２０００年情報論的学習理論ワークショップ
2000 Workshop on Information-Based Induction Sciences (IBIS2000)
Izu, Japan, July 17-18, 2000.

Model Selection in Statistical Inference and Geometric Fitting

Kenichi Kanatani*

Abstract: Taking line fitting to points in two dimensions as a typical example, we point out the inherent difference between *statistical inference* and *geometric fitting*. We describe their *duality* in the sense that the *asymptotic* properties of statistical inference in the limit of an infinite number of observations hold for geometric fitting in the limit of *infinitesimal perturbations*. We contrast *stochastic model selection* with *geometric model selection* and describe the difference between *Akaike's AIC* and the *geometric AIC* in their derivations.

1. Introduction

Model selection has been one of the central themes in the traditional statistical domains such as agriculture, medicine, politics, and economics. Recently, interest in model selection for computer vision and robotics applications has been quickly growing, partly due to the introduction of the *geometric AIC* by Kanatani [8, 9]. Already, it has been applied to a wide range of practical problems [10, 11, 12, 14, 17, 18, 19].

Since the geometric AIC appears very similar to Akaike's *AIC* (*Akaike information criterion*) [1], there has been some confusion about their distinction. This confusion stems from the concept of "observation", which plays different roles in the traditional framework of statistics and in computer vision and robotics problems. To clarify this is the main purpose of this paper.

We first take line fitting to points in two dimensions as a typical example and point out the inherent difference between *statistical inference* and *geometric fitting*. We describe their *duality* in the sense that the *asymptotic* properties of statistical inference in the limit of an infinite number of observations hold for geometric fitting in the limit of *infinitesimal perturbations*. We contrast *stochastic model selection* with *geometric model selection* and describe the difference between *Akaike's AIC* and the *geometric AIC* in their derivations.

2. Line Fitting

Suppose we want to fit a line $Ax + By + C = 0$ to N points (x_α, y_α), $\alpha = 1, ..., N$ (Fig. 1). We assume that each point is perturbed from its true position $(\bar{x}_\alpha, \bar{y}_\alpha)$ by random noise subject to an independent Gaussian distribution of mean 0 and variance σ^2 (unknown). The likelihood of the data is

$$P = \prod_{\alpha=1}^{N} \frac{e^{-(x_\alpha-\bar{x}_\alpha)^2/2\sigma^2-(y_\alpha-\bar{y}_\alpha)^2/2\sigma^2}}{2\pi\sigma^2} = \frac{e^{-J/2\sigma^2}}{(2\pi\sigma^2)^N}, \tag{1}$$

where

$$J = \sum_{\alpha=1}^{N} \Big((x_\alpha - \bar{x}_\alpha)^2 + (y_\alpha - \bar{y}_\alpha)^2 \Big). \tag{2}$$

* Department of Computer Science, Gunma University, Kiryu, Gunma, 376-8515 Japan, Tel: (0277)30-1841, Fax: (0277)30-1801, E-mail: kanatani@cs.gunma-u.ac.jp.

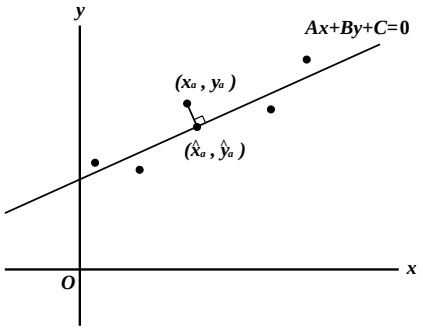

Figure 1: Optimal line fitting.

Maximum likelihood estimation (*MLE*) is to compute the solution, called the *maximum likelihood estimator* (*ML-estimator*), that maximizes the likelihood P with respect to the $2N + 4$ unknowns $\{(\bar{x}_\alpha, \bar{y}_\alpha)\}$, A, B, C, and σ^2, which are constrained by

$$A\bar{x}_\alpha + B\bar{y}_\alpha + C = 0. \tag{3}$$

Maximizing eq. (1) is equivalent to minimizing the negative logarithmic likelihood

$$-\log P = \frac{J}{2\sigma^2} + N \log 2\pi\sigma^2 \tag{4}$$

subject to the constraint (3). Evidently, minimization can be carried out for σ^2 and for the remaining variables separately. Differentiating eq. (4) with respect to σ^2 and letting the result be zero, we obtain the following ML-estimator:

