is, we selected each coordinate x_i and y_i independently from the unit interval $[0,1]$. In this experiment ellipses turned up with probability 28% and hyperbolas with probability 72%, again an overwhelming domination of hyperbolas!

And again, these percentages will remain the same if we generate points (x_i, y_i) according to any other rectangular (uniform) distribution. Indeed, any rectangle can be transformed to the square $[0,1] \times [0,1]$ by a linear map of the plane, and then the same argument as before applies.

We also conducted numerical tests for $n > 5$ points. In these tests we generated points $P_1, \dots, P_n$, independently according to four predefined distributions: normal with covariance matrix $\mathbf{V}_1 = \left[\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right]$, normal with covariance matrix $\mathbf{V}_2 = \left[\begin{smallmatrix} 4 & 0 \\ 0 & 1 \end{smallmatrix}\right]$, uniform in the square $[0,1] \times [0,1]$, uniform in the rectangle $[0,2] \times [0,1]$. Table 2 shows the percentage of randomly generated samples of n points for which the best-fitting ellipse exists.

We see that in most cases the best-fitting ellipse fails to exist with a high probability. Only for the uniform distribution is the square $[0,1] \times [0,1]$ the failures decrease as $n \to \infty$, and apparently the probability of existence of the best-fitting ellipse grows to 100%. We will explain this in Section 7.

We note that for $n \geq 6$ points there is no interpolating conic, so one has to find a global minimum of $\mathcal{F}$ via an extensive search over the model space $\mathbb{M}_Q$ of conics. This is a very time-consuming procedure and its results are never totally reliable. Thus our estimates, especially for larger n's, are quite approximate, but we believe that they are accurate to within 5%.

Insufficiency of Ellipses

We conclude that the collection of ellipses is *not sufficient* for fitting purposes. This means that there is a real chance that for a given set of data points no ellipse could be selected as the best fit to the points; that is, the ellipse-fitting problem would have no solution. More precisely, for any ellipse E there will be another ellipse E' providing a better fit, in the sense $\mathcal{F}(E') < \mathcal{F}(E)$. If one constructs a sequence of ellipses that fit the given points progressively better and on which the objective function $\mathcal{F}$ converges to its infimum, then those ellipses will grow in size and converge to something different than an ellipse (most likely, to a parabola).

In practical terms, one usually runs a computer algorithm that executes an iterative procedure such as Gauss-Newton or Levenberg-Marquardt. It produces a sequence of ellipses E_m (here m denotes the iteration number) such that $\mathcal{F}(E_m) < \mathcal{F}(E_{m-1})$; that is, the quality of approximations improves at every step, but those ellipses would keep growing in size and approach a parabola. This situation commonly occurs in practice [16, 25, 26], and one either accepts a parabola as the best fit or choose an ellipse by an alternative procedure (e.g., by sirect ellipse fit [27]).

We see that whenever the best-fitting ellipse fails to exist, the ellipse-fitting procedure will attempt to move beyond the collection of ellipses, so it will end up on the border of that collection and then return a secondary object (a parabola or a pair of lines). In a sense, the scope of the collection of ellipses is too narrow for fitting purposes. In other words, this collection is deficient, or badly incomplete. Its deficiency indicates that it should be substantially extended for the fitting problem to have a reasonable (adequate) solution with probability one.

Conics

Such an extension is achieved by the collection of all conics (ellipses, hyperbolas, and parabolas) denoted by $\mathbb{M}_Q$; see Section 5. If one searches for the best-fitting object S_{best} in the entire collection of conics, then S_{best} will exist with probability one. This was confirmed in numerous computer experiments that we have conducted. In fact, the best-fitting object has always been either an ellipse or hyperbola, so parabolas apparently can be excluded. Thus our numerical tests strongly suggest that the collection $\mathbb{M}_{EH} = \mathbb{M}_E \cup \mathbb{M}_H$ of ellipses, and hyperbolas is sufficient (though we do not have a proof).

7. Generalizations

In the basic formula (1.1) all the n points make equal contributions to $\mathcal{F}$. In some applications, though, one needs to minimize a weighted sum

$$\mathcal{F}_w(S) = \sum_{i=1}^{n} w_i [\text{dist}(P_i, S)]^2, \tag{7.1}$$

for some fixed $w_i > 0$. This problem is known as *weighted least squares*. All our results on the existence of the best fit apply to it as well.

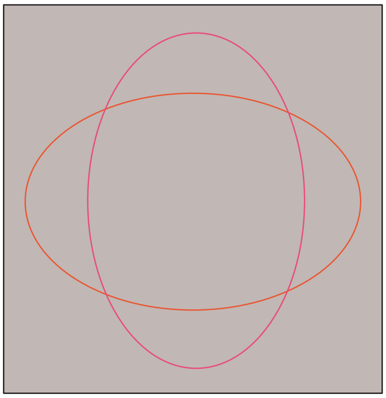
More generally, given a probability distribution $\mathrm{dP}(x, y)$ on $\mathbb{R}^2$, one can "fit" an object S to it (or approximate it) by minimizing

$$\mathcal{F}_P(S) = \iint d^2(x, y)\,\mathrm{dP}(x, y), \tag{7.2}$$

where $d(x, y) = \text{dist}[(x, y), S]$. Note that (7.1) is a particular version of (7.2) in which dP is a discrete measure with a finite support. In order to guarantee the finiteness of the integral in (7.2) one needs to assume that the distribution dP has finite second moments. After that all our results apply.

