Statistical efficiency of curve fitting algorithms

N. Chernov and C. Lesort
Department of Mathematics
University of Alabama at Birmingham
Birmingham, AL 35294, USA

February 1, 2008

Abstract

We study the problem of fitting parametrized curves to noisy data. Under certain assumptions (known as Cartesian and radial functional models), we derive asymptotic expressions for the bias and the covariance matrix of the parameter estimates. We also extend Kanatani's version of the Cramer-Rao lower bound, which he proved for unbiased estimates only, to more general estimates that include many popular algorithms (most notably, the orthogonal least squares and algebraic fits). We then show that the gradient-weighted algebraic fit is statistically efficient and describe all other statistically efficient algebraic fits.

Keywords: least squares fit, curve fitting, circle fitting, algebraic fit, Rao-Cramer bound, efficiency, functional model.

1 Introduction

In many applications one fits a parametrized curve described by an implicit equation $P(x, y; \Theta) = 0$ to experimental data (x_i, y_i), $i = 1, \dots, n$. Here Θ denotes the vector of unknown parameters to be estimated. Typically, P is a polynomial in x and y, and its coefficients are unknown parameters (or functions of unknown parameters). For example, a number of recent publications [2, 10, 11, 16, 19] are devoted to the problem of fitting quadrics $Ax^2+Bxy+Cy^2+Dx+Ey+F = 0$, in which case $\Theta = (A, B, C, D, E, F)$ is the parameter vector. The problem of fitting circles, given by equation $(x-a)^2+(y-b)^2-R^2 = 0$ with three parameters a, b, R, also attracted attention [8, 14, 15, 18].

We consider here the problem of fitting general curves given by implicit equations $P(x, y; \Theta) = 0$ with $\Theta = (\theta_1, \dots, \theta_k)$ being the parameter vector. Our goal is to investigate statistical properties of various fitting algorithms. We are interested in their biasedness, covariance matrices, and the Cramer-Rao lower bound.

First, we specify our model. We denote by $\bar{\Theta}$ the true value of Θ. Let $(\bar{x}_i, \bar{y}_i)$, $i = 1, \ldots, n$, be some points lying on the true curve $P(x, y; \bar{\Theta}) = 0$. Experimentally observed data points (x_i, y_i), $i = 1, \ldots, n$, are perceived as random perturbations of the true points $(\bar{x}_i, \bar{y}_i)$. We use notation $\mathbf{x}_i = (x_i, y_i)^T$ and $\bar{\mathbf{x}}_i = (\bar{x}_i, \bar{y}_i)^T$, for brevity. The random vectors $\mathbf{e}_i = \mathbf{x}_i - \bar{\mathbf{x}}_i$ are assumed to be independent and have zero mean. Two specific assumptions on their probability distribution can be made, see [4]:

> *Cartesian model*: Each $\mathbf{e}_i$ is a two-dimensional normal vector with covariance matrix $\sigma_i^2 I$, where I is the identity matrix.
>
> *Radial model*: $\mathbf{e}_i = \xi_i \mathbf{n}_i$ where ξ_i is a normal random variable $\mathcal{N}(0, \sigma_i^2)$, and $\mathbf{n}_i$ is a unit normal vector to the curve $P(x, y; \bar{\Theta}) = 0$ at the point $\mathbf{x}_i$.

Our analysis covers both models, Cartesian and radial. For simplicity, we assume that $\sigma_i^2 = \sigma^2$ for all i, but note that our results can be easily generalized to arbitrary $\sigma_i^2 > 0$.

Concerning the true points $\bar{\mathbf{x}}_i$, $i = 1, \ldots, n$, two assumptions are possible. Many researchers [6, 13, 14] consider them as fixed, but unknown, points on the true curve. In this case their coordinates $(\bar{x}_i, \bar{y}_i)$ can be treated as additional parameters of the model (nuisance parameters). Chan [6] and others [3, 4] call this assumption a *functional model*. Alternatively, one can assume that the true points $\bar{\mathbf{x}}_i$ are sampled from the curve $P(x, y; \bar{\Theta}) = 0$ according to some probability distribution on it. This assumption is referred to as a *structural model* [3, 4]. We only consider the functional model here.

It is easy to verify that maximum likelihood estimation of the parameter Θ for the functional model is given by the orthogonal least squares fit (OLSF), which is based on minimization of the function

$$\mathcal{F}_1(\Theta) = \sum_{i=1}^{n} [d_i(\Theta)]^2 \tag{1.1}$$

where $d_i(\Theta)$ denotes the distance from the point $\mathbf{x}_i$ to the curve $P(x, y; \Theta) = 0$. The OLSF is the method of choice in practice, especially when one fits simple curves such as lines and circles. However, for more general curves the OLSF becomes intractable, because the precise distance d_i is hard to compute. For example, when P is a generic quadric (ellipse or hyperbola), the computation of d_i is equivalent to solving a polynomial equation of degree four, and its direct solution is known to be numerically unstable, see [2, 11] for more detail. Then one resorts to various approximations. It is often convenient to minimize

$$\mathcal{F}_2(\Theta) = \sum_{i=1}^{n} [P(x_i, y_i; \Theta)]^2 \tag{1.2}$$

instead of (1.1). This method is referred to as a (simple) *algebraic fit* (AF), in this case one calls $|P(x_i, y_i; \Theta)|$ the *algebraic distance* [2, 10, 11] from the point (x_i, y_i) to the curve. The AF is computationally cheaper than the OLSF, but its accuracy is often unacceptable, see below.

The simple AF (1.2) can be generalized to a *weighted algebraic fit*, which is based on minimization of

$$\mathcal{F}_3(\Theta) = \sum_{i=1}^{n} w_i \left[P(x_i, y_i; \Theta)\right]^2 \tag{1.3}$$

where $w_i = w(x_i, y_i; \Theta)$ are some weights, which may balance (1.2) and improve its performance. One way to define weights w_i results from a linear approximation to d_i:

$$d_i \approx \frac{|P(x_i, y_i; \Theta)|}{\|\nabla_{\mathbf{x}} P(x_i, y_i; \Theta)\|}$$

where $\nabla_{\mathbf{x}} P = (\partial P/\partial x, \partial P/\partial y)$ is the gradient vector, see [20]. Then one minimizes the function

$$\mathcal{F}_4(\Theta) = \sum_{i=1}^{n} \frac{[P(x_i, y_i; \Theta)]^2}{\|\nabla_{\mathbf{x}} P(x_i, y_i; \Theta)\|^2} \tag{1.4}$$

This method is called the *gradient weighted algebraic fit* (GRAF). It is a particular case of (1.3) with $w_i = 1/\|\nabla_{\mathbf{x}} P(x_i, y_i; \Theta)\|^2$.