$$\hat{\sigma}^2 = \frac{\hat{J}}{2N}. \tag{5}$$

Here, $\hat{J}$ is the *residual* (i.e., the minimum) of the function J. It can be computed as follows. Introducing Lagrange multipliers λ_α for the constraint (3), differentiating $J/2 + \sum_\alpha \lambda_\alpha (A\bar{x}_\alpha + B\bar{y}_\alpha + C)$ with respect to $\bar{x}_\alpha$ and $\bar{y}_\alpha$, and letting the result be zero, we obtain

$$x_\alpha - \bar{x}_\alpha = \lambda_\alpha A, \qquad y_\alpha - \bar{y}_\alpha = \lambda_\alpha B. \tag{6}$$

These equations imply that the line segment connecting the two points (x_α, y_α) and $(\bar{x}_\alpha, \bar{y}_\alpha)$ should be orthogonal to the line $Ax + By + C = 0$. Since $(\bar{x}_\alpha, \bar{y}_\alpha)$

should be be on that line, we obtain λ_α in the form

$$\lambda_\alpha = \frac{Ax_\alpha + By_\alpha + C}{A^2 + B^2}. \tag{7}$$

From eq. (6), we obtain the ML-estimators $(\hat{x}_\alpha, \hat{y}_\alpha)$ of $(\bar{x}_\alpha, \bar{y}_\alpha)$ in the form

$$\begin{aligned} \hat{x}_\alpha &= x_\alpha - \frac{(Ax_\alpha + By_\alpha + C)A}{A^2 + B^2}, \\ \hat{y}_\alpha &= y_\alpha - \frac{(Ax_\alpha + By_\alpha + C)B}{A^2 + B^2}. \end{aligned} \tag{8}$$

These define the foot of the perpendicular line drawn from the data point (x_α, y_α) to the line $Ax + By + C = 0$ (Fig. 1). Substituting eqs. (8) for $(\bar{x}_\alpha, \bar{y}_\alpha)$ in eq. (2), we can rewrite the function J as

$$J = \sum_{\alpha=1}^{N} \frac{(Ax_\alpha + By_\alpha + C)^2}{A^2 + B^2}. \tag{9}$$

This is a homogeneous equation of degree zero in A, B, and C, reflecting the fact that $Ax + By + C = 0$ defines the same line if A, B, and C are multiplied by an arbitrary nonzero constant. This indeterminacy can be removed (except for sign) by imposing the normalization $A^2 + B^2 = 1$.

Differentiating eq. (9) with respect to C and letting the result be 0, we see that the line $Ax + By + C = 0$ should pass through the "centroid" of the data points. If we take a new coordinate system with its origin at the centroid, the function J becomes a quadratic form in A and B. Hence, the orientation of the line to be fitted is given as the eigenvector of the "moment (of inertia) matrix" with the largest eigenvalue [7].

3. Statistical Inference from Geometric Data

3.1 Inconsistency of MLE

The residual $\hat{J}$ is the sum of the square distances from the data points to the line. Because the distance from each point to the line is a Gaussian random variable of mean 0 and variance σ^2, the quotient $\hat{J}/\sigma^2$ is a χ^2 random variable. Since there are N points, its degree of freedom should be N for the true line, but we have fitted a line optimally by adjusting its two degrees of freedom. Thus, $\hat{J}/\sigma^2$ is subject to a χ^2 distribution with $N - 2$ degrees of freedom. The expectation of a χ^2 random variable is equal to its degree of freedom, so the residual $\hat{J}$ has expectation

$$E[\hat{J}] = (N - 2)\sigma^2, \tag{10}$$

where $E[\cdot]$ denotes expectation. Hence, the ML-estimator $\hat{\sigma}^2$ given by eq. (5) has expectation

$$E[\hat{\sigma}^2] = \frac{N - 2}{2N}\sigma^2. \tag{11}$$

In the asymptotic limit $N \to \infty$ of an infinite number of points, we have $E[\hat{\sigma}^2] \to \sigma^2/2$, which means that $\hat{\sigma}^2$ is a very poor estimate of σ^2. It is known in statistics that an ML-estimator converges to its true value in the asymptotic limit of an infinite number of data under a very mild condition (which is satisfied if the noise distribution is Gaussian). This property is called the *consistency* of MLE. How can we resolve this paradox?