The generalization (7.2) is relevant to the numerical tests described in Section 6. Suppose again that n points are selected independently according to a probability distribution $\mathrm{dP}(x, y)$ on $\mathbb{R}^2$. As n grows, due to the law of large numbers of probability theory that the normalized objective function $(1/n)\mathcal{F}$ converges exactly to the integral (7.2).

Now the global minimum of the function $\mathcal{F}_P(S)$ would give us the best-fitting object corresponding to the given probability distribution dP, which would be the best "asymptotic" fit to a random sample of size n selected from the distribution dP, as $n \to \infty$. This relation helps us explain some experimental results reported in Section 6.



Fitting object	Value of $\mathcal{F}$
Best ellipse (either of the two)	0.02001
Best circle (of radius 0.38265)	0.02025
Two diagonals of the square	0.02083

Figure 5: The best fit to a uniform distribution in a square.

Suppose that dP is a uniform distribution in a rectangle $R = [0, L] \times [0, 1]$. Then the integral in (7.2) becomes

$$\mathcal{F}_P(S) = L^{-1} \int_0^L \int_0^1 d^2(x, y)\, dy\, dx. \tag{7.3}$$

We have computed it numerically in order to find the best-fitting conic S. The results are given below.

Perfect Square ($R = [0, 1] \times [0, 1]$)

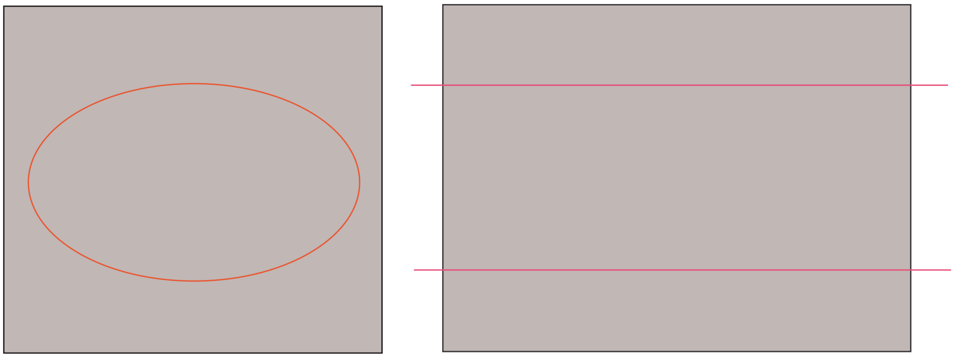
One would naturally expect that the best fit to a mass uniformly distributed in a perfect square would be either a circle or a pair of lines (say, the diagonals of the square).

What we found was totally unexpected; the best-fitting conic is an ellipse! Its center coincides with the center of the square, and its axes are aligned with the sides of the square, but its axes are different; the major axis is $2a = 0.9489$ and the minor axis is $2b = 0.6445$ (assuming that the square has unit side). In fact, there are exactly *two* such ellipses; one is oriented horizontally and the other—vertically; see the left panel of Figure 5. These two ellipses beat any circle and any pair of lines (see the table in Figure 5); they provide the global minimum of the objective function (7.3).

The fact that the best-fitting conic to a mass distributed in a square is an ellipse explains why for large samples generated from the uniform distribution in a square ellipses dominate over hyperbolas (Section 6).

As the rectangle R gets extended horizontally, that is, as $L > 1$ increases the best-fitting conic changes. We found that for $L < L_1 \approx 1.2$, the best-fitting conic is still an ellipse (though it gets elongated compared to the one found for the square, when $L = 1$). But for $L > L_1$, the best-fitting conic abruptly changes from an ellipse to a pair of parallel lines. Those two lines are running through the rectangle horizontally; their equations are $y = 0.25$ and $y = 0.75$ (See Figure 6).

This fact might explain why for longer rectangles, such as $R = [0, 2] \times [0, 1]$ used in Section 6, the percentage of samples for which the best-fitting ellipse exists does not grow



$1 < L < 1.2$ $L > 1.2$

Figure 6: The best fit to a uniform distribution in a rectangle.

much as n increases. Though it remains unclear which type of conics (ellipses or hyperbolas) dominates for large samples taken from longer rectangles $[0, L] \times [0, 1]$ with $L > L_1$.

8. Megaspace

Our conclusions can be illustrated by an interesting construction in a multidimensional space "megaspace". It was first used by Malinvaud [28, Chapter 10] and then by Chernov [4, Sections 1.5 and 3.4].