The GRAF is known since at least 1974 [21] and recently became standard for polynomial curve fitting [20, 16, 10]. The computational cost of GRAF depends on the function $P(x, y; \Theta)$, but, generally, the GRAF is much faster than the OLSF. It is also known from practice that the accuracy of GRAF is almost as good as that of the OLSF, and our analysis below confirms this fact. The GRAF is often claimed to be a *statistically optimal* weighted algebraic fit, and we will prove this fact as well.

Not much has been published on statistical properties of the OLSF and algebraic fits, apart from the simplest case of fitting lines and hyperplanes [12]. Chan [6], Berman and Culpin [4] investigated circle fitting by the OLSF and the simple algebraic fit (1.2) assuming the structural model. Kanatani [13, 14] used the Cartesian functional model and considered a general curve fitting problem. He established an analogue of the Rao-Cramer lower bound for unbiased estimates of Θ, which we call here Kanatani-Cramer-Rao (KCR) lower bound. He also showed that the covariance matrices of the OLSF and the GRAF attain, to the leading order in σ, his lower bound. We note, however, that in most cases the OLSF and algebraic fits are *biased* [4, 5], hence the KCR lower bound, as it is derived in [13, 14], does not immediately apply to these methods.

In this paper we extend the KCR lower bound to biased estimates, which include the OLSF and all weighted algebraic fits. We prove the KCR bound for estimates satisfying the following mild assumption:

Precision assumption. For precise observations (when $\mathbf{x}_i = \bar{\mathbf{x}}_i$ for all $1 \le i \le n$), the estimate $\hat{\Theta}$ is precise, i.e.

$$\hat{\Theta}(\bar{\mathbf{x}}_1, \dots, \bar{\mathbf{x}}_n) = \bar{\Theta} \tag{1.5}$$

It is easy to check that the OLSF and algebraic fits (1.3) satisfy this assumption. We will also show that all unbiased estimates of $\hat{\Theta}$ satisfy (1.5).

We then prove that the GRAF is, indeed, a statistically efficient fit, in the sense that its covariance matrix attains, to the leading order in σ, the KCR lower bound. On the

other hand, rather surprisingly, we find that GRAF is not the only statistically efficient algebraic fit, and we describe all statistically efficient algebraic fits. Finally, we show that Kanatani's theory and our extension to it remain valid for the radial functional model. Our conclusions are illustrated by numerical experiments on circle fitting algorithms.

2 Kanatani-Cramer-Rao lower bound

Recall that we have adopted the functional model, in which the true points $\bar{\mathbf{x}}_i$, $1 \le i \le n$, are fixed. This automatically makes the sample size n fixed, hence, many classical concepts of statistics, such as consistency and asymptotic efficiency (which require taking the limit $n \to \infty$) lose their meaning. It is customary, in the studies of the functional model of the curve fitting problem, to take the limit $\sigma \to 0$ instead of $n \to \infty$, cf. [13, 14]. This is, by the way, not unreasonable from the practical point of view: in many experiments, n is rather small and cannot be (easily) increased, so the limit $n \to \infty$ is of little interest. On the other hand, when the accuracy of experimental observations is high (thus, σ is small), the limit $\sigma \to 0$ is quite appropriate.

Now, let $\hat{\Theta}(\mathbf{x}_1, \dots, \mathbf{x}_n)$ be an arbitrary estimate of Θ satisfying the precision assumption (1.5). In our analysis we will always assume that all the underlying functions are regular (continuous, have finite derivatives, etc.), which is a standard assumption [13, 14].

The mean value of the estimate $\hat{\Theta}$ is

$$E(\hat{\Theta}) = \int \cdots \int \hat{\Theta}(\mathbf{x}_1, \dots, \mathbf{x}_n) \prod_{i=1}^{n} f(\mathbf{x}_i)\, d\mathbf{x}_1 \cdots d\mathbf{x}_n \tag{2.1}$$

where $f(\mathbf{x}_i)$ is the probability density function for the random point $\mathbf{x}_i$, as specified by a particular model (Cartesian or radial).

We now expand the estimate $\hat{\Theta}(\mathbf{x}_1, \dots, \mathbf{x}_n)$ into a Taylor series about the true point $(\bar{\mathbf{x}}_1, \dots, \bar{\mathbf{x}}_n)$ remembering (1.5):

$$\hat{\Theta}(\mathbf{x}_1, \dots, \mathbf{x}_n) = \bar{\Theta} + \sum_{i=1}^{n} \Theta_i \times (\mathbf{x}_i - \bar{\mathbf{x}}_i) + \mathcal{O}(\sigma^2) \tag{2.2}$$

where

$$\Theta_i = \nabla_{\mathbf{x}_i} \hat{\Theta}(\bar{\mathbf{x}}_1, \dots, \bar{\mathbf{x}}_n), \qquad i = 1, \dots, n \tag{2.3}$$

and $\nabla_{\mathbf{x}_i}$ stands for the gradient with respect to the variables x_i, y_i. In other words, Θ_i is a $k \times 2$ matrix of partial derivatives of the k components of the function $\hat{\Theta}$ with respect to the two variables x_i and y_i, and this derivative is taken at the point $(\bar{\mathbf{x}}_1, \dots, \bar{\mathbf{x}}_n)$,

Substituting the expansion (2.2) into (2.1) gives

$$E(\hat{\Theta}) = \bar{\Theta} + \mathcal{O}(\sigma^2) \tag{2.4}$$

since $E(\mathbf{x}_i - \bar{\mathbf{x}}_i) = 0$. Hence, the bias of the estimate $\hat{\Theta}$ is of order σ^2.

