On Bias Correction for Geometric Parameter Estimation in Computer Vision

Takayuki Okatani and Koichiro Deguchi
Tohoku University, Japan
okatani@fractal.is.tohoku.ac.jp

Abstract

Maximum likelihood (ML) estimation is widely used in many computer vision problems involving the estimation of geometric parameters, from conic fitting to bundle adjustment for structure and motion. This paper presents a detailed discussion on the bias of ML estimates derived for these problems. Statistical theory states that although ML estimates attain maximum accuracy in the limit as the sample size goes to infinity, they can have non-negligible bias with small sample sizes. In the case of computer vision problems, the ML optimality holds when regarding variance in observation errors as the sample size. A natural question is how large the bias will be for a given strength of observation errors. To answer this for a general class of problems, we analyze the mechanism of how the bias of ML estimates emerges, and show that the differential geometric properties of geometric constraints used in the problems determines the magnitude of bias. Based on this result, we present a numerical method of computing bias-corrected estimates.

1. Introduction

The problems of estimating geometric parameters from image(s), such as ellipse fitting, the estimation of a fundamental matrix, and the problem of structure from motion (SFM), are all formulated as follows. When observed data $\boldsymbol{X}$ are given, assuming an error model $p(\boldsymbol{X};\boldsymbol{\theta})$ of observation, estimate the parameter $\boldsymbol{\theta}$. Maximum likelihood (ML) estimation is often used to solve this problem, which is to compute a parameter value $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$ maximizing the likelihood

$$l(\boldsymbol{\theta};\boldsymbol{X}) \equiv p(\boldsymbol{X};\boldsymbol{\theta}). \tag{1}$$

ML estimation is commonly used because it is optimal in that the estimate is consistent and efficient for a sample size tending to infinity. Roughly speaking, it is the most accurate of all estimators when the sample size is sufficiently large. In the case of computer vision problems, this optimality holds true when we assume that the observation is repeated multiple times (e.g., multiple images are captured for the same scene) and then regard the number of repetitions as the sample size. In this case, a large sample size means small variance of errors in observation. (Note that this optimality does not hold under a natural interpretation of the sample size whereby the number of observed data is the sample size [6, 4].)

ML estimates are optimal only when the sample size is sufficiently large; they may be inaccurate for small sample sizes. In the latter case, a major issue is that estimates can be biased. For example, consider the problem of estimating unknown mean μ and variance ψ of a normal distribution $N(\mu,\psi)$ from N observations $x_1,\dots,x_N$ that follow the distribution. The ML estimate $\hat{\psi}$ of ψ has expectation $E[\hat{\psi}] = \psi - \psi/N$, where ψ/N is bias. Thus, although bias may be negligible for large N, it causes serious error for small N. In statistics, it is well known that ML estimates can generally be biased for a small sample size; when bias is not negligible, its correction is usually taken into consideration.

In the computer vision community, few studies have focused on the bias of ML estimates. (In contrast, many studies have dealt with the bias of least squares fitting; see [7].) Considering the above basic properties of ML estimates, we believe that this issue should be investigated in greater detail. What if it were possible to improve the accuracy of the widely used bundle adjustment technique for multi-view SFM, which is based on ML estimates?

Recently, Kanatani has dealt with this issue [4]. He analytically derived high-order error terms of an ML estimate and removed them from the estimate to make it more accurate. (Although he called it a "hyperaccurate" method, it can be classified as one of the standard approaches for bias correction used by statisticians.) However, there are several problems with his approach. The first is that it assumes a probabilistic model of observations that is not physically meaningful. The observation model that has a physical background and thus is widely used is such that the observations have a normal distribution in their space. In his approach, the geometric constraint (e.g., $\boldsymbol{x}'^{\top}\boldsymbol{F}\boldsymbol{x} = 0$) is linearized by transforming the observations into another space by nonlinear transformation. Therefore, the transformed observations will have a new distribution in the transformed space, and it is no longer is a normal distribution. Nevertheless, the approach approximates a new distribution with a normal distribution. Moreover, the approach cannot be applied to all problems.

Werman and Keren considered related issues and proposed a Bayesian approach to them [9]. It is reported that the proposed method can fit a circle to noisy data points in an unbiased manner. However, such Bayesian approaches

978-1-4244-3991-1/09/$25.00 ©2009 IEEE

are limited in that the prior distribution of the observations (more rigorously, the latent variables) is required. For example, in the estimation of a fundamental matrix, it is required to know in advance how 3D points distribute in space.

In this paper, we present a detailed description of the bias of ML estimates that emerges in computer vision problems. Unlike [4], we deal with the case where the observation noises have a normal distribution in the observation space. Our study makes three main contributions. First, for a particular class of problems, which includes ellipse fitting and the estimation of a fundamental matrix, we show that the curvature of the hypersurface constraining the true values of the observations has a direct relation to the bias of the ML estimates. Second, based on this result, for the same class of problems, we propose a method of numerically computing a bias-corrected estimate. Third, our study provides some linkage to researches in statistics that has dealt with this issue.

2. Maximum likelihood estimates for computer vision problems

2.1. Estimation of geometric parameters

First, we describe the formulation of the problems that are being considered. Denoting an observation by k-vector $\boldsymbol{x}_i$, we assume that n observations, $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$, are given. Assuming an additive Gaussian error for the observation process, we assume $\boldsymbol{x}_i$ to have a normal distribution $N_k(\boldsymbol{\eta}_i, \sigma^2\boldsymbol{I})$, where $\boldsymbol{\eta}_i$ is the true value of $\boldsymbol{x}_i$. The true value $\boldsymbol{\eta}_i$ is constrained by a geometric constraint given as

$$\boldsymbol{f}(\boldsymbol{\eta}_i, \boldsymbol{\theta}) = \boldsymbol{0}, \quad i = 1, \ldots, n, \tag{2}$$

where $\boldsymbol{f}$ is a function representing the geometry of interest and $\boldsymbol{\theta}$ is an unknown parameter we want to estimate. For example, in ellipse fitting, $\boldsymbol{f}$ is given as $f([x, y]^\top, [a, b, x_0, y_0]^\top) = (x - x_0)^2/a^2 + (y - y_0)^2/b^2 - 1$. In the estimation of a fundamental matrix, $\boldsymbol{f}$ is given as $f([x, y, x', y']^\top, \boldsymbol{F}) = [x, y, 1]\boldsymbol{F}[x', y', 1]^\top$.

The geometric constraint on $\boldsymbol{\eta}_i$ can be represented in an explicit manner by incorporating a variable $\boldsymbol{\xi}_i$ for each observation as

$$\boldsymbol{\eta}_i = \boldsymbol{g}(\boldsymbol{\xi}_i, \boldsymbol{\theta}), \quad i = 1, \ldots, n. \tag{3}$$

In ellipse fitting, using a scalar variable ξ_i, the true ellipse point is given as $[\eta_{i1}, \eta_{i2}]^\top = [a\cos\xi_i + x_0, b\sin\xi_i + y_0]^\top$. In the estimation of a fundamental matrix, a similar expression can be obtained.