3.2 Observation vs. the number of data

Carefully examining the origin of the inconsistency of $\hat{\sigma}^2$, we can trace it back to the inconsistency of the ML-estimators $(\hat{x}_\alpha, \hat{y}_\alpha)$: they do not converge to their true positions $(\bar{x}_\alpha, \bar{y}_\alpha)$ as $N \to \infty$. If they did, the residual $\hat{J}$ would approach the sum of the square distances of the data points to their true positions, whose expectation is $2N\sigma^2$. Hence, eq. (5) would converge to σ^2.

The reason why $(\hat{x}_\alpha, \hat{y}_\alpha)$ do not converge to $(\bar{x}_\alpha, \bar{y}_\alpha)$ when the number N of data is increased is evident: the added points do not have any information about the displacements of the rest of the points *along* the fitted line, because each point is independent of the others except that their true positions should be collinear. Thus, $(\hat{x}_\alpha, \hat{y}_\alpha)$ are pathological parameters sometimes referred to as *nuisance parameters*.

Then, is the proposition that an ML-estimator converges to its true value as $n \to \infty$ as proved in textbooks wrong? If we carefully examine the proof, we find that n in the proposition means the *number of observations*, not the number of data.

These two usually coincide in the traditional framework of statistics, because conquering randomness by repeated observations is the paradigm of statistics. It follows that each datum is regarded as a sample from the same probability distribution (*stochastic model*) involving unknown parameters. It is intuitively evident that ML-estimators of these unknowns converge to their true values as the number n of observations (= the number of data) increases.

In line fitting, on the other hand, the unknown parameters are $(\bar{x}_\alpha, \bar{y}_\alpha)$, A, B, C, and σ^2, which satisfy the constraint (3). Also, multiplication of A, B, and C by a nonzero constant does not alter the problem. Hence, the effective number of unknowns is $N + 3$. From this stochastic model, we observe *one* sample $\{(x_\alpha, y_\alpha)\}_{\alpha=1}^{N}$.

If we add a new point $(\bar{x}_{N+1}, \bar{y}_{N+1})$, we observe *one* sample $\{(x_\alpha, y_\alpha)\}_{\alpha=1}^{N+1}$ from a new model parameterized by $N + 4$ unknowns. The number of unknowns increases each time a new point is added, and we observe *one* sample from a new model with an increased number of unknowns.

What is required in computer vision and robotics is inference from *one pattern*, however complex it may be. This is a big contrast to the traditional domain of statistics, where one is required to infer about a random phenomenon from repeated observations.

Since the number n of observations is always 1 in geometric fitting, the asymptotic properties of MLE described in textbooks on statistics never materialize. In other words, it does not make sense to consider the "asymptotic" properties of line fitting in the limit N

$\to \infty$ of an infinite number of data points.

4. Duality of Statistical Estimation

Statisticians have tried to reconcile the anomaly of nuisance parameters by altering the problem to fit in the traditional framework. A typical expedient for this is to introduce artificial *hidden variables*. In line fitting, for instance, the true positions $\{(\bar{x}_\alpha, \bar{y}_\alpha)\}$ are regarded as a *sample* from an unknown distribution parameterized by hidden variables. The task now is to estimate these hidden variables, not $\{(\bar{x}_\alpha, \bar{y}_\alpha)\}$ themselves. As a result, the number N of data can be identified with the number n of observations, and one can study the asymptotic behavior in the limit $N \to \infty$ in the traditional framework. In statistical inference, this type of approach is generalized and called *semi-parametric models* [3, 4],

In computer vision and robotics, however, the number N of data is not very large, and the distribution of the true positions $\{(\bar{x}_\alpha, \bar{y}_\alpha)\}$ depend on the scene or objects that we are viewing; no information is available a priori about their distribution. We now present a formalism that does not introduce hidden variables or semi-parametric models.

Let us do the following thought experiment. Suppose we observe the N points to fit a line multiple times. Of course, we always observe the same points as long as the imaging devices and the image processing operations are the same, but let us hypothetically assume that we use different devices and operations with different characteristics each time. Let $(x_\alpha^{(a)}, y_\alpha^{(a)})$ be the position of the αth point in the ath observation. Compute the averages

$$x_\alpha = \frac{1}{n}\sum_{a=1}^{n} x_\alpha^{(a)}, \qquad y_\alpha = \frac{1}{n}\sum_{a=1}^{n} y_\alpha^{(a)}. \tag{12}$$

Then, (x_α, y_α) converge to $(\bar{x}_\alpha, \bar{y}_\alpha)$ as the number n of hypothetical observations increases. Hence, the asymptotic properties of MLE materialize.

