3rd Int. Workshop on Statistical and Computational Theories of Vision, October 2003, Nice, France.

How Are Statistical Methods for Geometric Inference Justified?

Kenichi Kanatani

Department of Information Technology, Okayama University, Okayama 700-8530 Japan

kanatani@suri.it.okayama-u.ac.jp

Abstract

This paper investigates the meaning of "statistical methods" for geometric inference based on image feature points. Tracing back the origin of feature uncertainty to image processing operations for computer vision in general, we discuss implications of asymptotic analysis in reference to "geometric fitting" and "geometric model selection". For the latter, we point out the prominent characteristics of the "geometric AIC" and the "geometric MDL" as compared with Akaike's AIC and Rissanen's MDL and present a dual interpretation between the standard and geometric inferences. We also evaluate their degeneracy detection performance by simulation, showing that their asymptotic characteristics are very contrasting. Finally, we discuss some issues concerning "nuisance parameters" and "semiparametric models". We conclude that application of statistical methods requires careful considerations about the peculiar nature of the geometric inference problem.

1. Introduction

Statistical inference from images has been one of the key components of computer vision research. Traditionally, statistical methods have been used for recognition and classification purposes. Recently, however, there are many studies of statistical analysis for *geometric inference* based on geometric primitives such as points and lines extracted by image processing operations.

However, the term "statistical" has somewhat a different meaning for such geometric inference than for recognition and classification purposes. This difference has often been overlooked, causing controversies over the validity of the statistical approach for geometric problems in general.

In this paper, we take a close look at this problem, tracing back the origin of feature uncertainty to image processing operations for computer vision. We then discuss implications of asymptotic analysis in reference to "geometric fitting" and "geometric model selection". For the latter, we point out the prominent characteristics of the "geometric AIC" and the "geometric MDL" as compared with Akaike's AIC and Rissanen's MDL present a dual interpretation between the standard and geometric inferences. We also evaluate their degeneracy detection performance by simulation, showing that their asymptotic characteristics are very contrasting. Finally, we discuss some issues concerning "nuisance parameters" and "semiparametric models" in relation to geometric inference. We conclude that application of statistical methods requires careful considerations about the peculiar nature of the geometric inference problem.

2. Statistical Methods for Geometric Inference

First, we clarify the meaning of a "statistical method".

2.1 What is a statistical method?

The goal of statistical methods is not to study the properties of observed data themselves but to infer the properties of the *ensemble* from which we regard the observed data as having been sampled. The ensemble may be a collection of existing entities (e.g., the entire population), but often it is a hypothetical set of conceivable possibilities.

When a statistical method is employed, the underlying ensemble is often taken for granted. For character recognition, for instance, it is understood that we are thinking of an ensemble of all prints and scripts of individual characters. Since some characters are more likely to appear than others, a *probability distribution* is naturally defined over the ensemble.

For handwritten character recognition, our attention is restricted to the set of all handwritten characters. The ensemble is further restricted if we want to recognize characters written by a specific writer (e.g., his/her signatures), but these restrictions are too obvious to be mentioned. However, this issue is very crucial for geometric inference from image features. To show this is the main purpose of this paper.

2.2 Ensembles for geometric inference

What we call *geometric inference* in this paper deals with a *single* image (or a single *set* of images). For example, we observe an image of a building and extract *feature points* such as isolated points, corners, and intersections of lines. Our task is to test if a particular geometric constraint exists on them. If so, we estimate the parameters of the constraint and evaluate the degree of uncertainty of that estimation.

The reason why we need a statistical method is that *the extracted feature positions have uncertainty*. So, we have to judge the extracted feature points as collinear if they are sufficiently aligned. We can also evaluate the degree of uncertainty of the fitted line by propagating the uncertainty of the individual points. What is the ensemble that underlies this type of inference?

This question reduces to the question of *why the uncertainty of the feature points occurs at all*. After all, statistical methods are not necessary if the data are exact. Using a statistical method means regarding the current feature position as randomly sampled from a set of its *possible positions*. But *where else could it be if not in the current position*? This is the key question of this paper.

2.3 Uncertainty of feature extraction

Many algorithms have been proposed for extracting feature points including the Harris operator [9] and SUSAN [38], and their performance has been extensively compared [4, 33, 37]. However, if we use, for example, the Harris operator to extract a particular corner of a particular building image, the output is unique (Fig. 1). No matter how many times we repeat the extraction, we obtain the same point because no external disturbances exist and the internal parameters (e.g., thresholds for judgment) are unchanged. It follows that the current position is the sole possibility. How can we find it elsewhere? In the past, no satisfactory answer seems to have been given to this question.

If we closely examine the situation, we are compelled to conclude that other possibilities should exist *because the extracted position is not necessarily correct*. But if the extracted position is not correct, why did we extract it? Why didn't we extract the correct

(a) (b)

Figure 1: (a) A feature point in an image of a building. (b) Its enlargement and the uncertainty of the feature location.

position in the first place? The answer is: *we cannot*. Why is this impossible? This is the core of our discussion.

2.4 Image processing for computer vision

The reason why there exist so many feature extraction algorithms, none of them being definitive, is that they are aiming at an *intrinsically impossible task*. If we were to extract a point around which, say, the intensity varies to the largest degree measured in such and such a criterion, the algorithm would be unique (variations may exist in intermediate steps, but the final output should be the same).

However, what we want is not "image properties" but "3-D properties" such as corners of a building, but the way a 3-D property is translated into an image property is intrinsically heuristic. As a result, as many algorithms can exist as the number of heuristics for its 2-D interpretation. This fact has not been given much attention in the past, and feature extraction has often been regarded as an objective image processing task.

If we specify a 3-D feature that we want to extract, its appearance in the image is not unique. It is affected by various properties of the scene including the details of its 3-D shape, the viewing orientation, the illumination condition, and the light reflectance properties of the material. A slight difference of any of them can result in a substantial difference on the image plane.

Theoretically, exact feature extraction would be possible if all the properties of the scene were exactly known, but *to infer them from images is the very task of computer vision*. It follows that we must make a *guess* in the image processing stage. For the current image, some guesses may be correct, but others may be wrong.

This means that the exact feature position could be found only by an (non-existing) "ideal" algorithm that could guess everything correctly. In reality, however, a wrong position may be located if the algorithm is based on wrong guesses. This observation allows us to interpret the "possible feature positions" to be the positions that would be located by different (non-ideal) algorithms based on different guesses. So far, this interpretation does not seem to have been explicitly noticed by computer vision researchers.

Our assertion is that the set of hypothetical positions should be associated with *the set of hypothetical algorithms*. The current position is regarded as produced by an algorithm sampled from it. This explains why one always obtains the same position no matter how many times one repeats extraction using that algorithm. To obtain a different position, one has to sample another algorithm.

3. Statistical Model of Feature Location

Next, we examine the meaning of "uncertainty" of feature locations.

3.1 Covariance matrix of a feature point

For doing statistical analysis based on the above interpretation, we hypothesize that the "mean" of the potential positions coincides with the true position. In other words, all hypothetical algorithms as a whole are assumed to be unbiased. Without this hypothesis, efforts to devise good algorithms would be meaningless.

The performance of feature point extraction depends on the image properties around that point. If, for example, we want to extract a point in a region with an almost homogeneous intensity, the resulting position may be ambiguous whatever algorithm is used. In other words, the positions that the hypothetical algorithms would extract should have a large spread around the true position. If, on the other hand, the intensity greatly varies around that point, any algorithm could easily locate it accurately, meaning that the positions that the hypothetical algorithms would extract should have a strong peak at the true position. This observation suggests that we may introduce for each feature point its *covariance matrix* that measures the spread of its potential positions.

Let $V[p_\alpha]$ be the covariance matrix of the αth feature point p_α. The above argument implies that we can determine the qualitative characteristics of uncertainty in relative terms but not its absolute magnitude. If, for example, the intensity variations around p_α are almost the same in all directions, we can think of the probability distribution as isotropic, a typical equiprobability line, known as the *uncertainty ellipses*, being a circle (Fig. 1(b)). If, on the other hand, p_α is on an object boundary, distinguishing it from nearby points should be difficult whatever algorithm is used, so its covariance matrix should have an elongated uncertainty ellipse along that boundary.