In both equations (2) and (3), assuming a fixed $\boldsymbol{\theta}$, the equation constrains $\boldsymbol{\eta}_i$ in k-dimensional space; the set of constrained $\boldsymbol{\eta}_i$ forms a $(k - l)$-dimensional submanifold in the space, where l is the dimensionality of $\boldsymbol{f}$ or $\boldsymbol{\xi}_i$. In both cases, the ML estimates of the parameters are obtained by minimizing the negative log-likelihood

$$L \propto \sum_{i=1}^{n} (\boldsymbol{x}_i - \boldsymbol{\eta}_i)^2. \tag{4}$$

In the case of (2), L is minimized with respect to $\boldsymbol{\theta}$ and $\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_n$ subject to constraint (2). In the case of (3), L is minimized simply with respect to $\boldsymbol{\theta}$ and $\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_n$. Note that in both cases, a variable exists for each observation $\boldsymbol{x}_i$, i.e., $\boldsymbol{\eta}_i$ in (2) and $\boldsymbol{\xi}_i$ in (3).

2.2. Two asymptotics

As mentioned earlier, ML estimation is proven to be optimal in the asymptotic sense that the sample size increases to infinity [2, 8]. Two types of asymptotics are possible for the problems formulated above. The first is such that the number of observations n increases to infinity, i.e., $n \to \infty$. The second is such that the variance of observation errors decreases to 0, i.e., $\sigma^2 \to 0$. In the former, the number of observations n is the sample size. In the latter, assuming each point $\boldsymbol{\eta}_i$ was observed repeatedly, we regard the number of repetitions m of the observation as the sample size. When an identical $\boldsymbol{\eta}_i$ is observed m times, we may regard the mean of the m observations as a new observation $\boldsymbol{x}_i$. Then, the variance σ_m^2 of the errors of the new observation is given by $\sigma_m^2 = \sigma_1^2/m$, where σ_1^2 is the original variance for a single observation. Thus, $m \to \infty$ is equivalent to $\sigma^2 \to 0$ [4, 6].

These two asymptotics lead to different conclusions. It is easy to show that in the latter case of $\sigma^2 \to 0$ (or equivalently, $m \to \infty$), the above optimality of ML estimates holds [2]. However, in the former case of $n \to \infty$, it is known that an ML estimate is generally not optimal [6, 4]. This is because the number of unknowns increases with the number of observations n, since there exists a latent variable per observation, as described above. Problems having this structure are referred to as Neyman-Scott problems [5]. It remains unsolved in statistics how to obtain an optimal estimate when $n \to \infty$.

In this paper, we consider only the former asymptotics that is easier to deal with, i.e. the case of $\sigma^2 \to 0$ (or equivalently, $m \to \infty$), in which we discuss the bias of a ML estimate. However, the bias itself may be related to the issue with the Neyman-Scott structure; as described below, the presence of latent variables is responsible for the emergence of the bias. Thus, the present study could also provide some insights into the issue with the case of $n \to \infty$.

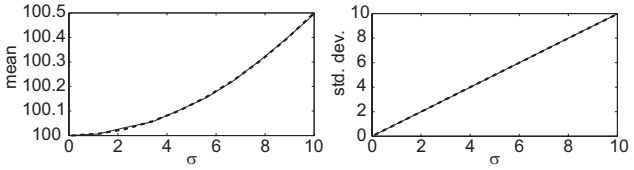
3. Bias of maximum likelihood estimates

3.1. Example: Bias of circle fitting

To illustrate by example the bias of an ML estimate and then discuss its properties, here, we consider a problem of fitting a circle to a set of points [1, 8].

Example 1. Consider a circle with radius r centered at the origin of the xy coordinate system: $x^2 + y^2 = r^2$. Assume that for $i = 1, \ldots, n$, an observation $\boldsymbol{x}_i = [x_i, y_i]^\top$ of a point $\boldsymbol{\eta}_i$ on the circle distributes according to $N(\boldsymbol{\eta}_i, \sigma^2\boldsymbol{I})$. Then, we want to estimate the radius r from the n observations $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$.

The observation $\boldsymbol{x}_i$ $(i = 1, \ldots, n)$ is explicitly expressed



(a) (b)

Figure 1. Results of ML estimation for the radius $r(= 100)$ of a circle. (a) The mean of the estimates over 1,000,000 trials. (b) The standard deviation. The horizontal axis indicates the standard deviation σ of the observation error. The solid lines indicate the sample mean and sample variance, and the chained lines represent their predictions by Efron's curvature.

as follows:

$$\begin{cases} x_i = r\cos\xi_i + \varepsilon_i, \\ y_i = r\sin\xi_i + \varepsilon'_i, \end{cases} \tag{5}$$

where ε_i and ε'_i are i.i.d. random variables generated according to $N(0, \sigma^2)$. The ML estimates of r and $\xi_1, \ldots, \xi_n$ are calculated by minimizing the negative log-likelihood: $\sum_i (x_i - r\cos\xi_i)^2 + (y_i - r\sin\xi_i)^2$. First, the ML estimate of ξ_i is derived as $\hat{\xi}_i = \arctan(y_i/x_i)$. Substituting this into the log-likelihood, we have a function only of r. The minimization of the function yields the ML estimate of r: $\hat{r} = \sum_i \sqrt{x_i^2 + y_i^2}/n$.

Suppose we estimate r from the minimum number of observations, i.e., a single point ($n = 1$). Fig.1 shows the result of experiments performed using synthetically generated observations. Fig.1(a) shows the mean of estimates $\hat{r}$ over 1,000,000 trials when σ varies from 0.1 to 10. We can see from the plot that the bias (i.e., $\hat{r} - 100$) increases with σ. This demonstrates the aforementioned asymptotic properties of ML estimates.

3.2. Bias and number of observations

In the above experiment, r is estimated from a single observation. This result applies to the case in which n observations are used to estimate r since the expectation of the ML estimate $\hat{r}$ is independent of the number of observations n. Let $\hat{r}_n$ be the ML estimate using n observations and $\hat{r}_1$ be that using a single observation. Then, it can be shown that $E[\hat{r}_n] = E[\sum_i \sqrt{x_i^2 + y_i^2}/n] = nE[\hat{r}_1]/n = E[\hat{r}_1]$. This means that even if an ML estimate is obtained using an infinite number of observations (i.e., $n \to \infty$), the resulting estimate will have the same magnitude of bias. Considering the mechanism of how the bias emerges, which is shown in the next section, we conjecture the following:

Conjecture 1. The bias of ML estimates is not directly dependent on the number of observations n.