It easily follows from the expansion (2.2) that the covariance matrix of the estimate $\hat{\Theta}$ is given by

$$\mathcal{C}_{\hat{\Theta}} = \sum_{i=1}^{n} \Theta_i E[(\mathbf{x}_i - \bar{\mathbf{x}}_i)(\mathbf{x}_i - \bar{\mathbf{x}}_i)^T]\Theta_i^T + \mathcal{O}(\sigma^4)$$

(it is not hard to see that the cubical terms $\mathcal{O}(\sigma^3)$ vanish because the normal random variables with zero mean also have zero third moment, see also [13]). Now, for the Cartesian model

$$E[(\mathbf{x}_i - \bar{\mathbf{x}}_i)(\mathbf{x}_i - \bar{\mathbf{x}}_i)^T] = \sigma^2 I$$

and for the radial model

$$E[(\mathbf{x}_i - \bar{\mathbf{x}}_i)(\mathbf{x}_i - \bar{\mathbf{x}}_i)^T] = \sigma^2 \mathbf{n}_i \mathbf{n}_i^T$$

where $\mathbf{n}_i$ is a unit normal vector to the curve $P(x, y; \bar{\Theta}) = 0$ at the point $\bar{\mathbf{x}}_i$. Then we obtain

$$\mathcal{C}_{\hat{\Theta}} = \sigma^2 \sum_{i=1}^{n} \Theta_i \Lambda_i \Theta_i^T + \mathcal{O}(\sigma^4) \tag{2.5}$$

where $\Lambda_i = I$ for the Cartesian model and $\Lambda_i = \mathbf{n}_i \mathbf{n}_i^T$ for the radial model.

Lemma. *We have $\Theta_i \mathbf{n}_i \mathbf{n}_i^T \Theta_i^T = \Theta_i \Theta_i^T$ for each $i = 1, \ldots, n$. Hence, for both models, Cartesian and radial, the matrix $\mathcal{C}_{\hat{\Theta}}$ is given by the same expression:*

$$\mathcal{C}_{\hat{\Theta}} = \sigma^2 \sum_{i=1}^{n} \Theta_i \Theta_i^T + \mathcal{O}(\sigma^4) \tag{2.6}$$

This lemma is proved in Appendix.

Our next goal is now to find a lower bound for the matrix

$$\mathcal{D}_1 := \sum_{i=1}^{n} \Theta_i \Theta_i^T \tag{2.7}$$

Following [13, 14], we consider perturbations of the parameter vector $\bar{\Theta} + \delta\Theta$ and the true points $\bar{\mathbf{x}}_i + \delta\bar{\mathbf{x}}_i$ satisfying two constraints. First, since the true points must belong to the true curve, $P(\bar{\mathbf{x}}_i; \bar{\Theta}) = 0$, we obtain, by the chain rule,

$$\langle \nabla_{\mathbf{x}} P(\bar{\mathbf{x}}_i; \bar{\Theta}), \delta\bar{\mathbf{x}}_i \rangle + \langle \nabla_{\Theta} P(\bar{\mathbf{x}}_i; \bar{\Theta}), \delta\Theta \rangle = 0 \tag{2.8}$$

where $\langle \cdot, \cdot \rangle$ stands for the scalar product of vectors. Second, since the identity (1.5) holds for all Θ, we get

$$\sum_{i=1}^{n} \Theta_i \, \delta\bar{\mathbf{x}}_i = \delta\Theta \tag{2.9}$$

by using the notation (2.3).

Now we need to find a lower bound for the matrix (2.7) subject to the constraints (2.8) and (2.9). That bound follows from a general theorem in linear algebra:

Theorem (Linear Algebra). *Let $n \geq k \geq 1$ and $m \geq 1$. Suppose n nonzero vectors $u_i \in \mathbb{R}^m$ and n nonzero vectors $v_i \in \mathbb{R}^k$ are given, $1 \leq i \leq n$. Consider $k \times m$ matrices*

$$X_i = \frac{v_i u_i^T}{u_i^T u_i}$$

for $1 \leq i \leq n$, and $k \times k$ matrix

$$B = \sum_{i=1}^{n} X_i X_i^T = \sum_{i=1}^{n} \frac{v_i v_i^T}{u_i^T u_i}$$

Assume that the vectors $v_1, \ldots, v_n$ span $\mathbb{R}^k$ (hence B is nonsingular). We say that a set of n matrices $A_1, \ldots, A_n$ (each of size $k \times m$) is **proper** *if*

$$\sum_{i=1}^{n} A_i w_i = r \tag{2.10}$$

for any vectors $w_i \in \mathbb{R}^m$ and $r \in \mathbb{R}^k$ such that

$$u_i^T w_i + v_i^T r = 0 \tag{2.11}$$

for all $1 \leq i \leq n$. Then for any proper set of matrices $A_1, \ldots, A_n$ the $k \times k$ matrix $D = \sum_{i=1}^{n} A_i A_i^T$ is bounded from below by B^{-1} in the sense that $D - B^{-1}$ is a positive semidefinite matrix. The equality $D = B^{-1}$ holds if and only if $A_i = -B^{-1} X_i$ for all $i = 1, \ldots, n$.

This theorem is, probably, known, but we provide a full proof in Appendix, for the sake of completeness.

As a direct consequence of the above theorem we obtain the lower bound for our matrix $\mathcal{D}_1$:

Theorem (Kanatani-Cramer-Rao lower bound). *We have $\mathcal{D}_1 \geq \mathcal{D}_{\min}$, in the sense that $\mathcal{D}_1 - \mathcal{D}_{\min}$ is a positive semidefinite matrix, where*

$$\mathcal{D}_{\min}^{-1} = \sum_{i=1}^{n} \frac{(\nabla_\Theta P(\bar{\mathbf{x}}_i; \Theta))(\nabla_\Theta P(\bar{\mathbf{x}}_i; \Theta))^T}{\|\nabla_{\mathbf{x}} P(\bar{\mathbf{x}}_i; \Theta)\|^2} \tag{2.12}$$

In view of (2.6) and (2.7), the above theorem says that the lower bound for the covariance matrix $\mathcal{C}_{\hat{\Theta}}$ is, to the leading order,

$$\mathcal{C}_{\hat{\Theta}} \geq \mathcal{C}_{\min} = \sigma^2 \mathcal{D}_{\min} \tag{2.13}$$

The standard deviations of the components of the estimate $\hat{\Theta}$ are of order $\sigma_{\hat{\Theta}} = \mathcal{O}(\sigma)$. Therefore, the bias of $\hat{\Theta}$, which is at most of order σ^2 by (2.4), is infinitesimally small, as $\sigma \to 0$, compared to the standard deviations. This means that the estimates satisfying (1.5) are practically unbiased.