From these observations, we write the covariance matrix $V[p_\alpha]$ in the form

$$V[p_\alpha] = \epsilon^2 V_0[p_\alpha], \tag{1}$$

where ϵ is an unknown magnitude of uncertainty, which we call the *noise level*. The matrix $V_0[p_\alpha]$, which we call the *normalized covariance matrix*, describes the relative magnitude and the dependence on orientations.

Most existing feature extraction algorithms are designed to output those points that have large image variations around them, so points in a region with an almost homogeneous intensity or on object boundaries are rarely chosen as feature points. As a result, the covariance matrix of a feature point extracted by such an algorithm can be regarded as nearly isotropic. This has also been confirmed by experiments [22], justifying the use of the identity as the normalized covariance matrix.

Remark 1. The intensity variations around different feature points are usually unrelated, so their uncertainty can be regarded as statistically independent. However, if we track feature points over consecutive video frames, it has been observed that the uncertainty of each point has strong correlations over the frames [39].

Remark 2. Many interactive applications require humans to extract feature points by manipulating a mouse. Extraction by a human is also an "algorithm", and it has been

shown by experiments that humans are likely to choose "easy-to-see" points such as isolated points and intersections, avoiding points in a region with an almost homogeneous intensity or on object boundaries [22]. In this sense, the statistical characteristics of human extraction are very similar to machine extraction. This is no surprise if we recall that image processing for computer vision is essentially a heuristic that simulates human perception. It has also been reported that strong microscopic correlations exist when humans manually select corresponding feature points over multiple images [28].

3.2 Image processing in computer vision

We have observed that the ensemble behind geometric inference is the set of algorithms and that statistical assumptions such as normality, independence, unbiasedness, and correlations are *properties of the underlying set of algorithms*. In the past, however, they were taken to be properties of the image. This has caused a lot of confusion. The main cause of this confusion may be the tradition that the uncertainty of feature points has simply been referred to as "image noise", giving a misleading impression as if the feature locations fluctuate because of *random intensity perturbations of individual pixels*.

Of course, we may obtain better results if we have higher-quality images whatever algorithm is used. The performance of any algorithm naturally depends on the image quality. However, the task of computer vision is not to analyze "image properties" but to study the "3-D properties" of the objects that we are viewing. As long as the image properties and the 3-D properties do not correspond one to one, any image processing for computer vision inevitably entails some degree of uncertainty, however high the image quality may be.

Thus, we conclude that as long as a hiatus exists between what we really want to extract and what we are actually computing, *any* process of computer vision accompanies uncertainty independent of the image quality, and the result must be interpreted *statistically*. The underlying ensemble is the set of *hypothetical (inherently imperfect) algorithms of image processing*, which should be distinguished from "image noise" caused by random intensity fluctuations of individual pixels. Yet, it has been customary to evaluate the performance of image processing algorithms for computer vision by *adding independent Gaussian noise to individual pixels*.

Remark 3. The above argument also applies to *edge detection*, whose goal is to find the boundaries of 3-D objects in the scene. In reality, however, all existing algorithms seek *edges*, i.e., lines and curves across which the intensity changes discontinuously. From the above observations, it can be argued that edge detection is also intrinsically a heuristic and hence no definitive algorithm will ever be found.

4. Asymptotic Analysis for Geometric Fitting

Here, we clarify the meaning of "asymptotic analysis".

4.1 What is asymptotic analysis?

As stated earlier, *statistical estimation* refers to estimating the properties of an ensemble from a finite number of samples chosen from it, assuming some knowledge, or a *model*, about the ensemble.

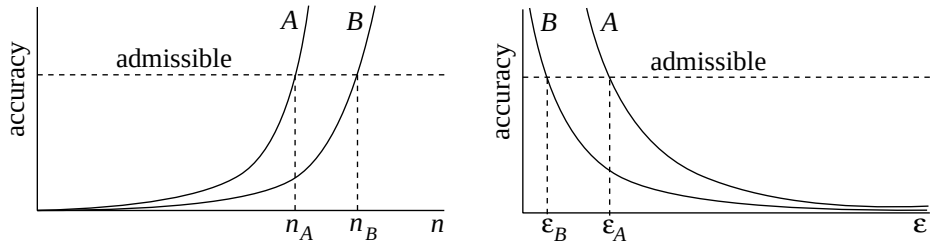


Figure 2: (a) For the standard statistical estimation problem, it is desired that the accuracy increases rapidly as the number of experiments $n \to \infty$, because admissible accuracy can be reached with a smaller number of experiments. (b) For geometric inference, it is desired that the accuracy increases rapidly as the noise level $\epsilon \to 0$, because admissible accuracy can be reached in the presence of larger uncertainty.

If the uncertainty originates from external conditions, as in experiments in physics, the estimation accuracy can be increased by controlling the measurement devices and environments. For internal uncertainty, on the other hand, there is no way of increasing the accuracy except by repeating the experiment and doing statistical inference. However, repeating experiments usually entails costs, and often the number of experiments is limited in practice.

Taking account of such practical considerations, statisticians usually evaluate the performance of estimation *asymptotically*, analyzing the growth in accuracy as the number n of experiments increases. This is justified because a method whose accuracy increases more rapidly as $n \to \infty$ than others can reach admissible accuracy *with a fewer number of experiments* (Fig. 2(a)).

In contrast, the ensemble for geometric inference based on feature points is, as we have seen, the set of potential feature positions that could be located if other (hypothetical) algorithms were used. The goal is to estimate geometric quantities as closely as possible to their expectations, which we assume are their true values. In other words, we want to minimize the discrepancy between obtained estimates and their true values *on average over all hypothetical algorithms*.

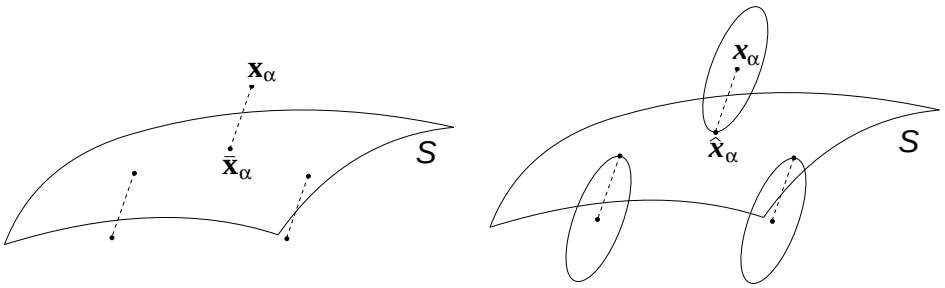
However, the crucial fact is, as stated earlier, we can choose only *one* sample from the ensemble as long as we use a particular image processing algorithm. In other words, the number n of experiments is 1. Then, how can we evaluate the performance of statistical estimation? This is the second main question of this paper.

Evidently, we want a method whose accuracy is sufficiently high *even for large feature uncertainty*. This implies that we need to analyze the growth in accuracy as the noise level ϵ decreases, because a method whose accuracy increases more rapidly as $\epsilon \to 0$ than others can reach admissible accuracy with larger uncertainty of feature extraction (Fig. 2(b)).

4.2 Geometric fitting

We now illustrate our assertion in more specific terms. Let p_1, ..., p_N be the extracted feature points. Suppose each point should satisfy a parameterized constraint

$$F(p_\alpha, \boldsymbol{u}) = 0 \tag{2}$$



(a) (b)

Figure 3: (a) Fitting a manifold $\mathcal{S}$ to the data $\{\boldsymbol{x}_\alpha\}$. (b) Estimating $\{\bar{\boldsymbol{x}}_\alpha\}$ and $\boldsymbol{u}$ by minimizing the sum of squared Mahalanobis distance with respect to the normalized covariance matrices $V_0[\boldsymbol{x}_\alpha]$.

when no uncertainty exists. In the presence of uncertainty, eq. (2) may not hold exactly. Our task is to estimate the parameter $\boldsymbol{u}$ from observed positions p_1, ..., p_N in the presence of uncertainty.