Although it may be independent of the bias, increasing the number of observations n improves the accuracy of the estimate; more rigorously, the variance of the estimate will monotonically decrease with n. For example, in the above problem of circle fitting, because of the independence among observations the variance $V[\hat{r}_n]$ of the estimate from n observations is $1/n$ times the variance $V[\hat{r}_1]$ of that from a single observation.

Now, we consider whether it is necessary to correct the bias of an estimate. Obviously, the answer depends on the magnitude of the bias. However, the absolute magnitude of the bias is not so important; it should be compared with the variance (or the standard deviation) of the estimate. In other words, if the bias is much smaller than the standard deviation of the estimate, it is not necessary to correct it. However, if the bias is comparable to the standard deviation, it should be corrected. From the above observation that the variance of an estimate depends on the number of observations n while its bias does not, we can conjecture the following:

Conjecture 2. As the number of observations n increases, so does the necessity for correcting bias.

Let us take the above case of circle fitting as an example. Fig.1(b) shows the standard deviation of the estimate $\hat{r}$ versus σ of the observation errors for $n = 1$; its standard deviation increases with σ. Then, the necessity for correcting the bias can be evaluated for each value of σ by comparing the bias shown in (a) with the standard deviation shown in (b). Comparing the two, the bias is approximately three digits smaller than the standard deviation, and therefore we can determine that it would be meaningless to correct it. When $\sigma = 10$, although the difference becomes smaller, it is still greater than a single digit, and therefore, bias correction remains unnecessary. However, as mentioned above, the estimate has a smaller standard deviation for a larger number of observations. For example, for $n = 100$, the vertical axis of Fig.1(b) is scaled down by $1/10$, and for $n = 10000$, it is scaled down by $1/100$. In the latter case, when $\sigma = 2$, the bias and the standard deviation of the estimate are comparable (approximately 0.02); therefore bias correction becomes necessary.

3.3. Profile likelihood

In some problems, there exist multiple unknown parameters that are inseparable in their estimation. It can be shown that for these problems, ML estimates can always be biased [2]. When there exist multiple parameters, their estimates are calculated as follows. For the sake of simplicity, consider the case of two parameters where the log-likelihood is given by $L(\theta_1, \theta_2)$; $\boldsymbol{\theta} = (\theta_1, \theta_2)$ are unknown parameters. The ML estimate of (θ_1, θ_2) is given by the parameter values maximizing L. In order to calculate them, unless L is separable as $L(\theta_1, \theta_2) = L(\theta_1)L(\theta_2)$, we first consider, say, θ_1 as a constant and maximize L with respect to θ_2. The resulting maximizer $\hat{\theta}_2$ can be considered to be a function of θ_1 as $\hat{\theta}_2(\theta_1)$. Then, plugging this in L, we have $\tilde{L}(\theta) = L(\theta_1, \hat{\theta}_2(\theta_1))$. Maximizing $\tilde{L}(\theta)$, we finally obtain $\hat{\theta}_1$, which is the ML estimate of θ_1; the ML estimate of θ_2 is calculated as $\hat{\theta}_2 = \hat{\theta}_2(\hat{\theta}_1)$. The function $\tilde{L}$ defined above is called a profile likelihood function [8].

It can be shown [2] that if the ML estimate of a parameter is calculated independently of other parameters, the estimate is unbiased. Otherwise, the ML estimate can be bi-

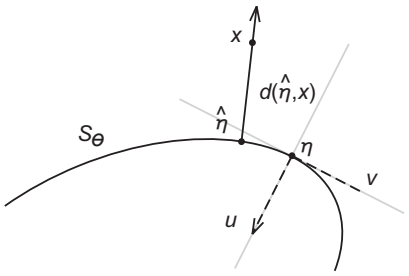


Figure 2. Local geometry of hypersurface S_θ. $\boldsymbol{\eta}$ on S_θ is true value of observation $\boldsymbol{x}$ that follows $N(\boldsymbol{\eta}, \sigma^2\boldsymbol{I})$, and $\hat{\boldsymbol{\eta}}$ is the nearest point on S_θ to $\boldsymbol{x}$. d is a signed distance from $\hat{\boldsymbol{\eta}}$ to $\boldsymbol{x}$.

ased. Mathematically, if $\partial L/\partial\theta_1$ is independent of θ_2, then the ML estimate of θ_1 is unbiased; otherwise, it can be biased. This can be confirmed for the problem mentioned in Section 1, which is to estimate the mean μ and variance ψ of a normal distribution from $x_1, \ldots, x_N$, each of which follows $N(\mu, \psi)$. It is easy to see that $\partial L/\partial\mu$ is independent of ψ, whereas $\partial L/\partial\psi$ depends on μ. In fact, $\hat{\mu} = \sum_i x_i/N$ is unbiased and $\hat{\psi} = \sum_i (x_i - \hat{\mu})^2/N$ is biased.

A major characteristic of computer vision problems formulated in Section 2.1 is that there exists the same number of latent variables as the observations. These latent variables (i.e., $\boldsymbol{\eta}_i$'s in Eq.(2) and $\boldsymbol{\xi}_i$'s in Eq.(3)) play the same role as θ_2 in the above discussion. Thus, ML estimates will generally be biased for the computer vision problems.

4. Bias and geometric constraints

In the last section, we stated that the magnitude of the bias depends on the variance of observation errors (and not directly on the number of observations). In this section, we show that the magnitude of the bias also depends on the geometric constraint. Specifically, we show a relation between the bias and a local differential-geometric property of the geometric constraint. It is based on a study by Efron [1].

4.1. Efron's curvature

Suppose a scalar function of a k-vector $\boldsymbol{\eta}$, $\theta = t(\boldsymbol{\eta})$. Assuming a constant value for θ, the function gives a hypersurface in k-dimensional space:

$$S_\theta = \{\boldsymbol{\eta} \,|\, t(\boldsymbol{\eta}) = \theta\}. \tag{6}$$

Consider a point $\boldsymbol{\eta}$ on S_θ. Assume that a k-vector $\boldsymbol{x}$ has a normal distribution with mean $\boldsymbol{\eta}$ and covariance $\boldsymbol{I}$. Let $\hat{\boldsymbol{\eta}}$ be the nearest point on S_θ to $\boldsymbol{x}$, as shown in Fig.2. Then, $\hat{\boldsymbol{\eta}}$ is given by

$$\hat{\boldsymbol{\eta}} = \underset{\boldsymbol{\eta}' \in S_\theta}{\operatorname{argmin}} |\boldsymbol{x} - \boldsymbol{\eta}'|^2. \tag{7}$$

Here, we assume that an orientation can be given to the hypersurface S_θ so that a signed distance $d(\hat{\boldsymbol{\eta}}, \boldsymbol{x})$ from $\hat{\boldsymbol{\eta}}$ to $\boldsymbol{x}$ can be defined. (It does not matter which sign convention is chosen but it needs to be consistently defined.) When S_θ and the original point $\boldsymbol{\eta}$ are fixed, $d(\hat{\boldsymbol{\eta}}, \boldsymbol{x})$ is a function only of $\boldsymbol{x}$. (Note the relation of d to the profile likelihood described in Section 3.3; specifically, $\sum d^2$ corresponds to the profile likelihood described in Section 3.3.)