On the other hand, if the variance of the data due to using different devices and operations is σ^2, the variance of the averages given by eq. (12) is σ^2/n. This means that increasing the number n of hypothetical observations *effectively reduces the noise level to* $1/n$. This is equivalent to using devices and operations of higher quality. Thus, *all the properties of MLE that hold in the asymptotic limit of an infinite number of observations hold for infinitesimal perturbations* [8].

Statistical inference is to conquer randomness by *repeated observations*. Hence, we desire an estimation method whose accuracy increases rapidly as the number of observations increases, i.e., *a method that requires a smaller number of observations for warranting specified accuracy*. In geometric fitting, in contrast, we do a *single measurement* using an electronic device. Hence, we desire an estimation method whose accuracy increases rapidly as the noise level decreases, i.e., *a method that tolerates a larger amount of noise for warranting specified accuracy* [8]. This is the inherent *duality* between statistical inference and geometric fitting in regard to the concept of "observation".

5. Modeling and Inference

5.1 Stochastic models of noise

In statistical inference, the probability density of data is defined to be the distribution of their values that are observed by repeated observations under the same condition. Since exactly the same condition can never be realized, statisticians focus on parameters that they believe are important, called *control parameters*, and define the fluctuations due to the other (or *unmodeled*) factors to be *noise*, which is expected to be Gaussian if the other factors are large in number and independent of each other (the *central limit theorem*). The resulting description is called a (*stochastic*) *model*. The purpose of statistical inference is to *validate or invalidate such a model by observing data*.

If we observe an image, on the other hand, we have only one sample, so the traditional framework of statistical inference does not apply. If we detect feature points such as corners and vertices by applying image processing operations, the detected positions may not be the true positions because the image is incomplete due to digitization, quantization, and illumination variations. Moreover, any image processing operation is an approximate process. We call the resulting uncertainty *noise* and model it in terms of probability distributions.

But however many times we repeat the image processing operations, the result is always the same as long as we use the same image and the same image processing operations. This means that we can never validate the noise distribution by experiments; the uncertainty description ought to be empirical. For example, if we expect the errors in the true positions to be about 2 pixels by experience, we assume that the noise is subject to a Gaussian distribution of mean 0 and standard deviation 2 and regard the observed data as a sample from it. In this sense, the noise model is intrinsically Bayesian. Our goal is *not* to validate or invalidate this model but to *compute the properties of the object that we see in the image as accurately as possible using that knowledge*.

Although the noise model is empirical, we can "measure" the characteristics of the noise quantitatively. For example, if the gray level of a feature point is very close to that of the surrounding points, like a white spot in a white region, we can regard that point as having a large covariance. The magnitude of the covariance can be estimated from the curvature of the residual surface of template matching based on local correlation [5, 15, 16]. Using such a method, we obtain a small covariance in a textured region and a large covariance in a region of almost uniform gray levels; for a point on an "edge" (i.e., object boundary), we obtain a distribution elongated along that edge.

However, the "covariance" measured by such a means does not have a true probabilistic meaning; it is simply a convenient measure of uncertainty implied

by the gray level distributions. The use of such covariance is justified, because doing statistical optimization based on it means effectively discarding the vague points and focusing only on clearly visible points.

5.2 Models for inference

In statistical inference, a "model" is a stochastic description of the imperfection of our description capability, and its validation or invalidation is the goal.

In contrast, two types of models coexist in computer vision and robotics applications. One is the stochastic model of the noise resulting from the imperfection of our sensing capability. This is, as we have pointed out, simply an assumption introduced by experience for the convenience of analysis; we are not interested in validating or invalidating it.

What we are interested in is to understand the scene that the data imply. To this end, we introduces various possibilities as "models", e.g., we may hypothesize that the feature points detected in the image are collinear. In order to distinguish such models from stochastic models of noise, let us call them *geometric models*. This distinction plays a key role in the subsequent formulation of geometric model selection.