A typical problem of this form is to fit a line or a curve (e.g., a circle or an ellipse) to given N points in the image, but this can be straightforwardly extended to multiple images. For example, if a point (x_α, y_α) in one image corresponds to a point (x'_α, y'_α) in another, we can regard them as a single point p_α in a 4-dimensional joint space with coordinates $(x_\alpha, y_\alpha, x'_\alpha, y'_\alpha)$. If the camera imaging geometry is modeled as perspective projection, the constraint (2) corresponds to the *epipolar equation*; the parameter $\boldsymbol{u}$ is the *fundamental matrix* [10].

4.3 General geometric fitting

The above problem can be stated in more general terms. We view a feature point in the image plane or a set of feature points in the joint space as an m-dimensional vector $\boldsymbol{x}$; we call it a "datum". Its covariance matrix is denoted by $V[\boldsymbol{x}]$, which is an $m \times m$ symmetric matrix. Assuming that multiple constraints exist on each datum $\boldsymbol{x}$, we define the *geometric fitting* problem as follows.

Let $\boldsymbol{x}_1$, ..., $\boldsymbol{x}_N$ be N observed data. Their true values $\bar{\boldsymbol{x}}_1$, ..., $\bar{\boldsymbol{x}}_N$ are supposed to satisfy r constraint equations

$$F^{(k)}(\bar{\boldsymbol{x}}_\alpha, \boldsymbol{u}) = 0, \quad k = 1, ..., r, \tag{3}$$

parameterized by a p-dimensional vector $\boldsymbol{u}$. We call eq. (3) the (*geometric*) *model*. The domain $\mathcal{X}$ of the data $\{\boldsymbol{x}_\alpha\}$ is called the *data space*; the domain $\mathcal{U}$ of the parameter $\boldsymbol{u}$ is called the *parameter space*. The number r of the constraint equations is called the *rank* of the constraint. The r equations $F^{(k)}(\boldsymbol{x}, \boldsymbol{u}) = 0$, $k = 1$, ..., r, are assumed to be mutually independent, defining a manifold $\mathcal{S}$ of codimension r parameterized by $\boldsymbol{u}$ in the data space $\mathcal{X}$. Eq. (3) requires that the true values $\{\bar{\boldsymbol{x}}_\alpha\}$ be all in the manifold $\mathcal{S}$. Our task is to estimate the parameter $\boldsymbol{u}$ from the noisy data $\{\boldsymbol{x}_\alpha\}$ (Fig. 3(a)).

4.4 Maximum likelihood estimation

Let

$$V[\boldsymbol{x}_\alpha] = \epsilon^2 V_0[\boldsymbol{x}_\alpha] \tag{4}$$

be the covariance matrix of $\boldsymbol{x}_\alpha$, where ϵ and $V_0[\boldsymbol{x}_\alpha]$ are the noise level and the normalized covariance matrix, respectively. If the distribution of uncertainty is Gaussian, which we assume hereafter, the probability density of the data $\{\boldsymbol{x}_\alpha\}$ is given by

$$P(\{\boldsymbol{x}_\alpha\}) = C \prod_{\alpha=1}^{N} e^{-(\boldsymbol{x}_\alpha - \bar{\boldsymbol{x}}_\alpha, V[\boldsymbol{x}_\alpha]^{-1}(\boldsymbol{x}_\alpha - \bar{\boldsymbol{x}}_\alpha))/2} = Ce^{-J/2\epsilon^2}, \tag{5}$$

where C is a normalization constant and we put

$$J = \sum_{\alpha=1}^{N} (\boldsymbol{x}_\alpha - \bar{\boldsymbol{x}}_\alpha, V_0[\boldsymbol{x}_\alpha]^{-1}(\boldsymbol{x}_\alpha - \bar{\boldsymbol{x}}_\alpha)), \tag{6}$$

which is known as the sum of squared *Mahalanobis distance* with respect to the normalized covariance matrices $V_0[\boldsymbol{x}_\alpha]$.

Maximum likelihood estimation (*MLE*) is to find the values of $\{\bar{\boldsymbol{x}}_\alpha\}$ and $\boldsymbol{u}$ that minimize eq. (6) subject to the constraint (3) (Fig. 3(b)). The solution is called the *maximum likelihood* (*ML*) *estimator*. If the uncertainty is small, which we assume hereafter, the constraint (3) can be eliminated by introducing Lagrange multipliers and applying first order approximation. After some manipulations, we obtain the following form [11]:

$$J = \sum_{\alpha=1}^{N} \sum_{k,l=1}^{r} W_\alpha^{(kl)} F^{(k)}(\boldsymbol{x}_\alpha, \boldsymbol{u}) F^{(l)}(\boldsymbol{x}_\alpha, \boldsymbol{u}). \tag{7}$$

Here, $W_\alpha^{(kl)}$ is the (kl) element of the inverse of the $r \times r$ matrix whose (kl) element is $(\nabla_{\boldsymbol{x}} F_\alpha^{(k)}, V_0[\boldsymbol{x}_\alpha] \nabla_{\boldsymbol{x}} F_\alpha^{(l)})$; we symbolically write

$$\left(W_\alpha^{(kl)}\right) = \left((\nabla_{\boldsymbol{x}} F_\alpha^{(k)}, V_0[\boldsymbol{x}_\alpha] \nabla_{\boldsymbol{x}} F_\alpha^{(l)})\right)^{-1}, \tag{8}$$

where $\nabla_{\boldsymbol{x}} F^{(k)}$ is the gradient of the function $F^{(k)}$ with respect to $\boldsymbol{x}$. The subscript α means that $\boldsymbol{x} = \boldsymbol{x}_\alpha$ is to be substituted.

Remark 4. The data $\{\boldsymbol{x}_\alpha\}$ may be subject to some constraints. For example, each $\boldsymbol{x}_\alpha$ may be normalized to be a unit vector. The above formulation still holds in such a case if the inverse $V_0[\boldsymbol{x}_\alpha]^{-1}$ in eq. (6) is replaced by the (Moore-Penrose) generalized (or pseudo) inverse $V_0[\boldsymbol{x}_\alpha]^-$ [11].

Remark 5. The r constraints in eq. (3) may be redundant, and only r' $(< r)$ of them may be algebraically independent. The above formulation still holds in such a case if the inverse in eq. (8) is replaced by the generalized inverse with rank r', i.e., the generalized inverse computed after all but r' largest eigenvalues are replaced by zero [11].

4.5 Accuracy of the ML estimator

It can be shown [11] that the covariance matrix of the ML estimator $\hat{\boldsymbol{u}}$ has the form

$$V[\hat{\boldsymbol{u}}] = \epsilon^2 \boldsymbol{M}(\hat{\boldsymbol{u}})^{-1} + O(\epsilon^4), \tag{9}$$

where

$$\boldsymbol{M}(\boldsymbol{u}) = \sum_{\alpha=1}^{N} \sum_{k,l=1}^{r} W_\alpha^{(kl)} \nabla_{\boldsymbol{u}} F_\alpha^{(k)} \nabla_{\boldsymbol{u}} F_\alpha^{(k)\top}. \tag{10}$$

Here, $\nabla_{\boldsymbol{u}} F^{(k)}$ is the gradient of the function $F^{(k)}$ with respect to $\boldsymbol{u}$. The subscript α means that $\boldsymbol{x} = \boldsymbol{x}_\alpha$ is to be substituted.

Remark 6. It can be proved that no other estimators could reduce the covariance matrix further than eq. (9) except for the higher order term $O(\epsilon^4)$ [11, 14]. The ML estimator is optimal in this sense. Note that we are focusing on the asymptotic analysis for small ϵ, as we argued in Sec. 4.1. So, what we call the "ML estimator" should be understood to be a first approximation to the true ML estimator for small ϵ. This is justified because the effects of the first approximation is squeezed into the term $O(\epsilon^4)$.