To describe Efron's theorem, we introduce a matrix $\boldsymbol{D}$ that is associated with the curvature of S_θ at $\boldsymbol{\eta}$. It is assumed here that a local coordinate system u-$\boldsymbol{v}$ can be defined such that the u-axis is normal to S_θ at $\boldsymbol{\eta}$ and the $\boldsymbol{v}$-axes are within the tangent space to S_θ at $\boldsymbol{\eta}$. There are two choices for the positive directions of u, and we choose the opposite direction to the positive direction of the signed distance $d(\hat{\boldsymbol{\eta}}, \boldsymbol{x})$ that has already been chosen, as shown in Fig.2. Then, $\boldsymbol{D}$, a $(k-1) \times (k-1)$ symmetric matrix, is defined such that S_θ is locally expressed as $u = \boldsymbol{v}^\top \boldsymbol{D}\boldsymbol{v}$ in the u-$\boldsymbol{v}$ coordinates. As in [1], we will call $\boldsymbol{D}$ the curvature matrix in what follows.

The theorem below is asymptotic in that it becomes increasingly accurate as the curvature approaches 0, i.e., $\boldsymbol{D} \to \boldsymbol{O}$, or S_θ approaches a locally flat surface. Instead of $\boldsymbol{D} \to \boldsymbol{O}$, we can assume N observations and take the limit $N \to \infty$. Specifically, we assume that $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_N$ are independent and identically distributed according to $N(\boldsymbol{\eta}, \boldsymbol{I})$. Then, let $\boldsymbol{x}$ be the mean of $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_N$; it is seen that $\boldsymbol{x}$ follows $N(\boldsymbol{\eta}, \boldsymbol{I}/N)$. We now scale the space by $\sqrt{N}$ and define $\boldsymbol{x}' = \sqrt{N}\boldsymbol{x}$ so that $\boldsymbol{x}'$ follows $N(\boldsymbol{\eta}', \boldsymbol{I})$, where $\boldsymbol{\eta}' = \sqrt{N}\boldsymbol{\eta}$. This scaling also results in $\boldsymbol{D}' = \boldsymbol{D}/\sqrt{N}$ since $u' = \boldsymbol{v}'^\top \boldsymbol{D}'\boldsymbol{v}'$ (or equivalently, $\sqrt{N}u = (\sqrt{N}\boldsymbol{v})^\top \boldsymbol{D}'(\sqrt{N}\boldsymbol{v})$) should be equivalent to $u = \boldsymbol{v}^\top \boldsymbol{D}\boldsymbol{v}$. Thus, $N \to \infty$ is equivalent to $\boldsymbol{D} \to \boldsymbol{O}$.

The following is shown in the above asymptotic sense [1]:

Theorem 1. When $\boldsymbol{x}$ has a normal distribution $N(\boldsymbol{\eta}, \boldsymbol{I})$, the signed distance $d(\hat{\boldsymbol{\eta}}, \boldsymbol{x})$ is asymptotically normal with the first four cumulants

$$[\mathrm{tr}(\boldsymbol{D}), (1 - \mathrm{tr}(\boldsymbol{D}^2))^2, 0, 0] \tag{8}$$

to $O(N^{-1})$; the errors of (8) are $O(N^{-3/2})$.

Note that $\mathrm{tr}(\boldsymbol{D})$ is $(k-1)/2$ times the mean curvature of S_θ. Even if S_θ is given only in an implicit form, $\mathrm{tr}(\boldsymbol{D})$ is calculated by using a formula for the mean curvature for an implicit hypersurface [3].

In [1], using this result, confidence intervals for ML estimates having higher-order accuracies than first-order standard intervals are analytically derived. For any function $\theta = t(\boldsymbol{\eta})$, the ML estimate $\hat{\theta}$ of θ is given by $\hat{\theta} = t(\hat{\boldsymbol{\eta}})$, where $\hat{\boldsymbol{\eta}}$ is the ML estimate of $\boldsymbol{\eta}$; then, the analytical confidence interval for $\hat{\theta}$ is calculated directly from the above result. The original objective of [1] was to show a close agreement with bootstrap confidence intervals and the analytical intervals.

In the above asymptotic analysis, the observation $\boldsymbol{x}$ is assumed to have a constant covariance $\boldsymbol{I}$. However, in our discussion, it is more convenient to assume $\sigma^2\boldsymbol{I}$. Then, as discussed earlier, $N \to \infty$ transforms to $\sigma^2 \to 0$. In this asymptotic discussion, Theorem 1 can be restated as follows:

Corollary 1. When $\boldsymbol{x}$ has a normal distribution $N(\boldsymbol{\eta}, \sigma^2\boldsymbol{I})$, the signed distance $d(\hat{\boldsymbol{\eta}}, \boldsymbol{x})$ is asymptotically normal as

$\sigma^2 \to 0$ with first four cumulants:

$$[\sigma^2 \mathrm{tr}(\boldsymbol{D}), \sigma^2(\boldsymbol{I} - \sigma^2 \mathrm{tr}(\boldsymbol{D}^2))^2, 0, 0]. \quad (9)$$

4.2. Bias for a particular class of problems

The equation $\theta = t(\boldsymbol{\eta})$ considered in the above analysis can be thought of as a particular case of the implicit geometric constraint (2). Thus, when Eq.(2) is represented in the form $\theta = t(\boldsymbol{\eta})$, or equivalently, when the parameter to be estimated is a scalar and it can be estimated from only a single observation, the bias of the estimate is approximately given by $\sigma^2 \mathrm{tr}(\boldsymbol{D})$. Thus, in this class of problems, the bias is proportional to the variance σ^2 of the observation errors as well as the mean curvature of the hypersurface (or rigorously the trace $\mathrm{tr}(\boldsymbol{D})$ of the curvature matrix).