The bound (2.13) was first derived by Kanatani [13, 14] for the Cartesian functional model and strictly unbiased estimates of Θ, i.e. satisfying $E(\hat{\Theta}) = \bar{\Theta}$. One can easily derive (1.5) from $E(\hat{\Theta}) = \bar{\Theta}$ by taking the limit $\sigma \to 0$, hence our results generalize those of Kanatani.

3 Statistical efficiency of algebraic fits

Here we derive an explicit formula for the covariance matrix of the weighted algebraic fit (1.3) and describe the weights w_i for which the fit is statistically efficient. For brevity, we write $P_i = P(x_i, y_i; \Theta)$. We assume that the weight function $w(x, y, ; \Theta)$ is regular, in particular has bounded derivatives with respect to Θ, the next section will demonstrate the importance of this condition. The solution of the minimization problem (1.3) satisfies

$$\sum P_i^2 \, \nabla_\Theta w_i + 2 \sum w_i \, P_i \, \nabla_\Theta P_i = 0 \tag{3.1}$$

Observe that $P_i = \mathcal{O}(\sigma)$, so that the first sum in (3.1) is $\mathcal{O}(\sigma^2)$ and the second sum is $\mathcal{O}(\sigma)$. Hence, to the leading order, the solution of (3.1) can be found by discarding the first sum and solving the reduced equation

$$\sum w_i \, P_i \, \nabla_\Theta P_i = 0 \tag{3.2}$$

More precisely, if $\hat{\Theta}_1$ and $\hat{\Theta}_2$ are solutions of (3.1) and (3.2), respectively, then $\hat{\Theta}_1 - \bar{\Theta} = \mathcal{O}(\sigma)$, $\hat{\Theta}_2 - \bar{\Theta} = \mathcal{O}(\sigma)$, and $\|\hat{\Theta}_1 - \hat{\Theta}_2\| = \mathcal{O}(\sigma^2)$. Furthermore, the covariance matrices of $\hat{\Theta}_1$ and $\hat{\Theta}_2$ coincide, to the leading order, i.e. $\mathcal{C}_{\hat{\Theta}_1} \mathcal{C}_{\hat{\Theta}_2}^{-1} \to I$ as $\sigma \to 0$. Therefore, in what follows, we only deal with the solution of equation (3.2).

To find the covariance matrix of $\hat{\Theta}$ satisfying (3.2) we put $\hat{\Theta} = \bar{\Theta} + \delta\Theta$ and $\mathbf{x}_i = \bar{\mathbf{x}}_i + \delta\mathbf{x}_i$ and obtain, working to the leading order,

$$\sum w_i (\nabla_\Theta P_i)(\nabla_\Theta P_i)^T \, (\delta\Theta) = -\sum w_i (\nabla_{\mathbf{x}} P_i)^T \, (\delta\mathbf{x}_i) \, (\nabla_\Theta P_i) + \mathcal{O}(\sigma^2)$$

hence

$$\delta\Theta = -\Big[\sum w_i (\nabla_\Theta P_i)(\nabla_\Theta P_i)^T\Big]^{-1} \Big[\sum w_i (\nabla_{\mathbf{x}} P_i)^T \, (\delta\mathbf{x}_i) \, (\nabla_\Theta P_i)\Big] + \mathcal{O}(\sigma^2)$$

The covariance matrix is then

$$\begin{aligned} \mathcal{C}_{\hat{\Theta}} &= E\Big[(\delta\Theta) \, (\delta\Theta)^T\Big] \\ &= \sigma^2 \Big[\sum w_i (\nabla_\Theta P_i)(\nabla_\Theta P_i)^T\Big]^{-1} \Big[\sum w_i^2 \|\nabla_{\mathbf{x}} P_i\|^2 (\nabla_\Theta P_i)(\nabla_\Theta P_i)^T\Big] \\ &\quad \times \Big[\sum w_i (\nabla_\Theta P_i)(\nabla_\Theta P_i)^T\Big]^{-1} + \mathcal{O}(\sigma^3) \end{aligned}$$

Denote by $\mathcal{D}_2$ the principal factor here, i.e.

$$\mathcal{D}_2 = \left[\sum w_i(\nabla_\Theta P_i)(\nabla_\Theta P_i)^T\right]^{-1} \left[\sum w_i^2 \|\nabla_{\mathbf{x}} P_i\|^2 (\nabla_\Theta P_i)(\nabla_\Theta P_i)^T\right] \left[\sum w_i(\nabla_\Theta P_i)(\nabla_\Theta P_i)^T\right]^{-1}$$

The following theorem establishes a lower bound for $\mathcal{D}_2$:

Theorem. *We have $\mathcal{D}_2 \geq \mathcal{D}_{\min}$, in the sense that $\mathcal{D}_2 - \mathcal{D}_{\min}$ is a positive semidefinite matrix, where $\mathcal{D}_{\min}$ is given by (2.12). The equality $\mathcal{D}_2 = \mathcal{D}_{\min}$ holds if and only if $w_i = \text{const}/\|\nabla_{\mathbf{x}} P_i\|^2$ for all $i = 1, \ldots, n$. In other words, an algebraic fit (1.3) is* **statistically efficient** *if and only if the weight function $w(x, y; \Theta)$ satisfies*

$$w(x, y; \Theta) = \frac{c(\Theta)}{\|\nabla_{\mathbf{x}} P(x, y; \Theta)\|^2} \tag{3.3}$$

for all triples x, y, Θ such that $P(x, y; \Theta) = 0$. Here $c(\Theta)$ may be an arbitrary function of Θ.