Remark 7. The p-dimensional parameter vector $\boldsymbol{u}$ may be constrained. For example, it may be normalized to a unit vector. If it has only p' ($< p$) degrees of freedom, the parameter space $\mathcal{U}$ is a p'-dimensional manifold in $\mathcal{R}^p$. In this case, the matrix $\boldsymbol{M}(\boldsymbol{u})$ in eq. (9) is replaced by $\boldsymbol{P}_{\mathbf{u}}\boldsymbol{M}(\boldsymbol{u})\boldsymbol{P}_{\mathbf{u}}$, where $\boldsymbol{P}_{\mathbf{u}}$ is the projection matrix onto the tangent space to the parameter space $\mathcal{U}$ at $\boldsymbol{u}$ [11]. Then inverse $\boldsymbol{M}(\hat{\boldsymbol{u}})^{-1}$ in eq. (9) is replaced by the generalized inverse $\boldsymbol{M}(\hat{\boldsymbol{u}})^{-1}$ with rank p' [11].

5. Asymptotic Analysis for Geometric Model Selection

Now, we turn to "model selection".

5.1 Geometric model selection

Geometric fitting is to estimate the parameter $\boldsymbol{u}$ of the model given in the form of (3). If, on the other hand, we have multiple candidate models

$$F_1^{(k)}(\bar{\boldsymbol{x}}_\alpha, \boldsymbol{u}_1) = 0, \quad F_2^{(k)}(\bar{\boldsymbol{x}}_\alpha, \boldsymbol{u}_2) = 0, \quad F_3^{(k)}(\bar{\boldsymbol{x}}_\alpha, \boldsymbol{u}_3) = 0, \quad ..., \tag{11}$$

from which we are to select an appropriate one for the observed data $\{\boldsymbol{x}_\alpha\}$, the problem is (*geometric*) *model selection* [11, 13, 15].

5.2 Geometric AIC

A naive idea for model selection is first to estimate the parameter $\boldsymbol{u}$ by MLE and compute the *residual* (*sum of squares*), i.e., the minimum value $\hat{J}$ of J in eq. (6), for each model. Then, we select the one that has the smallest residual $\hat{J}$. This does not work, however, because the ML estimator $\hat{\boldsymbol{u}}$ is determined so as to minimize the residual $\hat{J}$, so the residual $\hat{J}$ can be made arbitrarily smaller if the model is equipped with more parameters to adjust.

This observation leads to the idea of compensating for the negative bias of the residual caused by substituting the ML estimator. This is the principle of Akaike's *AIC* (*Akaike information criterion*) [1], which is derived from the asymptotic behavior of the *Kullback-Leibler information* (or *divergence*) as the number n of experiments goes to infinity. Doing a similar analysis to Akaike's and examining the asymptotic behavior as the noise level ϵ goes to zero, as discussed in Sec. 4.1, we obtain the following *geometric AIC* [11, 12]:

$$\text{G-AIC} = \hat{J} + 2(Nd + p)\epsilon^2 + O(\epsilon^4). \tag{12}$$

Here, d is the dimension of the manifold $\mathcal{S}$ defined by the constraint (3); p is the dimension of $\boldsymbol{u}$ (i.e., the number of unknowns). The model for which eq. (12) is the smallest is regarded as the best model. The derivation of eq. (12) is based on the following facts [11, 12]:

- The ML estimator $\hat{\boldsymbol{u}}$ converges to its true value as $\epsilon \to 0$.
- The ML estimator $\hat{\boldsymbol{u}}$ obeys a Gaussian distribution under linear constraints, because the noise is assumed to be Gaussian. For nonlinear constraints, linear approximation can be justified in the neighborhood of the solution if ϵ is sufficiently small, which we assume as discussed in Sec. 4.1.
- A quadratic form in standardized Gaussian random variables is subject to a χ^2 distribution, whose expectation is equal to its degree of freedom.

5.3 Geometric MDL

Another well known criterion for model selection is Rissanen's *MDL* (*minimum description length*) [34, 35, 36], which measures the goodness of a model by the minimum information theoretic code length of the data and the model. The idea is simple, but the following issues must be resolved for applying it in practice:

- Encoding a problem involving real numbers requires an infinitely long code length.
- The probability density, according to which a minimum length code can be obtained, involves unknown parameters.
- Evaluating an exact form of the minimum length code is very difficult.

Rissanen [34, 35, 36] avoided these difficulties by quantizing real numbers in a way that does not depend on individual models and substituting the ML estimators for the unknown parameters. They, too, are real numbers, so they are also quantized. The quantization width is chosen so as to minimize the total description length (*two-stage encoding*). The resulting code length is evaluated asymptotically as the data length n goes to infinity. If we analyzes the asymptotic behavior of encoding the geometric fitting problem as the noise level ϵ goes to zero, as discussed in Sec. 4.1, we obtain the following *geometric MDL* [16]:

$$\text{G-MDL} = \hat{J} - (Nd + p)\epsilon^2 \log\Big(\frac{\epsilon}{L}\Big)^2 + O(\epsilon^2). \tag{13}$$

Here, L is a reference length chosen so that its ratio to the magnitude of data is $O(1)$, e.g., L can be taken to be the image size for feature point data. Its exact determination requires an a priori distribution that specifies where the data are likely to appear, but it has been observed that the model selection is not very much affected by L as long as it is within the same order of magnitude [16].

6. Dual Interpretations for Statistical Inference

We now contrast geometric fitting (Sec. 4) and geometric model selection (Sec. 5) with the standard statistical estimation and the standard (stochastic) model selection. This section is another core of this paper.

6.1 Standard statistical inference

The asymptotic analysis in the preceding two sections bears a strong resemblance to the standard statistical analysis, although the basic premises and assumptions are different. In the standard statistical analysis, as pointed out in Sec. 2, an experiment is regarded as an instance of sampling from an ensemble with a probability density $P(\boldsymbol{x}|\boldsymbol{\theta})$

parameterized by a k-dimensional vector $\boldsymbol{\theta}$. Suppose we repeat the experiment independently n times and observe n outcomes $\boldsymbol{x}_1$, $\boldsymbol{x}_2$, ..., $\boldsymbol{x}_n$. Our task is to estimate the parameter $\boldsymbol{\theta}$ of the (*stochastic*) *model* $P(\boldsymbol{x}|\boldsymbol{\theta})$. *Maximum likelihood estimation* (*MLE*) is to find the value $\boldsymbol{\theta}$ that maximizes $\prod_{\alpha=1}^{n} P(\boldsymbol{x}_\alpha|\boldsymbol{\theta})$, or equivalently minimizes its negative logarithm $-\sum_{\alpha=1}^{n} \log P(\boldsymbol{x}_\alpha|\boldsymbol{\theta})$. It can be shown that the covariance matrix $V[\hat{\boldsymbol{\theta}}]$ of the resulting ML estimator $\hat{\boldsymbol{\theta}}$ converges, under a mild condition, to $\boldsymbol{O}$ as the number n of experiments goes to infinity (*consistency*) in the form

$$V[\hat{\boldsymbol{\theta}}] = \boldsymbol{I}(\boldsymbol{\theta})^{-1} + O\Big(\frac{1}{n^2}\Big), \tag{14}$$

where we define the *Fisher information matrix* $\boldsymbol{I}(\boldsymbol{\theta})$ by

$$\boldsymbol{I}(\boldsymbol{\theta}) = n\mathrm{E}[(\nabla_{\boldsymbol{\theta}} \log P(\boldsymbol{x}|\boldsymbol{\theta}))(\nabla_{\boldsymbol{\theta}} \log P(\boldsymbol{x}|\boldsymbol{\theta}))^\top]. \tag{15}$$

The operation $\mathrm{E}[\,\cdot\,]$ denotes expectation with respect to the density $P(\boldsymbol{x}|\boldsymbol{\theta})$. The first term in the right-hand side of eq. (14) is called the *Cramer-Rao lower bound*, describing the minimum degree of fluctuations in all estimators. Thus, the ML estimator is optimal if n is sufficiently large (*asymptotic efficiency*).