Recall that given n data points $P_1 = (x_1, y_1), \ldots, P_n = (x_n, y_n)$ and a model object S (a closed subset of $\mathbb{R}^2$), our objective function $\mathcal{F}(S)$ is defined by (1.1) and for the distances $\operatorname{dist}(P_i, S)$ we can write

$$[\operatorname{dist}(P_i, S)]^2 = \min_{(x_i', y_i') \in S} \left\{ \left(x_i - x_i'\right)^2 + \left(y_i - y_i'\right)^2 \right\}. \tag{8.1}$$

Thus we can express the objective function $\mathcal{F}(S)$ as follows:

$$\mathcal{F}(S) = \min \left\{ \sum_{i=1}^{n} \left[\left(x_i - x_i'\right)^2 + \left(y_i - y_i'\right)^2 \right];\ (x_i', y_i') \in S\ \forall i \right\}. \tag{8.2}$$

Now let us represent the n data points $P_1 = (x_1, y_1), \ldots, P_n = (x_n, y_n)$ by one point ("megapoint") $\mathcal{P}$ in the $2n$-dimensional space $\mathbb{R}^{2n}$ with coordinates $x_1, y_1, \ldots, x_n, y_n$. For the given object S, let us also define a multidimensional set "megaset" $\mathfrak{M}_S \subset \mathbb{R}^{2n}$ as follows:

$$\mathcal{P}' = \left(x_1', y_1', \ldots, x_n', y_n'\right) \in \mathfrak{M}_S \Longleftrightarrow \left(x_i', y_i'\right) \in S \quad \forall i. \tag{8.3}$$

Note that $\sum_{i=1}^{n} [(x_i - x_i')^2 + (y_i - y_i')^2]$ in (8.2) is the square of the distance from $\mathcal{P}$ to $\mathcal{P}' \in \mathfrak{M}_S$, in the megaspace $\mathbb{R}^{2n}$. Therefore consider

$$\mathcal{F}(S) = \min_{\mathcal{P}' \in \mathfrak{M}_S} \left[\operatorname{dist}(\mathcal{P}, \mathcal{P}')\right]^2 = [\operatorname{dist}(\mathcal{P}, \mathfrak{M}_S)]^2. \tag{8.4}$$

Next given a collection $\mathbb{M}$ of model objects we define a large megaset $\mathfrak{M}(\mathbb{M}) \subset \mathbb{R}^{2n}$ as follows: $\mathfrak{M}(\mathbb{M}) = \cup_{S\in\mathbb{M}}\mathfrak{M}_S$. Alternatively, it can be defined as

$$\left(x'_1, y'_1, \ldots, x'_n, y'_n\right) \in \mathfrak{M}(\mathbb{M}) \Longleftrightarrow \exists S \in \mathbb{M} : \left(x'_i, y'_i\right) \in S \quad \forall i. \tag{8.5}$$

The best-fitting object S_{best} minimizes the function $\mathcal{F}(S)$. Thus, due to (8.4), S_{best} minimizes the distance from the megapoint $\mathcal{P}$ representing the data set $P_1, \ldots, P_n$ to the megaset $\mathfrak{M}(\mathbb{M})$ representing the collection $\mathbb{M}$.

Thus, the problem of finding the best-fitting object S_{best} reduces to the problem of finding the megapoint $\mathcal{P}' \in \mathfrak{M}(\mathbb{M})$ that is closest to the given megapoint $\mathcal{P}$ (representing the given n points). In other words, we need to *project* the megapoint $\mathcal{P}$ onto the megaset $\mathfrak{M}(\mathbb{M})$, and the footpoint $\mathcal{P}'$ of the projection would correspond to the best-fitting object S_{best}.

We conclude that the best-fitting object $S_{\text{best}} \in \mathbb{M}$ exists if and only if there exists a megapoint $\mathcal{P}' \in \mathfrak{M}(\mathbb{M})$ that is closest to the given megapoint $\mathcal{P}$. It is a simple fact that given a set $D \subset \mathbb{R}^d$ with $d \geq 1$, the closest point $X' \in D$ to a given point $X \in \mathbb{R}^d$ exists for any $X \in \mathbb{R}^d$ if and only if the set D is closed. Thus the existence of the best-fitting object S_{best} requires the megaset $\mathfrak{M}(\mathbb{M})$ to be topologically closed. Again we see that the property of closedness is vital for the fitting problem to have a solution for every set of data points $P_1, \ldots, P_n$.

Theorem 8.1 (closedness of megasets). *If the model collection $\mathbb{M}$ is closed (in the sense of W convergence), then the megaset $\mathfrak{M}(\mathbb{M})$ is closed in the natural topology of $\mathbb{R}^{2n}$.*

Clearly, this theorem provides an alternative proof of the existence of the best-fitting object, provided that the model collection $\mathbb{M}$ is closed.