The bound $\mathcal{D}_2 \geq \mathcal{D}_{\min}$ here is a particular case of the previous theorem. It also can be obtained directly from the linear algebra theorem if one sets $u_i = \nabla_{\mathbf{x}} P_i$, $v_i = \nabla_\Theta P_i$, and

$$A_i = -w_i \left[\sum_{j=1}^{n} w_j(\nabla_\Theta P_j)(\nabla_\Theta P_j)^T\right]^{-1} (\nabla_\Theta P_i)\,(\nabla_{\mathbf{x}} P_i)^T$$

for $1 \leq i \leq n$.

The expression (3.3) characterizing the efficiency, follows from the last claim in the linear algebra theorem.

4 Circle fit

Here we illustrate our conclusions by the relatively simple problem of fitting circles. The canonical equation of a circle is

$$(x - a)^2 + (y - b)^2 - R^2 = 0 \tag{4.1}$$

and we need to estimate three parameters a, b, R. The simple algebraic fit (1.2) takes form

$$\mathcal{F}_2(a, b, R) = \sum_{i=1}^{n} [(x_i - a)^2 + (y_i - b)^2 - R^2]^2 \quad \to \quad \min \tag{4.2}$$

and the weighted algebraic fit (1.3) takes form

$$\mathcal{F}_3(a, b, R) = \sum_{i=1}^{n} w_i [(x_i - a)^2 + (y_i - b)^2 - R^2]^2 \quad \to \quad \min \tag{4.3}$$

In particular, the GRAF becomes

$$\mathcal{F}_4(a,b,R) = \sum_{i=1}^{n} \frac{[(x_i-a)^2+(y_i-b)^2-R^2]^2}{(x_i-a)^2+(y_i-b)^2} \quad \to \quad \min \tag{4.4}$$

(where the irrelevant constant factor of 4 in the denominator is dropped).

In terms of (2.12), we have

$$\nabla_\Theta P(\bar{\mathbf{x}}_i;\Theta) = -2(\bar{x}_i-a, \bar{y}_i-b, R)^T$$

and $\nabla_{\mathbf{x}} P(\bar{\mathbf{x}}_i;\Theta) = 2(\bar{x}_i-a, \bar{y}_i-b)^T$, hence

$$\|\nabla_{\mathbf{x}} P(\bar{\mathbf{x}}_i;\Theta)\|^2 = 4[(\bar{x}_i-a)^2+(\bar{y}_i-b)^2] = 4R^2$$

Therefore,

$$\mathcal{D}_{\min} = \begin{pmatrix} \sum u_i^2 & \sum u_i v_i & \sum u_i \\ \sum u_i v_i & \sum v_i^2 & \sum v_i \\ \sum u_i & \sum v_i & n \end{pmatrix}^{-1} \tag{4.5}$$

where we denote, for brevity,

$$u_i = \frac{\bar{x}_i - a}{R}, \qquad v_i = \frac{\bar{y}_i - b}{R}$$

The above expression for $\mathcal{D}_{\min}$ was derived earlier in [7, 14].

Now, our Theorem in Section 3 shows that the weighted algebraic fit (4.3) is statistically efficient if and only if the weight function satisfies $w(x,y;a,b,R) = c(a,b,R)/(4R^2)$. Since $c(a,b,R)$ may be an arbitrary function, then the denominator $4R^2$ here is irrelevant. Hence, statistically efficiency is achieved whenever $w(x,y;a,b,R)$ is simply independent of x and y for all (x,y) lying on the circle. In particular, the GRAF (4.4) is statistically efficient because $w(x,y;a,b,R) = [(x-a)^2+(y-b)^2]^{-1} = R^{-2}$. The simple AF (4.2) is also statistically efficient since $w(x,y;a,b,R) = 1$.

We note that the GRAF (4.4) is a highly nonlinear problem, and in its exact form (4.4) is not used in practice. Instead, there are two modifications of GRAF popular among experimenters. One is due to Chernov and Ososkov [8] and Pratt [17]:

$$\mathcal{F}_4'(a,b,R) = R^{-2} \sum_{i=1}^{n} [(x_i-a)^2+(y_i-b)^2-R^2]^2 \quad \to \quad \min \tag{4.6}$$

(it is based on the approximation $(x_i-a)^2+(y_i-b)^2 \approx R^2$), and the other due to Agin [1] and Taubin [20]:

$$\mathcal{F}_4''(a,b,R) = \frac{1}{\sum (x_i-a)^2+(y_i-b)^2} \sum_{i=1}^{n} [(x_i-a)^2+(y_i-b)^2-R^2]^2 \quad \to \quad \min \tag{4.7}$$

(here one simply averages the denominator of (4.4) over $1 \le i \le n$). We refer the reader to [9] for a detailed analysis of these and other circle fitting algorithms, including their numerical implementations.

We have tested experimentally the efficiency of four circle fitting algorithms: the OLSF (1.1), the simple AF (4.2), the Pratt method (4.6), and the Taubin method (4.7). We have generated $n = 20$ points equally spaced on a circle, added an isotropic Gaussian noise with variance σ^2 (according to the Cartesian model), and estimated the efficiency of the estimate of the center by

$$E = \frac{\sigma^2(\mathcal{D}_{11} + \mathcal{D}_{22})}{\langle (\hat{a} - a)^2 + (\hat{b} - b)^2 \rangle} \tag{4.8}$$

Here (a, b) is the true center, $(\hat{a}, \hat{b})$ is its estimate, $\langle \cdots \rangle$ denotes averaging over many random samples, and $\mathcal{D}_{11}$, $\mathcal{D}_{22}$ are the first two diagonal entries of the matrix (4.5). Table 1 shows the efficiency of the above mentioned four algorithms for various values of σ/R. We see that they all perform very well, and indeed are efficient as $\sigma \to 0$. One might notice that the OLSF slightly outperforms the other methods, and the AF is the second best.

σ/R	OLSF	AF	Pratt	Taubin
< 0.01	~ 1	~ 1	~ 1	~ 1
0.01	0.999	0.999	0.999	0.999
0.02	0.999	0.998	0.997	0.997
0.03	0.998	0.996	0.995	0.995
0.05	0.996	0.992	0.987	0.987
0.10	0.985	0.970	0.953	0.953
0.20	0.935	0.900	0.837	0.835
0.30	0.825	0.824	0.701	0.692

Table 1. Efficiency of circle fitting algorithms. Data are sampled along a full circle.