If we have multiple candidate models

$$P_1(\boldsymbol{x}|\boldsymbol{\theta}_1), \quad P_2(\boldsymbol{x}|\boldsymbol{\theta}_2), \quad P_3(\boldsymbol{x}|\boldsymbol{\theta}_3), \quad ..., \tag{16}$$

from which we are to select an appropriate one for the observations $\{\boldsymbol{x}_\alpha\}$, the problem is (*stochastic*) *model selection*. Akaike's AIC has the following form:

$$\mathrm{AIC} = -2\sum_{\alpha=1}^{N} \log P(\boldsymbol{x}_\alpha|\hat{\boldsymbol{\theta}}) + 2k + O\Big(\frac{1}{n}\Big). \tag{17}$$

The model for which this quantity is the smallest is regarded as the best model. The derivation of eq. (17) is based on the following facts [1]:

- The maximum likelihood estimator $\hat{\boldsymbol{\theta}}$ converges to its true value as $n \to \infty$ (the *law of large numbers*).
- The maximum likelihood estimator $\hat{\boldsymbol{\theta}}$ asymptotically obeys a Gaussian distribution as $n \to \infty$ (the *central limit theorem*).
- A quadratic form in standardized Gaussian random variables is subject to a χ^2 distribution, whose expectation equals its degree of freedom.

The Rissanen's MDL has the following form [35, 36]:

$$\mathrm{MDL} = -\sum_{\alpha=1}^{n} \log P(\boldsymbol{x}_\alpha|\hat{\boldsymbol{\theta}}) + \frac{k}{2}\log\frac{n}{2\pi} + \log\int_{\mathcal{T}} \sqrt{|\boldsymbol{I}(\boldsymbol{\theta})|}\,d\boldsymbol{\theta} + O(1). \tag{18}$$

Here, $\hat{\boldsymbol{\theta}}$ is the ML estimator; the symbol $O(1)$ denotes terms of order 0 in n in the limit $n \to \infty$. In order that the integration in the right-hand side of eq. (18) exists, the domain $\mathcal{T}$ of the parameter $\boldsymbol{\theta}$ must be compact. In other words, we must specify in the k-dimensional space of $\boldsymbol{\theta}$ a finite region $\mathcal{T}$ in which the true value of $\boldsymbol{\theta}$ is likely to exist. This is nothing but the *Bayesian* standpoint that requires a prior distribution for the parameter to estimate. If it is not known, we must introduce an appropriate expedient to suppress an explicit dependence on the prior. Such an expedient is also necessary for the geometric MDL, i.e., the introduction of the reference length L in eq. (18).

6.2 Dual interpretations of asymptotic analysis

We have seen above that the covariance matrix of the ML estimator for the standard statistical analysis converges to $\boldsymbol{O}$ as the number n of experiments goes to infinity and that it agrees with the Cramer-Rao lower bound expect for $O(1/n^2)$. If follows that $1/\sqrt{n}$ plays the same role as ϵ for geometric inference.

The same correspondence exists for model selection, too. Note that the unknowns are the p parameters of the constraint plus the N true positions specified by the d coordinates of the d-dimensional manifold $\mathcal{S}$ defined by the constraint. If eq. (12) is divided by ϵ^2, we have $\hat{J}/\epsilon^2 + 2(Nd + p) + O(\epsilon^2)$, which is ($-2$ times the logarithmic likelihood)+2(the number of unknowns), the same form as Akaike's AIC. The same holds for eq. (13), which reduces to Rissanen's MDL if ϵ is replaced by $1/\sqrt{n}$.

This correspondence can be interpreted as follows. Since the underlying ensemble is hypothetical, we can observe only one sample from it as long as a particular algorithm is used. Suppose we hypothetically apply n different algorithms to find n different positions. The optimal estimate of the true position under the Gaussian model is their sample mean. The covariance matrix of the sample mean is $1/n$ times that of the individual samples. Hence, this hypothetical estimation is equivalent to dividing the noise level ϵ in eq. (4) by $\sqrt{n}$.

In fact, there were attempts to generate a hypothetical *ensemble of algorithms* by randomly varying the internal parameters (e.g., the thresholds for judgments) and sample different points [5, 6], not adding any noise to the image. Then, one can compute their means and covariance matrix. Such a process as a whole can be regarded as one operation that effectively achieves higher accuracy.

Thus, the asymptotic analysis for $\epsilon \to 0$ is equivalent to the asymptotic analysis for $n \to \infty$, where n is the number of hypothetical observations. As a result, the expression $\cdots + O(1/\sqrt{n^k})$ in the standard statistical analysis turns into $\cdots + O(\epsilon^k)$ in geometric inference.

6.3 Noise level estimation

In order to use the geometric AIC or the geometric MDL, we need to know the noise level ϵ. If it is not known, it must be estimated. Here arises a sharp contrast to the traditional stochastic model selection.

For stochastic model selection, the noise magnitude is a *model parameter*, because "noise" is defined to be the random effects that cannot be accounted for by the assumed model. Hence, the noise magnitude must be estimated, if not known, *according to the assumed model*. For geometric inference, however, the noise level ϵ is a constant that reflects the *uncertainty of feature detection* (Sec. 3). So, it must be estimated *independently of individual models*.

If we know the true model, it can be estimated from the residual $\hat{J}$ using the knowledge that $\hat{J}/\epsilon^2$ is subject to a χ^2 distribution with $rN - p$ degrees of freedom in the first order [11]. Hence, we obtain an unbiased estimator of ϵ^2 in the form

$$\hat{\epsilon}^2 = \frac{\hat{J}}{rN - p}. \tag{19}$$

The validity of this formula has been confirmed by many simulations.

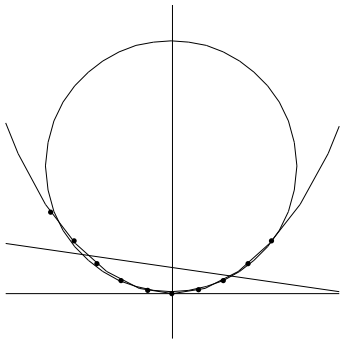


Figure 4: Fitting a line, a circle, and an ellipse.

Remark 8. Eq. (19) can be intuitively understood as follows. Recall that $\hat{J}$ is the sum of square distances from $\{\boldsymbol{x}_\alpha\}$ to the manifold $\hat{\mathcal{S}}$ defined by the constraint $F^{(k)}(\boldsymbol{x}, \boldsymbol{u}) = 0$, $k = 1$, ..., r, in the data space $\mathcal{X}$. Since $\hat{\mathcal{S}}$ has codimension r (the dimension of the orthogonal directions to it), the residual $\hat{J}$ should have expectation $rN\epsilon^2$. However, $\hat{\mathcal{S}}$ is fitted so as to minimize $\hat{J}$ by adjusting its p-dimensional parameter $\boldsymbol{u}$, so the expectation of $\hat{J}$ reduces to $(rN - p)\epsilon^2$.

One may wonder if model selection is necessary at all when the true model is known. In practice, however, a typical situation where model selection is called for is *degeneracy detection*. In 3-D analysis from images, for instance, the constraint (3) corresponds to our knowledge about the scene such as rigidity of motion. However, the computation fails if degeneracy occurs (e.g., the motion is zero). Even if exact degeneracy does not occur, the computation may become numerically unstable in near degeneracy conditions. In such a case, the computation can be stabilized by detecting degeneracy by model selection and switching to a specific model that describes the degeneracy [13, 17, 20, 21, 25, 31, 45].

Degeneracy means addition of new constraints, such as some quantity being zero. As a result, the manifold $\mathcal{S}$ defined by the general constraint degenerates into a submanifold $\mathcal{S}'$ of it. Since the general model holds irrespective of the degeneracy, i.e. $\mathcal{S}' \subset \mathcal{S}$, we can estimate the noise level ϵ from the residual $\hat{J}$ of the general model $\mathcal{S}$ by eq. (19).

7. Comparing the geometric AIC and the geometric MDL

We now compare the performance of the geometric AIC and the geometric MDL in detecting degeneracy.

7.1 Detection of circles and lines

Consider an ellipse that is tangent to the x-axis at the origin O with radius 50 in the y direction and eccentricity $1/\beta$. On it, we take eleven points that have equally spaced x coordinates. Adding random Gaussian noise of mean 0 and variance ϵ^2 to the x and y coordinates of each point independently, we fit an ellipse, a circle, and a line in a statistically optimal manner by a technique called *renormalization* [11, 18, 19]. Fig. 4 shows one instance for $\beta = 2.5$ and $\epsilon = 0.1$. Note that a line and a circle are both special cases (degeneracies) of an ellipse.