Proof. Suppose a sequence of megapoints

$$\mathcal{P}^{(k)} = \left(x_1^{(k)}, y_1^{(k)}, \ldots, x_n^{(k)}, y_n^{(k)}\right), \tag{8.6}$$

all belonging to the megaset $\mathfrak{M}(\mathbb{M})$, converges, as $k \to \infty$, to a megapoint

$$\mathcal{P}^{(\infty)} = \left(x_1^{(\infty)}, y_1^{(\infty)}, \ldots, x_n^{(\infty)}, y_n^{(\infty)}\right) \tag{8.7}$$

in the usual topology of $\mathbb{R}^{2n}$. Equivalently, for every $i = 1, \ldots, n$ the point $(x_i^{(k)}, y_i^{(k)})$ converges, as $k \to \infty$, to the point $(x_i^{(\infty)}, y_i^{(\infty)})$. We need to show that $\mathcal{P}^{(\infty)} \in \mathfrak{M}(\mathbb{M})$.

Now for every k there exists a model object $S^{(k)} \in \mathbb{M}$ that contains the n points $(x_1^{(k)}, y_1^{(k)}), \ldots, (x_n^{(k)}, y_n^{(k)})$. The sequence $\{S^{(k)}\}$ contains a convergent subsequence S^{k_r}, that is, such that $S^{k_r} \to S$ as $r \to \infty$ for some closed set $S \subset \mathbb{R}^2$. (The existence of a convergent subsequence can be verified by the diagonal argument as in the proof of the existence theorem in Section 4.)

As we assumed that the collection $\mathbb{M}$ is closed, it follows that $\mathbb{M}$ contains the limit object S, that is, $S \in \mathbb{M}$. It is intuitively clear (and can be verified by a routine calculus-type argument) that S contains all the limit points $(x_1^{(\infty)}, y_1^{(\infty)}), \ldots, (x_n^{(\infty)}, y_n^{(\infty)})$. Therefore the limit megapoint $\mathcal{P}^{(\infty)}$ belongs to the megaset $\mathfrak{M}(\mathbb{M})$. This proves that the latter is closed. □

Megaset for Lines

Let $\mathbb{M}_L$ consist of all lines in $\mathbb{R}^2$. The corresponding megaset $\mathfrak{M}(\mathbb{M}_L) \subset \mathbb{R}^{2n}$ is described in [28, Chapter 10]. A megapoint $(x_1, y_1, \ldots, x_n, y_n)$ belongs to $\mathfrak{M}(\mathbb{M}_L)$ if and only if all the n planar points $(x_1, y_1), \ldots, (x_n, y_n)$ belong to one line (i.e., they are collinear). This condition can be expressed by $C_{n,3}$ algebraic relations as follow:

$$\det \begin{bmatrix} x_i - x_j & y_i - y_j \\ x_i - x_k & y_i - y_k \end{bmatrix} = 0 \tag{8.8}$$

for all $1 \leq i < j < k \leq n$. Each of these relations means that the three points (x_i, y_i), (x_j, y_j), and (x_k, y_k) are collinear; cf. (6.1). All of these relations together mean that all the n points $(x_1, y_1), \ldots, (x_n, y_n)$ are collinear.

Note that $\mathbb{M}_L$ is specified by $n - 2$-independent relations; hence it is an $(n + 2)$-dimensional manifold (algebraic variety) in $\mathbb{R}^{2n}$. The relations (8.8) are quadratic, so $\mathbb{M}_L$ is a quadratic surface. It is closed in topological sense; hence the problem of finding the best-fitting line always has a solution.

Megaset for Circles

Let $\mathbb{M}_C$ consist of all circles in $\mathbb{R}^2$. The corresponding megaset $\mathfrak{M}(\mathbb{M}_C) \subset \mathbb{R}^{2n}$ is described in [4, Section 3.4], and we briefly repeat it here. A megapoint $(x_1, y_1, \ldots, x_n, y_n)$ belongs to $\mathfrak{M}(\mathbb{M}_C)$ if and only if all the n planar points $(x_1, y_1), \ldots, (x_n, y_n)$ belong to one circle (we say that they are "cocircular"). In that case all these points satisfy one quadratic equation of a special type as follow:

$$A\left(x^2 + y^2\right) + Bx + Cy + D = 0. \tag{8.9}$$

This condition can be expressed by $C_{n,4}$ algebraic relations as follow:

$$\det \begin{bmatrix} x_i - x_j & y_i - y_j & x_i^2 - x_j^2 + y_i^2 - y_j^2 \\ x_i - x_k & y_i - y_k & x_i^2 - x_k^2 + y_i^2 - y_k^2 \\ x_i - x_m & y_i - y_m & x_i^2 - x_m^2 + y_i^2 - y_m^2 \end{bmatrix} = 0 \tag{8.10}$$

for $1 \leq i < j < k < m \leq n$. Note that (8.10) include megapoints satisfying (8.8); hence they actually describe the union $\mathfrak{M}(\mathbb{M}_C) \cup \mathfrak{M}(\mathbb{M}_L)$.