The circle fitting problem discussed earlier belongs to this class of problems. Denoting the coordinates of a point on the circle by $\boldsymbol{\eta} = [\eta_1, \eta_2]^\top$ and defining $t(\boldsymbol{\eta}) = \sqrt{\eta_1^2 + \eta_2^2}$, the circle can be expressed as $\theta(= r) = t(\boldsymbol{\eta})$. When estimating from a single observation, the observation $\boldsymbol{x}$ gives the ML estimate of the true coordinates $\boldsymbol{\eta}$ of the circle point. Because of the invariance of ML estimates to parameter transformation (that is, if $\hat{\alpha}$ is the ML estimate of α and if h is any function of α, then the ML estimate of $\beta = h(\alpha)$ is simply given as $\hat{\beta} = h(\hat{\alpha})$), the ML estimate $\hat{r}$ of the radius r is given as $\hat{r} = t(\boldsymbol{x})$. Then, the signed distance d is given by $\hat{r} - r$. Since the hypersurface S_θ is simply a circle with radius r and its curvature is $1/(2r)$, Corollary 1 states that $d = \hat{r} - r$ approximately follows $N(\sigma^2/(2r), \sigma^2(1 - (\sigma/(2r))^2)^2)$; the bias of $\hat{r}$ is approximately $\sigma^2/(2r)$. In Fig.1(a), this analytically calculated bias is plotted along with the bias computed by simulation; we can see that there is close agreement between them.

4.3. Mechanism of the emergence of biases

The analytical calculation of biases given above is possible only for a particular class of problems (i.e., $\theta = t(\boldsymbol{\eta})$). However, Theorem 1 itself (and also Corollary 1) holds true for general hypersurfaces.

Corollary 2. Corollary 1 holds for a more general hypersurface $S_\theta = \{\boldsymbol{\eta} \,|\, f(\boldsymbol{\eta}, \boldsymbol{\theta}) = 0\}$.

Its proof is omitted here. Using this, we can consider the behavior of the bias for a wider class of problems in which constraint (2) is given in the form $f(\boldsymbol{\eta}, \boldsymbol{\theta}) = 0$. Note that although (2) in the most general form is given by a vector function $\boldsymbol{f}(\boldsymbol{\eta}, \boldsymbol{\theta})$, we require it to be a scalar function because of the necessity that S_θ should be a hypersurface. Many problems, such as ellipse fitting, the estimation of a fundamental matrix, etc belong to this class.

In this case, the ML estimate $\hat{\boldsymbol{\theta}}$ of the parameter $\boldsymbol{\theta}$ is calculated by minimizing $L(\boldsymbol{\theta}) = \sum_i (\boldsymbol{x}_i - \boldsymbol{\eta}_i)^2$ subject to $f(\boldsymbol{\eta}_i, \boldsymbol{\theta}) = 0$. Using the notation used above, $\hat{\boldsymbol{\theta}}$ can be expressed as follows:

$$\hat{\boldsymbol{\theta}} = \operatorname*{argmin}_{\boldsymbol{\theta}} L(\boldsymbol{\theta}) = \operatorname*{argmin}_{\boldsymbol{\theta}} \sum_{i=1}^{n} d(\hat{\boldsymbol{\eta}}(\boldsymbol{x}_i; \boldsymbol{\theta}), \boldsymbol{x}_i)^2, \quad (10)$$

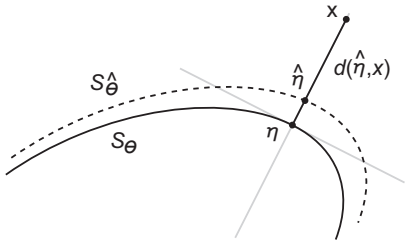


Figure 3. The estimated hypersurface $S_{\hat{\theta}}$ is "pulled out" from the true hypersurface S_θ in the direction that S_θ is convex.

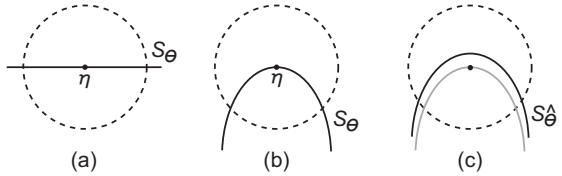


Figure 4. Relation between the local shape of a hypersurface and the distribution of observations. Depending on the curvature of the hypersurface, the distance from the hypersurface to the observations can differ in an average sense for the two regions divided by the hypersurface.

where $\hat{\boldsymbol{\eta}}(\boldsymbol{x}_i; \boldsymbol{\theta})$ is the nearest point on S_θ to $\boldsymbol{x}_i$ and d is the signed distance from $\hat{\boldsymbol{\eta}}$ to $\boldsymbol{x}_i$, as defined earlier. Then, we can consider $\hat{\boldsymbol{\theta}}$ to be the minimizer of the sum of the squared signed distance d^2.

A problem with this minimization is that when it is considered to be a probabilistic variable, the signed distance d has an asymmetric distribution with a nonzero mean, as stated in Corollary 1. (Another problem is that the variance of d is different for each observation, although this has a less significant impact.) As described above, the estimate $\hat{\boldsymbol{\theta}}$ is determined such that d^2 is minimized or, ideally, is equated to 0. Thus, there arises the following disagreement: in reality, d has an asymmetric nature for a true hypersurface S_θ, whereas the estimate $S_{\hat{\theta}}$ of S_θ is determined such that the square of the signed distance d to $S_{\hat{\theta}}$ approaches 0. Because of this disagreement, the estimated hypersurface $S_{\hat{\theta}}$ typically deviates from the true one S_θ. This is illustrated in Fig.3; $S_{\hat{\theta}}$ tends to be "pulled out" in the direction that the hypersurface is convex. (More rigorously, it is pulled out in the direction having the opposite sign to the sign of $\mathrm{tr}(\boldsymbol{D})$.)

Fig.4 illustrates the mechanism of how this deviation emerges. Each circle represents the distribution of an observation $\boldsymbol{x}$, which is given as $N(\boldsymbol{\eta}, \sigma^2 \boldsymbol{I})$. (In reality, it is a k-dimensional hypersphere, where k is the dimensionality of the observation vectors.) The true point $\boldsymbol{\eta}$ lies on the center of the circle, and the true hypersurface S_θ passes through this center.

Now, consider the two subregions of the circle divided by the hypersurface. The observation $\boldsymbol{x}$ is randomly generated equally in all directions from the circle center. Thus, each subregion corresponds to the sign of the signed distance d. More specifically, d is determined for each $\boldsymbol{x}$ in such a way that its sign is determined by which subregion $\boldsymbol{x}$ falls in and its absolute value $|d|$ equals the distance from S_θ

to $\boldsymbol{x}$. If the hypersurface is flat, as shown in Fig.4(a), $|d|$ is the same for each subregion in an average sense. However, if the hypersurface is curved, as shown in Fig.4(b), $|d|$ differs for each subregion in an average sense. In this case, d will have an asymmetric distribution. Moreover, the probability that $\boldsymbol{x}$ falls in each subregion is proportional to its volume, which reinforces this asymmetric nature. Hence, the distribution of d has a nonzero mean that is proportional to the curvature.