Table 2 shows the efficiency of the same algorithms as the data points are sampled along half a circle, rather than a full circle. Again, the efficiency as $\sigma \to 0$ is clear, but we also make another observation. The AF now consistently falls behind the other methods for all $\sigma/R \le 0.2$, but for $\sigma/R = 0.3$ the others suddenly break down, while the AF keeps afloat.

σ/R	OLSF	AF	Pratt	Taubin
< 0.01	~ 1	~ 1	~ 1	~ 1
0.01	0.999	0.996	0.999	0.999
0.02	0.997	0.983	0.997	0.997
0.03	0.994	0.961	0.992	0.992
0.05	0.984	0.902	0.978	0.978
0.10	0.935	0.720	0.916	0.916
0.20	0.720	0.493	0.703	0.691
0.30	0.122	0.437	0.186	0.141

Table 2. Efficiency of circle fitting algorithms with data sampled along half a circle.

The reason of the above turnaround is that at large noise the data points may occasionally line up along a circular arc of a very large radius. Then the OLSF, Pratt and Taubin dutifully return a large circle whose center lies far away, and such fits blow up the denominator of (4.8), a typical effect of large outliers. On the contrary, the AF is notoriously known for its systematic bias toward smaller circles [8, 11, 17], hence while it is less accurate than other fits for typical random samples, its bias safeguards it from large outliers.

This behavior is even more pronounced when the data are sampled along quarter[1] of a circle (Table 3). We see that the AF is now far worse than the other fits for $\sigma/R < 0.1$ but the others characteristically break down at some point ($\sigma/R = 0.1$).

σ/R	OLSF	AF	Pratt	Taubin
0.01	0.997	0.911	0.997	0.997
0.02	0.977	0.722	0.978	0.978
0.03	0.944	0.555	0.946	0.946
0.05	0.837	0.365	0.843	0.842
0.10	0.155	0.275	0.163	0.158

Table 3. Data are sampled along a quarter of a circle.

It is interesting to test smaller circular arcs, too. Figure 1 shows a color-coded diagram of the efficiency of the OLSF and the AF for arcs from 0^o to 50^o and variable σ (we set $\sigma = ch$, where h is the height of the circular arc, see Fig. 2, and c varies from 0 to 0.5). The efficiency of the Pratt and Taubin is virtually identical to that of the OLSF, so it is not shown here. We see that the OLSF and AF are efficient as $\sigma \to 0$ (both squares in the diagram get white at the bottom), but the AF loses its efficiency at moderate levels

[1]All our algorithms are invariant under simple geometric transformations such as translations, rotations and similarities, hence our experimental results do not depend on the choice of the circle, its size, and the part of the circle the data are sampled from.

of noise ($c > 0.1$), while the OLSF remains accurate up to $c = 0.3$ after which it rather sharply breaks down.

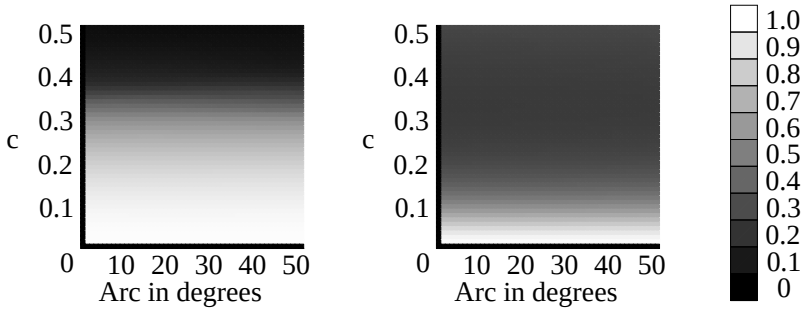

Figure 1: The efficiency of the simple OLSF (left) and the AF (center). The bar on the right explains color codes.

The following analysis sheds more light on the behavior of the circle fitting algorithms. When the curvature of the arc decreases, the center coordinates a, b and the radius R grow to infinity and their estimates become highly unreliable. In that case the circle equation (4.1) can be converted to a more convenient algebraic form

$$A(x^2 + y^2) + Bx + Cy + D = 0 \tag{4.9}$$

with an additional constrain on the parameters: $B^2+C^2-4AD = 1$. This parametrization was used in [17, 11], and analyzed in detail in [9]. We note that the original parameters can be recovered via $a = -B/2A$, $b = -C/2A$, and $R = (2\,|A|)^{-1}$. The new parametrization (4.9) is safe to use for arcs with arbitrary small curvature: the parameters A, B, C, D remain bounded and never develop singularities, see [9]. Even as the curvature vanishes, we simply get $A = 0$, and the equation (4.9) represents a line $Bx + Cy + D = 0$.

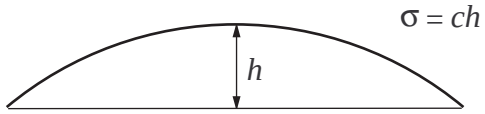

Figure 2: The height of an arc, h, and our formula for σ.

In terms of the new parameters A, B, C, D, the weighted algebraic fit (1.3) takes form

$$\mathcal{F}_3(A, B, C, D) = \sum_{i=1}^{n} w_i[A(x^2 + y^2) + Bx + Cy + D]^2 \quad \to \quad \min \tag{4.10}$$

(under the constraint $B^2+C^2-4AD=1$). Converting the AF (4.2) to the new parameters gives

$$\mathcal{F}_2(A,B,C,D)=\sum_{i=1}^{n} A^{-2}[A(x^2+y^2)+Bx+Cy+D]^2 \quad \to \quad \min \tag{4.11}$$

which corresponds to the weight function $w=1/A^2$. The Pratt method (4.6) turns to

$$\mathcal{F}_4(A,B,C,D)=\sum_{i=1}^{n} [A(x^2+y^2)+Bx+Cy+D]^2 \quad \to \quad \min \tag{4.12}$$

We now see why the AF is unstable and inaccurate for arcs with small curvature: its weight function $w=1/A^2$ develops a singularity (it explodes) in the limit $A \to 0$. Recall that, in our derivation of the statistical efficiency theorem (Section 3), we assumed that the weight function was regular (had bounded derivatives). This assumption is clearly violated by the AF (4.11). On the contrary, the Pratt fit (4.12) uses a safe choice $w=1$ and thus behaves decently on arcs with small curvature, see next.