The determinant in (8.10) is a polynomial of degree four, and $\mathfrak{M}(\mathbb{M}_C) \cup \mathfrak{M}(\mathbb{M}_L)$ is an $(n + 3)$-dimensional algebraic variety (manifold) in $\mathbb{R}^{2n}$ defined by quartic polynomial equations. Note that the dimension of $\mathfrak{M}(\mathbb{M}_C) \cup \mathfrak{M}(\mathbb{M}_L)$ is one higher than that of $\mathfrak{M}(\mathbb{M}_L)$, that is,

$$\dim \mathfrak{M}(\mathbb{M}_C) \cup \mathfrak{M}(\mathbb{M}_L) = \dim \mathfrak{M}(\mathbb{M}_L) + 1. \tag{8.11}$$

A closer examination shows that $\mathfrak{M}(\mathbb{M}_L)$ plays the role of the boundary of $\mathfrak{M}(\mathbb{M}_C)$, that is, $\mathfrak{M}(\mathbb{M}_C)$ terminates on $\mathfrak{M}(\mathbb{M}_L)$. The megaset $\mathfrak{M}(\mathbb{M}_C)$ is not closed, but if we add its boundary $\mathfrak{M}(\mathbb{M}_L)$ to it, it will become closed.

Megaset for Ellipses

Let $\mathbb{M}_E$ consist of all ellipses in $\mathbb{R}^2$. The corresponding megaset $\mathfrak{M}(\mathbb{M}_E) \subset \mathbb{R}^{2n}$ can be described in a similar manner as above. A point $(x_1, y_1, \ldots, x_n, y_n)$ belongs in $\mathfrak{M}(\mathbb{M}_E)$ if and only if all the n planar points $(x_1, y_1), \ldots, (x_n, y_n)$ belong to one ellipse (we say that they are "coelliptical"). In that case all these points satisfy one quadratic equation of a general type as follow:

$$Ax^2 + 2Bxy + Cy^2 + 2Dx + 2Ey + F = 0. \tag{8.12}$$

This equation actually means that the points belong to one conic (either regular or degenerate). This condition can be expressed by $C_{n,6}$ algebraic relations as follow:

$$\det \begin{bmatrix} x_i - x_j & y_i - y_j & x_i^2 - x_j^2 & y_i^2 - y_j^2 & x_i y_i - x_j y_j \\ x_i - x_k & y_i - y_k & x_i^2 - x_k^2 & y_i^2 - y_k^2 & x_i y_i - x_k y_k \\ x_i - x_m & y_i - y_m & x_i^2 - x_m^2 & y_i^2 - y_m^2 & x_i y_i - x_m y_m \\ x_i - x_l & y_i - y_l & x_i^2 - x_l^2 & y_i^2 - y_l^2 & x_i y_i - x_l y_l \\ x_i - x_r & y_i - y_r & x_i^2 - x_r^2 & y_i^2 - y_r^2 & x_i y_i - x_r y_r \end{bmatrix} = 0 \tag{8.13}$$

for $1 \le i < j < k < m < l < r \le n$. Each of these relations means that the six points (x_i, y_i), (x_j, y_j), (x_k, y_k), (x_m, y_m), (x_l, y_l), and (x_r, y_r) satisfy one quadratic equation of type (8.12); that is, they belong to one conic (either regular or degenerate). All of these relations together mean that all the n points $(x_1, y_1), \ldots, (x_n, y_n)$ satisfy one quadratic equation of type (8.12); that is, they all belong to one conic (regular or degenerate). Therefore the relations (8.13) describe a much larger megaset $\mathfrak{M}(\mathbb{M}_{\text{Conics}})$ corresponding to the collection of all conics, regular and degenerate, that is,

$$\mathbb{M}_{\text{Conics}} = \mathbb{M}_E \cup \mathbb{M}_H \cup \mathbb{M}_\cup \cup \mathbb{M}_{LL} \tag{8.14}$$

using our previous notation.

The determinant in (8.13) is a polynomial of the eighth degree, and $\mathfrak{M}(\mathbb{M}_{\text{Conics}})$ is a closed $(n+5)$-dimensional algebraic manifold in $\mathbb{R}^{2n}$. It is mostly made of two big megasets: $\mathfrak{M}(\mathbb{M}_E)$ and $\mathfrak{M}(\mathbb{M}_H)$ and they both are $(n+5)$-dimensional. Other megasets (parabolas and pairs of lines) have smaller dimensions and play the role of the boundaries of the bigger megasets $\mathfrak{M}(\mathbb{M}_E)$ and $\mathfrak{M}(\mathbb{M}_H)$.

Illustrations

The structure of the megaset $\mathfrak{M}(\mathbb{M}_{\text{Conics}})$ is schematically depicted in Figure 7, where it is shown as the xy plane $\{z = 0\}$ in the 3D space (the latter plays the role of the megaspace $\mathbb{R}^{2n}$). The positive half-plane $H_+ = \{y > 0, z = 0\}$ represents the elliptic megaset $\mathfrak{M}(\mathbb{M}_E)$, and the negative half-plane $H_- = \{y < 0, z = 0\}$ represents the hyperbolic megaset $\mathfrak{M}(\mathbb{M}_H)$. The x-axis $\{y = z = 0\}$ separating these two half-planes represents the lower-dimensional megasets $\mathfrak{M}(\mathbb{M}_\cup \cup \mathbb{M}_{LL})$ in the decomposition (8.14). True, the real structure of $\mathfrak{M}(\mathbb{M}_{\text{Conics}})$ is much more complicated, but our simplified picture still shows its most basic features.