As described above, the estimate $\hat{\boldsymbol{\theta}}$ is determined by minimizing the sum of d^2. Geometrically speaking, this is equivalent to searching for the hypersurface that cancels the asymmetric nature of d. The resulting hypersurface is such that the volumes of the two subregions are more balanced, as shown in Fig.3(c). This illustrates the mechanism of why the estimated hypersurface $S_{\hat{\boldsymbol{\theta}}}$ deviates from the true one $S_{\boldsymbol{\theta}}$.

In Section 3.3, we have described that when the estimation of the interest parameter $\boldsymbol{\theta}$ depends on that of the latent variable $\boldsymbol{\eta}$, the resulting ML estimate $\hat{\boldsymbol{\theta}}$ can be biased. The above discussion geometrically illustrates this in greater detail; the estimation of $\boldsymbol{\eta}$ at a curved hypersurface results in the asymmetric nature of the signed distance d, which makes the final estimate $\hat{\boldsymbol{\theta}}$ biased.

4.4. Remarks and example

An important implication of the above analysis is that when the hypersurface has curvatures of different magnitudes for different parts, the deviation of the estimated hypersurface can differ for each part; it could be large for a local part having large curvature. The interest parameter $\boldsymbol{\theta}$ is usually estimated from many data points, and thus, how each local part affects the final estimate of $\boldsymbol{\theta}$ may not be predictable. Nevertheless, it is possible to derive a few useful observations from this discussion. For example, even for the same problem, it is possible that the shape of the hypersurface changes drastically depending on the parameter values; therefore, local parts having large curvatures emerge and the final estimate is biased. In other words, the magnitude of the bias depends on the true parameter values we want to estimate, e.g., the shape of ellipses and the layout of stereo cameras. Even when it is confirmed that the ML estimate is unbiased for a particular configuration, it does not necessarily mean that the estimate is always guaranteed to be similarly unbiased for other configurations.

The problem of ellipse fitting is considered as an example.

Example 2. *For an ellipse* $(x-x_0)^2/a^2 + (y-y_0)^2/b^2 = 1$, *estimate the parameter* $\boldsymbol{\theta} = [a, b, x_0, y_0]$ *from* $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$.

Using a formula given in [3], the curvature κ at a point (x, y) on the ellipse is given as follows:

$$\kappa = \frac{a^2 b^2 (b^2 x^2 + a^2 y^2)}{(b^4 x^2 + a^4 y^2)^{3/2}}. \tag{11}$$

In this case, $\kappa = \mathrm{tr}(\boldsymbol{D})$. As described earlier, if $\mathrm{tr}(\boldsymbol{D})$ is large, the estimate will be more biased. If the major and minor axes of the ellipse are similar in length (i.e., $a \sim b$), the resulting bias will be similar to the case of circle fitting. Thus, we considered an ellipse with $a = 5$ and $b = 1$; it is assumed to be centered at the origin $(x_0, y_0) = (0, 0)$. Then, we computed the ML estimates of these parameters from noisy observations. Fig.5 shows the result. The data points (observations) are synthesized by choosing 300 ellipse points and adding Gaussian noises with $\sigma = 0.2$ to the x and y coordinates of each point. Fig.5(a) shows an example set of these points. The estimates are obtained by using the Levenberg-Marquardt algorithm. Figs.5(b) and (c) show magnified plots of the estimated ellipses for 30 trials; (b) shows a portion having the largest curvature and (c), a portion having the smallest curvature. The estimated ellipses are indicated by thin lines and the true ellipse, by a thick line. It is clearly observed that for the portion having a large curvature, the estimated ellipses (thin lines) significantly deviate from the true ellipse (thick lines) toward the outer side. On the other hand, for the portion having a small curvature, the estimated ellipses appear to have much smaller biases. These observations are supported by the histograms of the estimates of a and b shown in Fig.6.

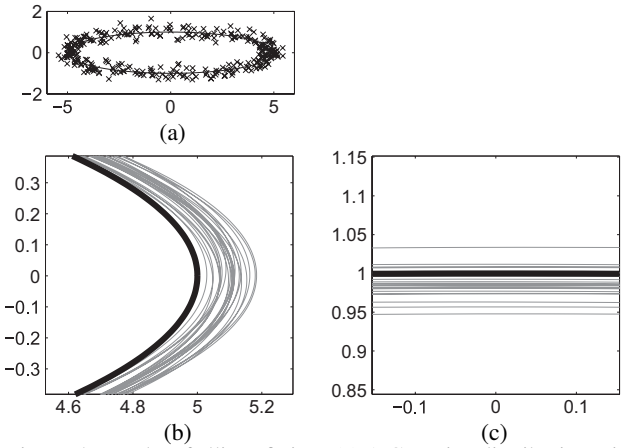


Figure 5. Results of ellipse fitting. (a) A Gaussian distribution with $\sigma = 0.2$ is assumed for the observation errors. (b) Estimated ellipses (thin lines) and the true ellipse (thick line) around the largest curvature. It is seen that the estimated ellipses are significantly biased. (c) Those around the smallest curvature.

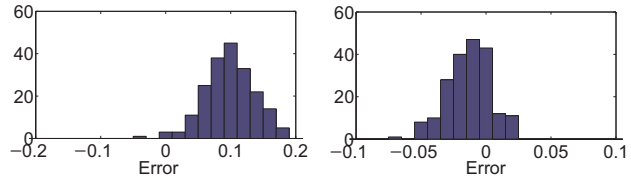


Figure 6. Histograms of estimation errors for the axis parameters a and b. Left: $\hat{a} - a$. Right: $\hat{b} - b$.

5. Correcting biases

In this section, we consider numerical methods for dealing with biases.

5.1. A method for bias correction

In statistics, two standard methods are used for correcting the bias of ML estimates. The first method is to analytically derive the bias and remove it from the ML estimate. (The method described in [4] may belong to this category.) The second is to use resampling-based methods such as bootstrap and jackknife. However, these methods are difficult or impossible to adopt for computer vision problems considered here. In many cases, analytical derivation of the bias is usually impossible because of its complexity. In fact, ML estimates themselves are usually obtained by only numerical computation, where the negative log-likelihood is numerically minimized, except for very simple problems such as Example 1. The resampling-based method also cannot be used, since it can only cope with a bias emerging due to a small number of observations; it reduces the bias by increasing the effective sample size by resampling the observed data. In our case, each datum $\boldsymbol{x}_i$ is observed only once and it cannot be resampled.

Thus, a different method is required. Considering the fact that ML estimates are obtained only by numerical minimization, the only promising strategy would be to 1) somehow modify the cost to be minimized so as to remove the bias and 2) obtain a bias-corrected estimate by numerically minimizing the modified cost.