Lines, circles, and ellipses define one-dimensional (geometric) models with 2, 3, and 5 degrees of freedom, respectively. Hence, their geometric AIC and the geometric MDL for

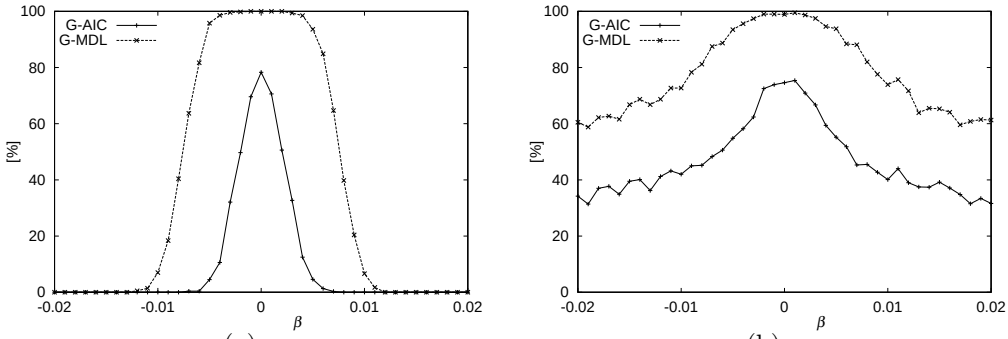


Figure 5: The ratio (%) of detecting a line by the geometric AIC (solid lines) and the geometric MDL (dashed lines) using (a) the true noise level and (b) the estimated noise level.

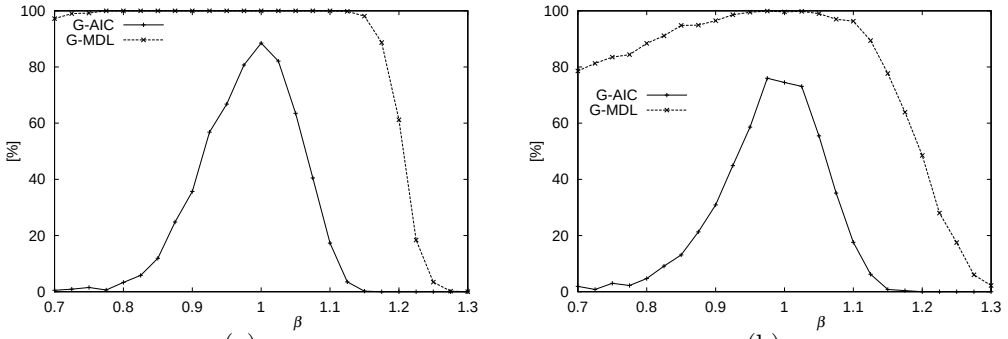


Figure 6: The ratio (%) of detecting a circle by the geometric AIC (solid lines) and the geometric MDL (dashed lines) using (a) the true noise level and (b) the estimated noise level.

N points are given as follows:

$$\begin{aligned}
\text{G-AIC}_l &= \hat{J}_l + 2(N+2)\epsilon^2, \qquad \text{G-MDL}_l = \hat{J}_l - (N+2)\epsilon^2 \log\Big(\frac{\epsilon}{L}\Big)^2, \\
\text{G-AIC}_c &= \hat{J}_c + 2(N+3)\epsilon^2, \qquad \text{G-MDL}_c = \hat{J}_c - (N+3)\epsilon^2 \log\Big(\frac{\epsilon}{L}\Big)^2, \\
\text{G-AIC}_e &= \hat{J}_e + 2(N+5)\epsilon^2, \qquad \text{G-MDL}_e = \hat{J}_e - (N+5)\epsilon^2 \log\Big(\frac{\epsilon}{L}\Big)^2.
\end{aligned} \tag{20}$$

The subscripts l, c, and e refer to lines, circles, and ellipses, respectively. For each β, we compute the geometric AIC and the geometric MDL of the fitted line, circle, and ellipse and choose the one that has the smallest geometric AIC or the smallest geometric MDL. We used the reference length $L = 1$.

Fig. 5(a) shows the percentage of choosing a line for $\epsilon = 0.01$ after 1000 independent trials for each β in the neighborhood of $\beta = 0$. If there were no noise, it should be 0% for $\beta \neq 0$ and 100% for $\beta = 0$. In the presence of noise, the geometric AIC gives a sharp peak, indicating a high capability of distinguishing a line from an ellipse. However, it judges a line to be an ellipse with some probability. The geometric MDL judges a line to be a line almost 100% for small noise, but it judges an ellipse to be a line over a wide range of β.

In Fig. 5(a), we used the true value of the noise variance ϵ^2. If it is unknown, it can be estimated from the residual of the general ellipse model by eq. (19). Fig. 5(b) shows

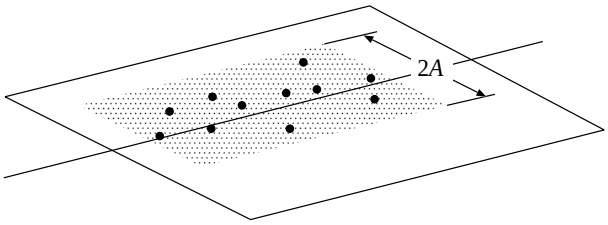


Figure 7: Fitting a space line and a plane to points in space.

the result using its estimate. Although the sharpness is somewhat lost, we observe similar performance characteristics of the geometric AIC and the geometric MDL.

Fig. 6 shows the percentage of choosing a circle for $\epsilon = 0.01$ in the neighborhood of $\beta = 1$. If there were no noise, it should be 0% for $\beta \neq 1$ and 100% for $\beta = 1$. In the presence of noise, as we see, it is difficult to distinguish a small circular arc from a small elliptic arc for $\beta < 1$. Yet, the geometric AIC can detect a circle very sharply, although it judges a circle to be an ellipse with some probability. In contrast, the geometric MDL almost always judges an ellipse to be a circle for $\beta < 1.1$.

7.2 Detection of space lines

Consider a rectangular region $[0, 10] \times [-1, 1]$ in the xy plane. We randomly take eleven points in it and enlarge them A times in the y direction (Fig. 7). Adding random Gaussian noise of mean 0 and variance ϵ^2 to the x, y, and z coordinates of each point independently, we fit a line and a plane in a statistically optimal manner. The rectangular region degenerates into a line segment as $A \to 0$.

A space line is a one-dimensional model with four degrees of freedom; a plane is a two-dimensional model with three degrees of freedom. Hence, their geometric AIC and geometric MDL have the following form:

$$\begin{aligned} \text{G-AIC}_l &= \hat{J}_l + 2(N+4)\epsilon^2, \qquad \text{G-MDL}_l = \hat{J}_l - (N+4)\epsilon^2 \log\Big(\frac{\epsilon}{L}\Big)^2, \\ \text{G-AIC}_p &= \hat{J}_p + 2(2N+3)\epsilon^2, \qquad \text{G-MDL}_p = \hat{J}_p - (2N+3)\epsilon^2 \log\Big(\frac{\epsilon}{L}\Big)^2. \end{aligned} \tag{21}$$

The subscripts l and p refer to lines and planes, respectively. For each A, we compare the geometric AIC and the geometric MDL of the fitted line and plane and choose the one that has the smaller geometric AIC or the smallest geometric MDL. We used the reference length $L = 1$ for the geometric MDL.

Fig. 8(a) shows the percentage of choosing a line for $\epsilon = 0.01$ after 1000 independent trials for each A in the neighborhood of $A = 0$. If there were no noise, it should be 0% for $A \neq 0$ and 100% for $A = 0$. In the presence of noise, the geometric AIC has a high capability of distinguishing a line from a plane, but it judges a line to be a plane with some probability. In contrast, the geometric MDL judges a line to be a line almost 100% for small noise, but it judges a plane to be a line over a wide range of A.

In Fig. 8(a), we used the true value of the noise variance ϵ^2. Fig. 8(b) shows the corresponding result using its estimate obtained from the general plane model by eq. (19). We observe somewhat degraded but similar performance characteristics.