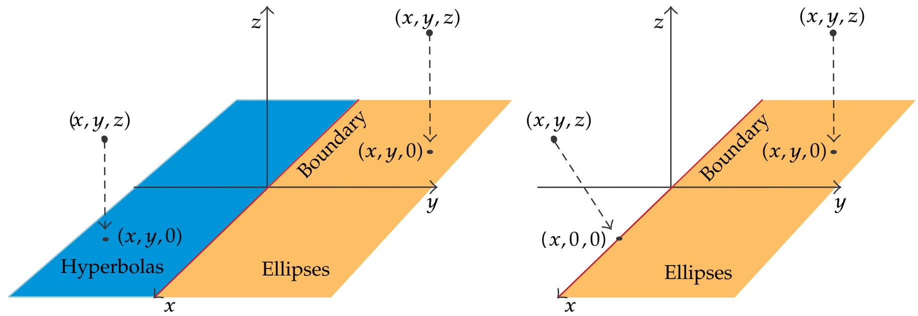


Figure 7: Schematic illustration of the megaset $\mathfrak{M}(\mathbb{M}_{\text{Conics}})$.

Now recall that finding the best-fitting conic corresponds to an orthogonal projection of the given megapoint $\wp$ in the megaspace $\mathbb{R}^{2n}$ (in our illustration, it would be a point $(x, y, z) \in \mathbb{R}^3$) onto the megaset $\mathfrak{M}(\mathbb{M}_{\text{Conics}})$ (in our illustration—onto the xy plane). In Figure 7 the point (x, y, z) is simply projected onto $(x, y, 0)$. What are the chances that the footpoint of the projection corresponds to a "boundary" object in $\mathbb{M}_{\cup} \cup \mathbb{M}_{LL}$ (i.e., to a secondary object)? Clearly, only the points of the xz plane $\{y = 0\}$ are projected onto the line $\{y = z = 0\}$. If the point $(x, y, z) \in \mathbb{R}^3$ is selected randomly with an absolutely continuous distribution, then a point on the xz plane would be chosen with probability zero. This fact illustrates the sufficiency of the model collection of nondegenerate conics (even the sufficiency of ellipses and hyperbolas alone, without parabolas); the best-fitting object will be a primary object with probability one.

But what if our model collection consists of ellipses only, without hyperbolas? Then in our illustration, the corresponding megaset would be the positive half-plane $H_+ = \{y > 0, z = 0\}$. Finding the best-fitting ellipse would correspond to an orthogonal projection of the given point $(x, y, z) \in \mathbb{R}^3$ onto the positive half-plane $H_+ = \{y > 0, z = 0\}$. Now if the given point (x, y, z) has a positive y-coordinate, then it is projected onto $(x, y, 0)$, as before, and we get the desired best-fitting ellipse. But if it has a negative y-coordinate, then it is projected onto $(x, 0, 0)$, which is on the *boundary* of the half-plane, so we get a boundary footpoint; that is, a *secondary* object will be the best fit. We see that all the points (x, y, z) with $y < 0$ (making a whole half-space!) are projected onto the boundary line, hence for all those points the best-fitting ellipse would not exist! This fact clearly illustrates the deficiency of the model collection of ellipses.

One may wonder: How is it possible that the collection of circles is sufficient (as we proved in Section 5), while the larger collection of ellipses is not? Indeed every circle is an ellipse, hence the collection of ellipses contains all the circles. So why the sufficiency of circles does not guarantee the sufficiency of the bigger, inclusive collection of ellipses? Well, this seemingly counterintuitive fact can be illustrated too (see Figure 8).

Suppose that the megaset $\mathfrak{M}(\mathbb{M}_C)$ for the collection of circles is represented by the set $U = \{y = x^2, x \neq 0, z = 0\}$ in our illustration. Note that U consists of two curves (branches of a parabola on the xy plane), both lie in the half-plane $H_+ = \{y > 0, z = 0\}$ that corresponds to the collection of ellipses. So the required inclusion $U \subset H_+$ does take place. The two curves making U terminate at the point $(0, 0, 0)$, which does not belong to U, so it plays the role of the boundary of U.