As shown in Eq.(10), the ML estimate is obtained by minimizing the sum of squares of the signed distances for all observations. Then, as described in the last section, the bias of ML estimates emerges because the signed distance d has an asymmetric distribution $N(\sigma^2\mathrm{tr}(\boldsymbol{D}), \sigma^2(\boldsymbol{I} - \sigma^2\mathrm{tr}(\boldsymbol{D}^2))^2)$. Therefore, we propose the use of the normalized signed distance e defined below instead of d:

$$e \equiv \frac{d - \sigma^2\mathrm{tr}(\boldsymbol{D})}{\sigma(\boldsymbol{I} - \sigma^2\mathrm{tr}(\boldsymbol{D}^2))}. \tag{12}$$

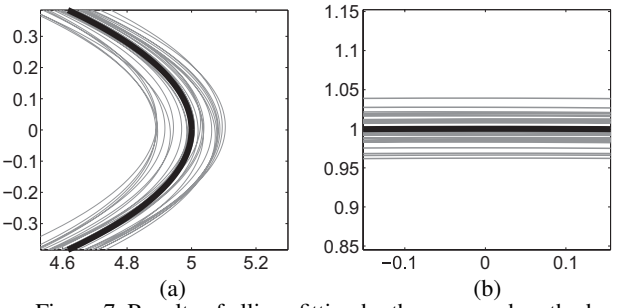
Since e approximately follows $N(0, 1)$, the minimization of the sum of squares of the e values is expected to yield a bias-corrected estimate. This requires true values of σ^2 and $\boldsymbol{D}$, which are unknown. We compute their estimates and use them in the minimization.

Specifically, the algorithm is as follows. We assume here that the geometric constraint is given in the explicit form $\boldsymbol{\eta}_i = \boldsymbol{g}(\boldsymbol{\xi}_i, \boldsymbol{\theta})$. Moreover, we assume that the unit normal vector to the hypersurface and the curvature matrix can be calculated at each point $\boldsymbol{\eta} = \boldsymbol{g}(\boldsymbol{\xi}, \boldsymbol{\theta})$ as $\boldsymbol{n} = \boldsymbol{n}(\boldsymbol{\xi}, \boldsymbol{\theta})$ and $\boldsymbol{D} = \boldsymbol{D}(\boldsymbol{\xi}, \boldsymbol{\theta})$, respectively. The main body of the algorithm consists of two steps, the computation of the latent variables and that of the interest parameter, which are performed alternatively.

0. Initialize $\hat{\boldsymbol{\theta}}$. For example, $\hat{\boldsymbol{\theta}} \leftarrow \hat{\boldsymbol{\theta}}_{ML}$.
1. Using $\hat{\boldsymbol{\theta}}$, calculate $\hat{\boldsymbol{\xi}}_i$ $(i = 1, \ldots, n)$ by the following minimization:

$$\hat{\boldsymbol{\xi}}_i \leftarrow \operatorname*{argmin}_{\boldsymbol{\xi}} |\boldsymbol{x}_i - \boldsymbol{g}(\boldsymbol{\xi}, \hat{\boldsymbol{\theta}})|^2. \tag{13}$$

   Moreover, calculate the estimate $\hat{\sigma}^2$ of the noise variance σ^2 using the residual squared differences.



(a) (b)

Figure 7. Results of ellipse fitting by the proposed method.

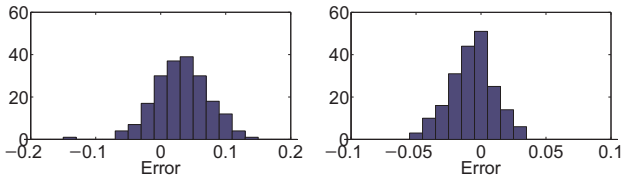


Figure 8. Histograms of errors of the bias-corrected estimates for the axis parameters a and b. Left: $\hat{a} - a$. Right: $\hat{b} - b$.

2. Using $\hat{\boldsymbol{\xi}}_i$, calculate $\hat{\boldsymbol{\theta}}$ $(i = 1, \ldots, n)$ by the following minimization:

$$\hat{\boldsymbol{\theta}} = \operatorname*{argmin}_{\boldsymbol{\theta}} \sum_{i=1}^{n} |\boldsymbol{e}_i|^2, \tag{14}$$

   where, for $i = 1, \ldots, n$,

$$\boldsymbol{e}_i = \frac{1}{\hat{\sigma}(\boldsymbol{I} - \hat{\sigma}^2\mathrm{tr}(\boldsymbol{D}_i^2))} \left(\boldsymbol{x}_i - \boldsymbol{g}(\hat{\boldsymbol{\xi}}_i, \boldsymbol{\theta}) + \hat{\sigma}^2\mathrm{tr}(\boldsymbol{D}_i)\boldsymbol{n}_i\right), \tag{15}$$

   where $\boldsymbol{D}_i = \boldsymbol{D}(\hat{\boldsymbol{\xi}}_i, \boldsymbol{\theta})$ and $\boldsymbol{n}_i = \boldsymbol{n}(\hat{\boldsymbol{\xi}}_i, \boldsymbol{\theta})$.
3. Go to Step 1 until convergence.

5.2. Experimental results: ellipse fitting

We applied the above algorithm to Example 2, i.e., ellipse fitting. An identical set of data and parameters were used. The results are shown in Figs.7 and 8. A comparison of these figures with Fig.5 and 6 reveals that the bias in the major axis direction mostly disappears; this shows the effectiveness of the proposed method. However, from Fig.8, it appears that a small bias remains in the estimates of a. This may be due to the inaccuracy of the values of $\boldsymbol{D}$ and $\boldsymbol{n}$ used in the algorithm, for which we use the values at the estimate $\hat{\boldsymbol{\eta}}_i \equiv \boldsymbol{g}(\hat{\boldsymbol{\xi}}_i, \boldsymbol{\theta})$ of $\boldsymbol{\eta}_i (= \boldsymbol{g}(\boldsymbol{\xi}_i, \boldsymbol{\theta}))$.

5.3. Curvature of the epipolar geometry

For some problems, it is possible that the bias of ML estimates is always negligible in practice. If so, it is obviously unnecessary to apply the above algorithm. However, the necessity of bias correction is usually unknown in advance. Thus, it is desirable to be able to judge in advance whether or not the bias will be nonnegligibly large.