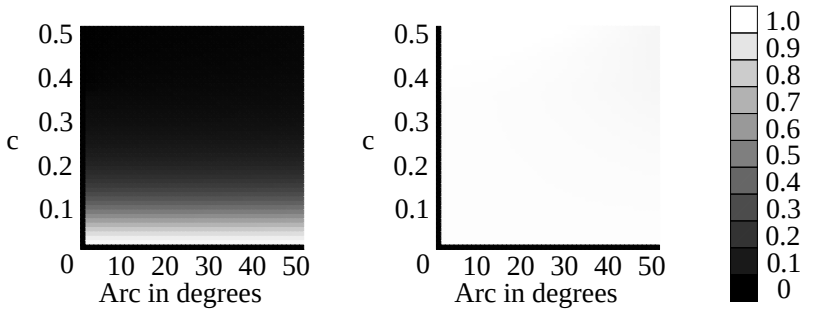

Figure 3: The efficiency of the simple AF (left) and the Pratt method (center). The bar on the right explains color codes.

Figure 3 shows a color-coded diagram of the efficiency of the estimate of the parameter[2] A by the AF (4.11) versus Pratt (4.12) for arcs from 0° to 50° and the noise level $\sigma=ch$, where h is the height of the circular arc and c varies from 0 to 0.5. The efficiency of the OLSF and the Taubin method is visually indistinguishable from that of Pratt (the central square in Fig. 3), so we did not include it here.

We see that the AF performs significantly worse than the Pratt method for all arcs and most of the values of c (i.e., σ). The Pratt's efficiency is close 100%, its lowest point is 89% for 50° arcs and $c=0.5$ (the top right corner of the central square barely gets grey). The AF's efficiency is below 10% for all $c>0.2$ and almost zero for $c>0.4$. Still,

[2]Note that $|A|=1/2R$, hence the estimation of A is equivalent to that of the curvature, an important geometric parameter of the arc.

the AF remains efficient as $\sigma \to 0$ (as the tiny white strip at the bottom of the left square proves), but its efficiency can be only counted on when σ is extremely small.

Our analysis demonstrates that the choice of the weights w_i in the weighted algebraic fit (1.3) should be made according to our theorem in Section 3, and, in addition, one should avoid singularities in the domain of parameters.

Appendix

Here we prove the theorem of linear algebra stated in Section 2. For the sake of clarity, we divide our proof into small lemmas:

Lemma 1. *The matrix B is indeed nonsingular.*

Proof. If $Bz = 0$ for some nonzero vector $z \in \mathbb{R}^k$, then $0 = z^T Bz = \sum_{i=1}^n (v_i^T z)^2/\|u_i\|^2$, hence $v_i^T z = 0$ for all $1 \le i \le k$, a contradiction.

Lemma 2. *If a set of n matrices $A_1, \dots, A_n$ is proper, then rank$(A_i) \le 1$. Furthermore, each A_i is given by $A_i = z_i u_i^T$ for some vector $z_i \in \mathbb{R}^k$, and the vectors $z_1, \dots, z_n$ satisfy $\sum_{i=1}^n z_i v_i^T = -I$ where I is the $k \times k$ identity matrix. The converse is also true.*

Proof. Let vectors $w_1, \dots, w_n$ and r satisfy the requirements (2.10) and (2.11) of the theorem. Consider the orthogonal decomposition $w_i = c_i u_i + w_i^\perp$ where $w_i^\perp$ is perpendicular to u_i, i.e. $u_i^T w_i^\perp = 0$. Then the constraint (2.11) can be rewritten as

$$c_i = -\frac{v_i^T r}{u_i^T u_i} \tag{A.1}$$

for all $i = 1, \dots, n$ and (2.10) takes form

$$\sum_{i=1}^n c_i A_i u_i + \sum_{i=1}^n A_i w_i^\perp = r \tag{A.2}$$

We conclude that $A_i w_i^\perp = 0$ for every vector $w_i^\perp$ orthogonal to u_i, hence A_i has a $(k-1)$-dimensional kernel, so indeed its rank is zero or one. If we denote $z_i = A_i u_i/\|u_i\|^2$, we obtain $A_i = z_i u_i^T$. Combining this with (A.1)-(A.2) gives

$$r = -\sum_{i=1}^n (v_i^T r) z_i = -\left(\sum_{i=1}^n z_i v_i^T\right) r$$

Since this identity holds for any vector $r \in \mathbb{R}^k$, the expression within parentheses is $-I$. The converse is obtained by straightforward calculations. Lemma is proved.

Corollary. *Let $\mathbf{n}_i = u_i/\|u_i\|$. Then $A_i \mathbf{n}_i \mathbf{n}_i^T A_i = A_i A_i^T$ for each i.*

This corollary implies our lemma stated in Section 2. We now continue the proof of the theorem.

Lemma 3. *The sets of proper matrices make a linear variety, in the following sense. Let $A'_1, \ldots, A'_n$ and $A''_1, \ldots, A''_n$ be two proper sets of matrices, then the set $A_1, \ldots, A_n$ defined by $A_i = A'_i + c(A''_i - A'_i)$ is proper for every $c \in \mathbb{R}$.*

Proof. According to the previous lemma, $A'_i = z'_i u_i^T$ and $A''_i = z''_i u_i^T$ for some vectors z'_i, z''_i, $1 \le i \le n$. Therefore, $A_i = z_i u_i^T$ for $z_i = z'_i + c(z''_i - z'_i)$. Lastly,

$$\sum_{i=1}^{n} z_i v_i^T = \sum_{i=1}^{n} z'_i v_i^T + c \sum_{i=1}^{n} z''_i v_i^T - c \sum_{i=1}^{n} z'_i v_i^T = -I$$

Lemma is proved.

Lemma 4. *If a set of n matrices $A_1, \ldots, A_n$ is proper, then $\sum_{i=1}^{n} A_i X_i^T = -I$, where I is the $k \times k$ identity matrix.*

Proof. By using Lemma 2 $\sum_{i=1}^{n} A_i X_i^T = \sum_{i=1}^{n} z_i v_i^T = -I$. Lemma is proved.