Thus, we can observe that the geometric AIC has a higher capability for distinguishing degeneracy than the geometric MDL, but the general model is chosen with some probabil-

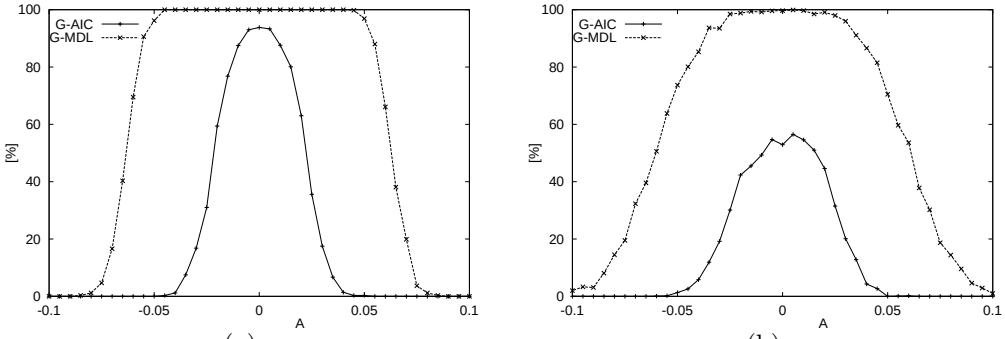


Figure 8: The rate (%) of detecting a space line by the geometric AIC (solid lines) and the geometric MDL (dashed lines) with (a) the true noise level and (b) the estimated noise level.

ity when the true model is degenerate. In contrast, the percentage for the geometric MDL to detect degeneracy when the true model is really degenerate approaches 100% as the noise decreases. This is exactly the dual statement to the well known fact, called the *consistency of the MDL*, that the percentage for Rissanen's MDL to identify the true model converges to 100% in the limit of an infinite number of observations. Rissanen's MDL is regarded by many as superior to Akaike's AIC because the latter lacks this property.

At the cost of this consistency, however, the geometric MDL regards a wide range of nondegenerate models as degenerate. This is natural, since the penalty $-(Nd + p)\epsilon^2 \log(\epsilon/L)^2$ for the geometric MDL in eq. (13) is heavier than the penalty $2(Nd + p)\epsilon^2$ for the geometric AIC in eq. (12). As a result, the geometric AIC is more faithful to the data than the geometric MDL, which is more likely to choose a degenerate model. This contrast in their behavior can also be observed in virtual studio applications for stabilizing the camera position estimation [25].

Remark 9. Despite the difference of geometric model selection from the standard (stochastic) model selection, many attempts have been made in the past to apply Akaike's AIC and their variants to computer vision problems directly [40, 41, 42, 43, 44] without much consideration of the meaning of the underlying asymptotic analysis. Rissanen's MDL is also used in computer vision applications. Its use may be justified if the problem has the standard form of linear/nonlinear regression [3, 26]. Often, however, the solution having a shorter description length was simply chosen with a rather arbitrary definition of the complexity [8, 23, 27].

8. Nuisance Parameters and Semiparametric Model

We now give a brief discussion about a different statistical approach based on the concepts of *nuisance parameters* and *semiparametric models*, which are recently attracting attentions of some computer vision researchers [32, 30]. The underlying assumptions are very contrasting to those of our approach.

8.1 Asymptotic parameters

The number n that appears in the asymptotic analysis of the standard statistical estimation problem is the *number of experiments*. It is also called the *number of trials*,

the *number of observations*, and the *number of samples*. Evidently, the properties of an ensemble are revealed more precisely as more elements are sampled from it.

However, the number n is often called the *number of data*, which has caused considerable confusion. For example, if we observe a 100-dimensional vector data in one experiment, one may think that the "number of data" is 100, but this is wrong: the number n of experiments is 1. We are observing one sample from an ensemble of 100-dimensional vectors.

For character recognition, the underlying ensemble is a set of character images, and the learning process concerns the number n of training steps necessary to establish satisfactory responses. This is independent of the dimension N of the vector that represents each character. The learning performance is evaluated asymptotically as $n \to \infty$, not $N \to \infty$.

For geometric inference, many researchers have taken the dimension of the data as the "number of data" perhaps because the ensemble is hypothetical and one cannot sample more than one datum from it. However, if we extract, for example, 50 feature points, they constitute a 100-dimensional vector consisting of their x and y coordinates. If no other information, such as the intensity value, is used, the image is completely characterized by that vector. Applying a statistical method means regarding it as a sample from a hypothetical ensemble of 100-dimensional vectors.

8.2 Neyman-Scott problem

Many studies of geometric inference for computer vision have analyzed the asymptotic behavior as $N \to \infty$ with respect to the number N of extracted feature points without explicitly mentioning what the underlying ensemble is. This is perhaps motivated by a similar formulation in the statistical literature. Suppose, for example, a rod-like structure lies on the ground in the distance. We emit a laser beam toward it and estimate its position and orientation by observing the reflection of the beam, which is contaminated by noise. We assume that the laser beam can be emitted in any orientation any number of times but the emission orientation is measured with noise. The task is to estimate the position and orientation of the structure as accurately as possible by emitting as small a number of beams as possible. Naturally, the estimation performance should be evaluated in the asymptotic limit $n \to \infty$ with respect to the number n of emissions.

Evidently, this problem has nested ensembles behind. Since we are interested in the position and orientation of the structure but not the exact orientation of each emission, the variables for the former are called the *structural parameters*, which are fixed in number, while the latter are called the *nuisance parameters*, which increase indefinitely as the number n of experiments increases [2]. This type of formulation is called the *Neyman-Scott problem* [29]. Since the constraint is an implicit function in the form of eq. (3), we are considering an *errors-in-variables model* [7]. If we linearize the constraint by changing variables for the data, the noise characteristics differs for each data component, so the problem is *heteroscedastic* [24].

To solve this problem, one can introduce a parametric model for the laser emission orientations and regarding the actual emissions as randomly sampled from it. This formulation is called a *semiparametric model* [2]. An optimal solution can be obtained by finding a good *estimating function* [2, 32].

8.3 Semiparametric model for geometric inference

Since the semiparametric model has something different from the geometric inference problem described in Sec. 2 and 3, a detailed analysis is required for examining if application of a semiparametric model for geometric inference will yield a desirable result [32, 30]. In any event, one should explicitly state what kind of ensemble (or ensemble of ensembles) is assumed before doing statistical analysis.

This is not merely a conceptual issue. It also affects the performance evaluation of simulation experiments. In doing a simulation, one can freely change the number N of feature points and the noise level ϵ. If the accuracy of Method A is higher than Method B for particular values of N and ϵ, one cannot conclude that Method A is superior to Method B, because opposite results may come out for other values of N and ϵ. Here, we have two alternatives for performance evaluation: fixing ϵ and varying N to see if admissible accuracy is attained for a smaller number of feature point; fixing N and varying ϵ to see if admissible accuracy is attained for less certain feature extraction. These two types of evaluation have different meanings. Our conclusion is that the results of one type of evaluation cannot directly be compared with the results of the other.

9. Conclusions

We have investigated the meaning of "statistical methods" for geometric inference based on image feature points. Tracing back the origin of feature uncertainty to image processing operations for computer vision in general, we discussed the implications of asymptotic analysis for performance evaluation in reference to "geometric fitting" and "geometric model selection". For the latter, we pointed out the prominent characteristics of the "geometric AIC" and the "geometric MDL" as compared with Akaike's AIC and Rissanen's MDL presented a dual interpretation between the standard and geometric inferences. We have also evaluated their degeneracy detection performance by simulation, showing that their asymptotic characteristics are very contrasting. Finally, we discussed some issues concerning "nuisance parameters", and "semiparametric models". We conclude that application of statistical methods requires careful considerations about the peculiar nature of the geometric inference problem.

Acknowledgments: This work was supported in part by the Ministry of Education, Culture, Sports, Science and Technology, Japan, under a Grant in Aid for Scientific Research C(2) (No. 15500113), the Support Center for Advanced Telecommunications Technology Research, and Kayamori Foundation of Informational Science Advancement.

References

[1] H. Akaike, A new look at the statistical model identification, IEEE Trans. Autom. Control, **16**-6 (1977-12), 716–723.

[2] S. Amari and M. Kawanabe, Information geometry of estimating functions in semiparametric statistical models, *Bernoulli*, **3** (1997), 29–54.