This is made possible by comparing the order of the correction term in Eq.(15) with other terms. The correction

term is based on the mean of the signed distance d, which is given by $\sigma^2\mathrm{tr}(\boldsymbol{D}) = (k-1)/2\sigma^2\kappa$, where k is the dimensionality of the observation vectors and κ is the mean curvature of the hypersurface. The magnitude of the term $\sigma^2\mathrm{tr}(\boldsymbol{D})$ is compared with the distance $|\boldsymbol{x}_i - \boldsymbol{g}(\hat{\boldsymbol{\xi}}, \boldsymbol{\theta}))|$. In an average sense, the distance has the order of the noise standard deviation σ. Thus, the criterion for the above judgment is given by whether the ratio of $\sigma^2\mathrm{tr}(\boldsymbol{D})$ and σ, i.e., $\sigma\mathrm{tr}(\boldsymbol{D})$, has the order of 1; if it is much smaller than 1, the ordinary ML estimate is considered to be sufficiently accurate. This procedure is also practical in that this ratio can be computed as long as the mean curvature κ can be calculated.

We apply this method to estimating a fundamental matrix.

Example 3. Let (u_1, v_1) and (u_2, v_2) be the image coordinates of a 3D point for the left and right cameras, respectively. An observation is given as $\tilde{\boldsymbol{x}} = [u_1, v_1, u_2, v_2]$. Defining the homogeneous vectors $\boldsymbol{x}_1 = [u_1, v_1, 1]^\top$ and $\boldsymbol{x}_2 = [u_2, v_2, 1]^\top$, the epipolar constraint is given as $f(\tilde{\boldsymbol{x}}; \boldsymbol{F}) = \boldsymbol{x}_1^\top \boldsymbol{F} \boldsymbol{x}_2 = 0$. The objective is to estimate $\boldsymbol{F}$.

Using a formula given in [3], the mean curvature κ of the hypersurface implicitly given by $f = 0$ can be calculated, from which we have

$$\mathrm{tr}(\boldsymbol{D})(= 3\kappa/2) = \frac{\boldsymbol{x}_1^\top \boldsymbol{F}\boldsymbol{U}\boldsymbol{F}^\top \boldsymbol{U}\boldsymbol{F}\boldsymbol{x}_2}{\{(\boldsymbol{U}\boldsymbol{F}^\top \boldsymbol{x}_1)^2 + (\boldsymbol{U}\boldsymbol{F}\boldsymbol{x}_2)^2\}^{3/2}}, \qquad (16)$$

where $\boldsymbol{U} = \mathrm{diag}[1, 1, 0]$. Then, our concern is how large $\mathrm{tr}(\boldsymbol{D})$ will be and furthermore whether it has a large value for a particular local part of the hypersurface.

First, several properties of the above quantity $\mathrm{tr}(\boldsymbol{D})$ can be analytically derived. One property is that when its denominator approaches 0, it has a large value. The denominator vanishes if and only if $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ are both on epipoles. This corresponds to a degenerate case where the 3D coordinates of the point cannot be uniquely determined. Since $\mathrm{tr}(\boldsymbol{D})$ has a large value around the point, it may be necessary to further investigate this case. Another property is that $\mathrm{tr}(\boldsymbol{D})$ can be 0 for some configurations. For example, when two cameras have parallel optical axes and for both cameras, the skew is 0 and the aspect ratio is 1, the numerator of Eq.(16) vanishes. In this case, the ML estimate $\hat{\boldsymbol{F}}$ is considered to be highly accurate.

For other general cases, it is possible to evaluate $\mathrm{tr}(\boldsymbol{D})$ only numerically. Fig.9 shows one such example. Assuming the image size to be 640×480 pixels, we generate images for multiple camera layouts. For every layout, we have confirmed that the value of $\mathrm{tr}(\boldsymbol{D})$ calculated at each observation $\tilde{\boldsymbol{x}}$ was less than 10^{-3}. The histogram of Fig.9 shows the distribution of $\mathrm{tr}(\boldsymbol{D})$ for the layout shown on the left. Assuming the noise strength to be $\sigma = 1$ pixel, the criterion of bias correction, i.e., $\sigma\mathrm{tr}(\boldsymbol{D})$, has an absolute value smaller than 10^{-3}. Thus, in this case, we can assume that the bias is negligible and the standard ML estimate is very accurate.

6. Summary

In this paper, we have discussed the bias of ML estimates for computer vision problems involving the estimation of geometric parameters. One characteristic of these problems is that there is the same number of latent variables as the number of observations. It is usually necessary to eliminate these latent variables when estimating the interest parameters, which contributes to the emergence of a bias. In relation to this, we show that for a class of problems in which the geometric constraint gives a hypersurface in the observation space, the magnitude of the bias depends on the mean curvature of the hypersurface. Based on this analysis, for the same class of problems, we present a method for computing a bias-corrected estimate. The estimate is computed by minimizing a cost function that is obtained by modifying the negative log-likelihood so as to reduce the bias. Using the problem of ellipse fitting and the estimation of a fundamental matrix as examples, we have demonstrated the effectiveness of the proposed approach.

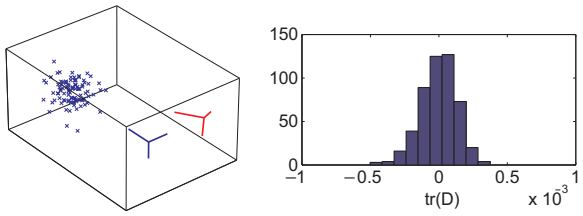


Figure 9. Left: An example of examined layouts. Right: Histogram of curvatures at each pair of corresponding points.

Acknowledgments

The authors would like to thank anonymous reviewers for their helpful comments on improving this paper.

References

[1] B. Efron. Bootstrap confidence intervals for a class of parametric problems. *Biometrika*, 72(1):45–58, 1985.

[2] V. P. Godambe, editor. *Estimating functions*. Oxford University Press, 1991.

[3] R. Goldman. Curvature formulas for implicit curves and surfaces. *Computer Aided Geometric Design*, 22:623–658, 2005.

[4] K. Kanatani. Statistical optimization for geometric fitting: Theoretical accuracy bound and high order error analysis. *International Journal of Computer Vision*, 80(2):167–188, 2007.

[5] J. Neyman and E. Scott. Consistent estimates based on partially consistent observations. *Econometrica*, 16(1):1–32, 1948.

[6] T. Okatani and K. Deguchi. Toward a statistically optimal method for estimating geometric relations from noisy data: Cases of linear relations. In *Proceedings of CVPR*, volume 1, pages 432–439, 2003.

[7] P. H. S. Torr. Bayesian model estimation and selection for epipolar geometry and generic manifold fitting. *International Journal of Computer Vision*, 50(1):35–61, 2002.

[8] R. P. Waterman and B. G. Lindsay. Projected score methods for approximating conditional scores. *Biometrika*, 83(1):1–13, 1996.

[9] M. Werman and D. Keren. A Bayesian method for fitting parametric and nonparametric models to noisy data. *IEEE Trans. Pattern Anal. Mach. Intell.*, 23(5):528–534, 2001.