Lemma 5. *We have indeed $D \ge B^{-1}$.*

Proof. For each $i = 1, \ldots, n$ consider the $2k \times m$ matrix $Y_i = \begin{pmatrix} A_i \\ X_i \end{pmatrix}$. Using the previous lemma gives

$$\sum_{i=1}^{n} Y_i Y_i^T = \begin{pmatrix} D & -I \\ -I & B \end{pmatrix}$$

By construction, this matrix is positive semidefinite. Hence, the following matrix is also positive semidefinite:

$$\begin{pmatrix} I & B^{-1} \\ 0 & B^{-1} \end{pmatrix} \begin{pmatrix} D & -I \\ -I & B \end{pmatrix} \begin{pmatrix} I & 0 \\ B^{-1} & B^{-1} \end{pmatrix} = \begin{pmatrix} D - B^{-1} & 0 \\ 0 & B^{-1} \end{pmatrix}$$

By Sylvester's theorem, the matrix $D - B^{-1}$ is positive semidefinite.

Lemma 6. *The set of matrices $A_i^{\circ} = -B^{-1} X_i$ is proper, and for this set we have $D = B^{-1}$.*

Proof. Straightforward calculation.

Lemma 7. *If $D = B^{-1}$ for some proper set of matrices $A_1, \ldots, A_n$, then $A_i = A_i^{\circ}$ for all $1 \le i \le n$.*

Proof. Assume that there is a proper set of matrices $A'_1, \ldots, A'_n$, different from $A_1^{\circ}, \ldots, A_n^{\circ}$, for which $D = B^{-1}$. Denote $\delta A_i = A'_i - A_i^{\circ}$. By Lemma 3, the set of matrices $A_i(\gamma) = A_i^{\circ} + \gamma(\delta A_i)$ is proper for every real γ. Consider the variable matrix

$$\begin{aligned} D(\gamma) &= \sum_{i=1}^{n} [A_i(\gamma)][A_i(\gamma)]^T \\ &= \sum_{i=1}^{n} A_i^{\circ}(A_i^{\circ})^T + \gamma \left(\sum_{i=1}^{n} A_i^{\circ}(\delta A_i)^T + \sum_{i=1}^{n} (\delta A_i)(A_i^{\circ})^T \right) + \gamma^2 \sum_{i=1}^{n} (\delta A_i)(\delta A_i)^T \end{aligned}$$

Note that the matrix $R = \sum_{i=1}^{n} A_i^{\circ}(\delta A_i)^T + \sum_{i=1}^{n} (\delta A_i)(A_i^{\circ})^T$ is symmetric. By Lemma 5 we have $D(\gamma) \ge B^{-1}$ for all γ, and by Lemma 6 we have $D(0) = B^{-1}$. It is then easy

to derive that $R = 0$. Next, the matrix $S = \sum_{i=1}^{n} (\delta A_i)(\delta A_i)^T$ is symmetric positive semidefinite. Since we assumed that $D(1) = D(0) = B^{-1}$, it is easy to derive that $S = 0$ as well. Therefore, $\delta A_i = 0$ for every $i = 1, \ldots, n$. The theorem is proved.

References

[1] G.J. Agin, *Fitting Ellipses and General Second-Order Curves*, Carnegi Mellon University, Robotics Institute, Technical Report 81-5, 1981.

[2] S.J. Ahn, W. Rauh, and H.J. Warnecke, Least-squares orthogonal distances fitting of circle, sphere, ellipse, hyperbola, and parabola, *Pattern Recog.*, **34**, 2001, 2283–2303.

[3] D. A. Anderson, The circular structural model, *J. R. Statist. Soc. B*, **27**, 1981, 131–141.

[4] M. Berman and D. Culpin, The statistical behaviour of some least squares estimators of the centre and radius of a circle, *J. R. Statist. Soc. B*, **48**, 1986, 183–196.

[5] M. Berman, Large sample bias in least squares estimators of a circular arc center and its radius, *Computer Vision, Graphics and Image Processing*, **45**, 1989, 126–128.

[6] N. N. Chan, On circular functional relationships, *J. R. Statist. Soc. B*, **27**, 1965, 45–56.

[7] Y. T. Chan and S. M. Thomas, Cramer-Rao Lower Bounds for Estimation of a Circular Arc Center and Its Radius, *Graph. Models Image Proc.* **57**, 1995, 527–532.

[8] N. I. Chernov and G. A. Ososkov, Effective algorithms for circle fitting, *Comp. Phys. Comm.* **33**, 1984, 329–333.

[9] N. Chernov and C. Lesort, Fitting circles and lines by least squares: theory and experiment, preprint, available at http://www.math.uab.edu/cl/cl1

[10] W. Chojnacki, M.J. Brooks, and A. van den Hengel, Rationalising the renormalisation method of Kanatani, *J. Math. Imaging & Vision*, **14**, 2001, 21–38.

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

[12] *Recent advances in total least squares techniques and errors-in-variables modeling*, Ed. by S. van Huffel, SIAM, Philadelphia, 1997.

[13] K. Kanatani, *Statistical Optimization for Geometric Computation: Theory and Practice*, Elsevier Science, Amsterdam, 1996.

[14] K. Kanatani, Cramer-Rao lower bounds for curve fitting, *Graph. Models Image Proc.* **60**, 1998, 93–99.

[15] U.M. Landau, Estimation of a circular arc center and its radius, *Computer Vision, Graphics and Image Processing*, **38** (1987), 317–326.

[16] Y. Leedan and P. Meer, Heteroscedastic regression in computer vision: Problems with bilinear constraint, *Intern. J. Comp. Vision*, **37**, 2000, 127–150.

[17] V. Pratt, Direct least-squares fitting of algebraic surfaces, *Computer Graphics* **21**, 1987, 145–152.

[18] H. Spath, Least-Squares Fitting By Circles, *Computing*, **57**, 1996, 179–185.

[19] H. Spath, Orthogonal least squares fitting by conic sections, in *Recent Advances in Total Least Squares techniques and Errors-in-Variables Modeling*, SIAM, 1997, pp. 259–264.

[20] 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*, **13**, 1991, 1115–1138.

[21] K. Turner, *Computer perception of curved objects using a television camera*, Ph.D. Thesis, Dept. of Machine Intelligence, University of Edinburgh, 1974.