[3] K. Bubna and C. V. Stewart, Model selection techniques and merging rules for range data segmentation algorithms, *Comput. Vision Image Understand.*, **80**-2 (2000), 215–245.

[4] F. Chabat, G. Z. Yang and D. M. Hansell, A corner orientation detector, *Image Vision Comput.*, **17** (1999), 761–769.

[5] K. Cho and P. Meer, Image segmentation form consensus information, *Comput. Vision Image Understand.*, **68**-1 (1997), 72–89.

[6] K. Cho, P. Meer, and J. Cabrera, Performance assessment through bootstrap, *IEEE Trans. Patt. Anal. Mach. Intell.*, **19**-11 (1997), 1185–1198.

[7] W. A. Fuller, *Measurement Error Models*, Wiley, New York, 1987.

[8] H. Gu, Y. Shirai and M. Asada, MDL-based segmentation and motion modeling in a long sequence of scene with multiple independently moving objects, *IEEE Trans. Patt. Anal. Mach. Intell.*, **18**-1 (1996), 58–64.

[9] C. Harris and M. Stephens, A combined corner and edge detector, *Proc. 4th Alvey Vision Conf.*, Aug. 1988, Manchester, U.K., pp. 147–151.

[10] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*, Cambridge University Press, Cambridge, U.K., 2000.

[11] K. Kanatani, *Statistical Optimization for Geometric Computation: Theory and Practice*, Elsevier, Amsterdam, the Netherlands, 1996.

[12] K. Kanatani, Geometric information criterion for model selection, *Int. J. Comput. Vision*, **26**-3 (1998), 171–189.

[13] K. Kanatani, Statistical optimization and geometric inference in computer vision, *Phil. Trans. Roy. Soc. Lond.*, A-**356** (1998), 1303–1320.

[14] K. Kanatani, Cramer-Rao lower bounds for curve fitting, *Graphical Models Image Process.*, **60**-2 (1988), 93–99.

[15] K. Kanatani, Model selection criteria for geometric inference, in A. Bab-Hadiashar and D. Suter (eds.), *Data Segmentation and Model Selection for Computer Vision: A Statistical Approach*, Springer, 2000, pp. 91–115.

[16] K. Kanatani, Model selection for geometric inference, plenary talk, *Proc. 5th Asian Conf. Comput. Vision*, January 2002, Melbourne, Australia, Vol. 1, pp. xxi–xxxii.

[17] K. Kanatani, Motion segmentation by subspace separation: Model selection and reliability evaluation, *Int. J. Image Graphics/*, **2**-2 (2002), 179–197.

[18] Y. Kanazawa and K. Kanatani, Optimal line fitting and reliability evaluation, *IEICE Trans. Inf. & Syst.*, **E79-D**-9 (1996), 1317–1322.

[19] Y. Kanazawa and K. Kanatani, Optimal conic fitting and reliability evaluation, *IEICE Trans. Inf. & Syst.*, **E79-D**-9 (1996), 1323–1328.

[20] Y. Kanazawa and K. Kanatani, Infinity and planarity test for stereo vision, *IEICE Trans. Inf. & Syst.*, **E80-D**-8 (1997), 774–779.

[21] Y. Kanazawa and K. Kanatani, Stabilizing image mosaicing by model selection, in M. Pollefeys, L. Van Gool, A. Zisserman and A. Fitzgibbon (eds.), *3D Structure from Images–SMILE 2000*, Springer, Berlin, 2001, pp. 35–51.

[22] Y. Kanazawa and K. Kanatani, Do we really have to consider covariance matrices for image features? *Proc. 8th Int. Conf. Comput. Vision*, July 2001, Vancouver, Canada, Vol. 2, pp. 586–591.

[23] Y. G. Leclerc, Constructing simple stable descriptions for image partitioning, *Int. J. Comput. Vision*, **3**-1 (1989), 73–102.

[24] Y. Leedan and P. Meer, Heteroscedastic regression in computer vision: Problems with bilinear constraint, *Int. J. Comput. Vision.*, **37**-2 (2000-6), 127–150.

[25] C. Matsunaga and K. Kanatani, Calibration of a moving camera using a planar pattern: Optimal computation, reliability evaluation and stabilization by model selection, *Proc. 6th Euro. Conf. Comput. Vision*, June–July, 2000, Dublin, Ireland, Vol. 2, pp. 595–609.

[26] B. A. Maxwell, Segmentation and interpretation of multicolored objects with highlights, *Comput. Vision Image Understand.*, **77**-1 (2000), 1–24.

[27] S. J. Maybank and P. F. Sturm, MDL, collineations and the fundamental matrix, *Proc. 10th British Machine Vision Conference*, September 1999, Nottingham, U.K., pp. 53–62.

[28] D. D. Morris, K. Kanatani and T. Kanade, Gauge fixing for accurate 3D estimation, *Proc. IEEE Conf. Comput. Vision Pattern Recog.*, December 2001, Kauai, Hawaii, U.S.A., Vol. 2, pp. 343–350.

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

[30] N. Ohta, Motion parameter estimation from optical flow without nuisance parameters, *3rd Int. Workshop on Statistical and Computational Theory of Vision* October 2003, Nice, France: `http://www.stat.ucla.edu/~sczhu/Workshops/SCTV2003.html`

[31] N. Ohta and K. Kanatani, Moving object detection from optical flow without empirical thresholds, *IEICE Trans. Inf. & Syst.*, **E81-D**-2 (1998), 243–245.

[32] T. Okatani and K. Deguchi, Toward a statistically optimal method for estimating geometric relations from noisy data: Cases of linear relations, *Proc. IEEE Conf. Comput. Vision Pattern Recog.*, June 2003, Madison, WI, U.S.A., Vol. 1, pp. 432–439.

[33] D. Reisfeld, H. Wolfson and Y. Yeshurun, Context-free attentional operators: The generalized symmetry transform, *Int. J. Comput. Vision*, **14** (1995), 119–130.

[34] J. Rissanen, Universal coding, information, prediction and estimation, *IEEE Trans. Inform. Theory*, **30**-4 (1984), 629–636.

[35] J. Rissanen, *Stochastic Complexity in Statistical Inquiry*, World Scientific, Singapore, 1989.

[36] J. Rissanen, Fisher information and stochastic complexity, IEEE Trans. Inform. Theory, **42**-1 (1996), 40–47.

[37] C. Schmid, R. Mohr and C. Bauckhage, Evaluation of interest point detectors, *Int J. Comput. Vision*, **37**-2 (2000), 151–172.

[38] S. M. Smith and J. M. Brady, SUSAN—A new approach to low level image processing, *Int. J. Comput. Vision*, **23**-1 (1997-5), 45–78.

[39] Y. Sugaya and K. Kanatani, Outlier removal for feature tracking by subspace separation, *IEICE Trans. Inf. & Syst.*, **E86**-D (2003-6), 1095–1102.

[40] P. H. S. Torr, An assessment of information criteria for motion model selection, *Proc. IEEE Conf. Comput. Vision Patt. Recog.*, June 1997, Puerto Rico, pp. 47–53.

[41] P. H. S. Torr, Geometric motion segmentation and model selection, *Phil. Trans. Roy. Soc. Lond.*, A-**356** (1998), 1321–1340.

[42] P. H. S. Torr, Model selection for structure and motion recovery from multiple images, in A. Bab-Hadiashar and D. Suter (eds.), *Data Segmentation and Model Selection for Computer Vision: A Statistical Approach*, Springer, 2000, pp. 143–183.

[43] P. H. S. Torr, A. FitzGibbon and A. Zisserman, Maintaining multiple motion model hypotheses through many views to recover matching and structure, *Proc. 6th Int. Conf. Comput. Vision*, January 1998, Bombay, India, pp. 485–492.

[44] P. H. S. Torr and A. Zisserman, Concerning Bayesian motion segmentation, model averaging, matching and the trifocal tensor, *Proc. 6th Euro. Conf. Comput. Vision*, June–July, 2000, Dublin, Ireland, Vol. 1, pp. 511–528.

[45] Iman Triono, N. Ohta and K. Kanatani, Automatic recognition of regular figures by geometric AIC, *IEICE Trans. Inf. & Syst.*, **E81-D**-2 (1998), 246–248.