Now suppose a randomly selected point $(x, y, z) \in \mathbb{R}^3$ is to be projected onto the set U. What are the chances that its projection will end up on the boundary of U, that is, at the origin

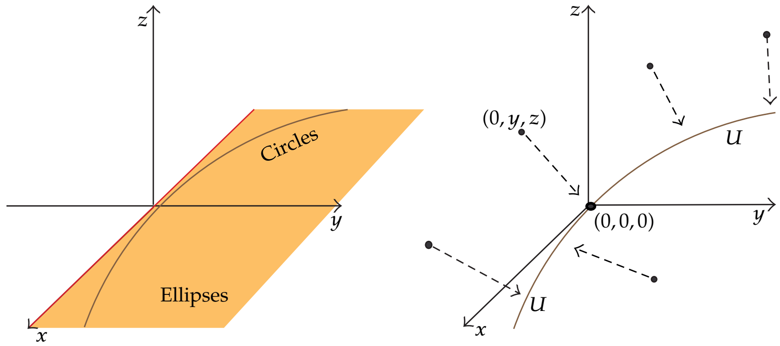


Figure 8: Schematic illustration of the megaset $\mathfrak{M}(\mathbb{M}_C) \subset \mathfrak{M}(\mathbb{M}_E)$.

$(0, 0, 0)$? It is not hard to see (and prove by elementary geometry) that only points on the yz plane may be projected onto the origin $(0, 0, 0)$ (and not even all of them; points with large positive y-coordinates would be projected onto some interior points of U). So the chance that the footpoint of the projection ends up at the boundary of U is zero. This illustrates the sufficiency of the *smaller* model collection of circles, despite the deficiency of the *larger* model collection of ellipses (which we have seen in Figure 7).

Appendix

Here we prove the theorem on no local minima for $n = 5$ (Section 6).

Proof. Suppose that S is a conic on which $\mathcal{F}$ takes a local minimum. We will assume that S is nondegenerate; that is, S is an ellipse, or hyperbola, or parabola. Degenerate conics are easier to handle, so we omit them.

Let P_1, P_2, P_3, P_4, P_5 denote the given points and Q_1, Q_2, Q_3, Q_4, and Q_5 their projections onto S. Since S does not interpolate all the given points simultaneously, there exists at least one which is not on S. Suppose that it is P_1, hence $P_1 \neq Q_1$.

Let us first assume that Q_1 is different from the other projections Q_2, Q_3, Q_4, andQ_5. Then let Q_1' be a point near Q_1 such that $\operatorname{dist}(Q_1', P_1) < \operatorname{dist}(Q_1, P_1)$. In other words, we perturb Q_1 slightly by moving it *in the direction of* P_1. Then there exists another non-degenerate conic S' that interpolates the points Q_1', Q_2, Q_3, Q_4, and Q_5. It is easy to see that

$$\begin{gathered} \operatorname{dist}(P_1, S') \leq \operatorname{dist}(P_1, Q_1') < \operatorname{dist}(P_1, Q_1) = \operatorname{dist}(P_1, S), \\ \operatorname{dist}(P_i, S') \leq \operatorname{dist}(P_i, Q_i) = \operatorname{dist}(P_i, S) \quad \text{for every } i = 2, \ldots, 5. \end{gathered} \tag{A.1}$$

As a result,

$$\mathcal{F}(S') = \sum_{i=1}^{5} [\operatorname{dist}(P_i, S')]^2 < \sum_{i=1}^{5} [\operatorname{dist}(P_i, S)]^2 = \mathcal{F}(S), \tag{A.2}$$

as desired. Since Q_1' can be placed arbitrarily close to Q_1, the new conic S' can be made arbitrarily close to S.

Now suppose that the projection point Q_1 coincides with some other projection point(s). In that case there are at most *four* distinct projection points. We will not move the point Q_1, but we will rotate (slightly) the tangent line to the conic S at the point Q_1. Denote that tangent line by T. Note that T is perpendicular to the line passing through P_1 and Q_1.

Now let T' be another line passing through the point Q_1 and making an arbitrarily small angle with T. It is not hard to show, by elementary geometry, that there exists another nondegenerate conic S' passing though the same (≤ 4) projection points and having the tangent line T' at Q_1.

Since the tangent line T' to the conic S' at the point Q_1 is *not* orthogonal to the line P_1Q_1, we easily have

$$\operatorname{dist}(P_1, S') < \operatorname{dist}(P_1, Q_1) = \operatorname{dist}(P_1, S). \tag{A.3}$$

At the same time we have

$$\operatorname{dist}(P_i, S') \leq \operatorname{dist}(P_i, Q_i) = \operatorname{dist}(P_i, S) \quad \text{for every } i = 2, \ldots, 5. \tag{A.4}$$

As a result, consider

$$\mathcal{F}(S') = \sum_{i=1}^{5} \left[\operatorname{dist}(P_i, S')\right]^2 < \sum_{i=1}^{5} \left[\operatorname{dist}(P_i, S)\right]^2 = \mathcal{F}(S), \tag{A.5}$$

as desired. Since the line T' can be selected arbitrarily close to T, the new conic S' will be arbitrarily close to S. $\square$

Acknowledgment

N. Chernov was partially supported by National Science Foundation, Grant DMS-0969187.

References

[1] S. J. Ahn, *Least Squares Orthogonal Distance Fitting of Curves and Surfaces in Space*, vol. 3151 of *Lecture Notes in Computer Science*, Springer, Berlin, Germany, 2004.

[2] Geometric Product Specification (GPS), "Acceptance and representation test for coordinate measuring machines (CMM)—Part 6: estimation of errors in computing Gaussian associated features," Int'l Standard ISO 10360-6, International standard, Geneva, Switzerland, 2001.

[3] C.-L. Cheng and J. W. Van Ness, *Statistical Regression with Measurement Error*, vol. 6, Arnold, London, UK, 1999.

[4] N. Chernov, *Circular and Linear Regression: Fitting Circles and Lines by Least Squares*, vol. 117 of *Monographs on Statistics and Applied Probability*, Chapman & Hall/CRC, London, UK, 2011.

[5] R. J. Adcock, "Note on the method of least squares," *Analyst*, vol. 4, pp. 183–184, 1877.

[6] C. H. Kummell, "Reduction of observation equations which contain more than one observed quantity," *Analyst*, vol. 6, pp. 97–105, 1879.

[7] A. Madansky, "The fitting of straight lines when both variables are subject to error," *Journal of the American Statistical Association*, vol. 54, pp. 173–205, 1959.

[8] C. Villegas, "Maximum likelihood estimation of a linear functional relationship," *Annals of Mathematical Statistics*, vol. 32, pp. 1048–1062, 1961.

[9] A. Wald, "The fitting of straight lines if both variables are subject to error," *Annals of Mathematical Statistics*, vol. 11, pp. 285–300, 1940.

[10] T. W. Anderson, "Estimation of linear functional relationships: approximate distributions and connections with simultaneous equations in econometrics," *Journal of the Royal Statistical Society B*, vol. 38, no. 1, pp. 1–36, 1976.

[11] T. W. Anderson and T. Sawa, "Exact and approximate distributions of the maximum likelihood estimator of a slope coefficient," *Journal of the Royal Statistical Society B*, vol. 44, no. 1, pp. 52–62, 1982.

[12] S. M. Robinson, "Fitting spheres by the method of least squares," *Communications of the Association for Computing Machinery*, vol. 4, p. 491, 1961.

[13] A. Thom, "A statistical examination of the megalithic sites in Britain," *Journal of the Royal Statistical Society A*, vol. 118, pp. 275–295, 1955.

[14] A. Albano, "Representation of digitized contours in terms of conic arcs and straight-line segments," *Computer Graphics and Image Processing*, vol. 3, pp. 23–33, 1974.

[15] R. H. Biggerstaff, "Three variations in dental arch form estimated by a quadratic equation," *Journal of Dental Research*, vol. 51, no. 5, article 1509, 1972.

[16] F. L. Bookstein, "Fitting conic sections to scattered data," *Computer Graphics and Image Processing*, vol. 9, no. 1, pp. 56–71, 1979.

[17] K. Paton, "Conic sections in chromosome analysis," *Pattern Recognition*, vol. 2, no. 1, pp. 39–51, 1970.

[18] G. Taubin, "Estimation of planar curves, surfaces, and nonplanar space curves defined by implicit equations with applications to edge and range image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 13, no. 11, pp. 1115–1138, 1991.

[19] S. C. Pei and J. H. Horng, "Optimum approximation of digital planar curves using circular arcs," *Pattern Recognition*, vol. 29, no. 3, pp. 383–388, 1996.

[20] B. Sarkar, L. K. Singh, and D. Sarkar, "Approximation of digital curves with line segments and circular arcs using genetic algorithms," *Pattern Recognition Letters*, vol. 24, no. 15, pp. 2585–2595, 2003.

[21] Y. Nievergelt, "A finite algorithm to fit geometrically all midrange lines, circles, planes, spheres, hyperplanes, and hyperspheres," *Numerische Mathematik*, vol. 91, no. 2, pp. 257–303, 2002.

[22] M. Berman and D. Culpin, "The statistical behaviour of some least squares estimators of the centre and radius of a circle," *Journal of the Royal Statistical Society B*, vol. 48, no. 2, pp. 183–196, 1986.

[23] N. Chernov and C. Lesort, "Least squares fitting of circles," *Journal of Mathematical Imaging and Vision*, vol. 23, no. 3, pp. 239–252, 2005.

[24] E. E. Zelniker and I. V. L. Clarkson, "A statistical analysis of the Delogne-Kåsa method for fitting circles," *Digital Signal Processing*, vol. 16, no. 5, pp. 498–522, 2006.

[25] Y. Nievergelt, "Fitting conics of specific types to data," *Linear Algebra and its Applications*, vol. 378, pp. 1–30, 2004.

[26] B. Matei and P. Meer, "Reduction of bias in maximum likelihood ellipse fitting," in *Proceedings of the 15th International Conference on Computer Vision and Pattern Recognition (ICCVPR '00)*, vol. 3, pp. 802–806, Barcelona, Spain, September 2000.

[27] A. Fitzgibbon, M. Pilu, and R. B. Fisher, "Direct least square fitting of ellipses," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 21, no. 5, pp. 476–480, 1999.

[28] E. Malinvaud, *Statistical Methods of Econometrics*, vol. 6, North-Holland Publishing, Amsterdam The Netherland, 3rd edition, 1980.