6. Geometric Fitting

Let N m-dimensional vectors $\boldsymbol{x}_1, ..., \boldsymbol{x}_N$ be given as data. Each $\boldsymbol{x}_\alpha$ is regarded as perturbed from its true value $\bar{\boldsymbol{x}}_\alpha$ by $\Delta\boldsymbol{x}_\alpha$, which we assume is an independent Gaussian random variable of mean $\boldsymbol{0}$ and covariance matrix $V[\boldsymbol{x}_\alpha]$. The true value $\bar{\boldsymbol{x}}_\alpha$ is assumed to satisfy r constraint equations

$$F^{(k)}(\bar{\boldsymbol{x}}_\alpha, \boldsymbol{u}) = 0, \qquad k = 1, ..., r, \tag{13}$$

parameterized by a p-dimensional vector $\boldsymbol{u}$. We call the domain of the vectors $\{\boldsymbol{x}_\alpha\}$ the *data space* and the domain of the vector $\boldsymbol{u}$ the *parameter space*; we call the number r of the constraint equations the *rank* of the constraint. The r equations $F^{(k)}(\boldsymbol{x}, \boldsymbol{u}) = 0$ define a manifold $\mathcal{S}$ of dimension $d = m - r$ (*codimension* r) in the m-dimensional data space; the constraint (13) constrains the true values $\{\bar{\boldsymbol{x}}_\alpha\}$ to be on that manifold. Our task is to estimate the parameter $\boldsymbol{u}$ from the noisy data $\{\boldsymbol{x}_\alpha\}$.

We assume that the covariance matrix $V[\boldsymbol{x}_\alpha]$ is known only *up to scale* and write

$$V[\boldsymbol{x}_\alpha] = \epsilon^2 V_0[\boldsymbol{x}_\alpha]. \tag{14}$$

We call the unknown magnitude ϵ the *noise level* and the known matrix $V_0[\boldsymbol{x}_\alpha]$ the *normalized covariance matrix*, which qualitatively characterizes the noise.

This assumption can be justified by two reasons. First, the absolute magnitude of noise is very difficult to predict a priori in many practical situations, while qualitative behavior, such as isotropy/anisotropy and homogeneity/inhomogeneity, of noise can be estimated in many ways; e.g., the template matching method mentioned earlier determines covariance only up to scale.

Another reason is that the scale of the covariance does not affect the ML-estimator, which can be obtained by minimizing the sum of the square *Mahalanobis distances* from the data points to their true positions associated with the normalized covariance matrices $V_0[\boldsymbol{x}_\alpha]$

$$J = \sum_{\alpha=1}^{N} (\boldsymbol{x}_\alpha - \bar{\boldsymbol{x}}, V_0[\boldsymbol{x}_\alpha]^{-1}(\boldsymbol{x}_\alpha - \bar{\boldsymbol{x}})) \tag{15}$$

subject to the constraint (13). Here and hereafter, the inner product of vectors $\boldsymbol{a}$ and $\boldsymbol{b}$ is denoted by $(\boldsymbol{a}, \boldsymbol{b})$.

The above formulation can be further generalized to admit redundancies in eq. (13) and constraints on the data space and the parameter space by incorporating projection operations and generalized inverses [8].

In minimizing eq. (15), we can eliminate the constraint (13) by introducing liner approximation and Lagrange multiplies, assuming that noise is small. After some manipulations, eq. (15) can be written in the form

$$J = \sum_{\alpha=1}^{N} \sum_{k,l=1}^{r} W_\alpha^{(kl)} F^{(k)}(\boldsymbol{x}_\alpha, \boldsymbol{u}) F^{(l)}(\boldsymbol{x}_\alpha, \boldsymbol{u}), \tag{16}$$

where $W_\alpha^{(kl)}$ is the (kl) element of the inverse of the $r \times r$ matrix whose (kl) element equals $(\nabla_{\mathbf{x}} F^{(k)}, V[\boldsymbol{x}_\alpha] \nabla_{\mathbf{x}} F^{(l)})$.

The solution $\hat{\boldsymbol{u}}$ that minimizes eq. (16) does not depend on the noise level ϵ. However, its covariance matrix $V[\hat{\boldsymbol{u}}]$ depends on ϵ. Evidently, we have $V[\hat{\boldsymbol{u}}] \to \boldsymbol{O}$ as $\epsilon \to 0$. Moreover, it can be proved that $V[\hat{\boldsymbol{u}}]$ attains a theoretical accuracy bound if terms of order $O(\epsilon^4)$ are ignored [8].

This is precisely the dual statement to the well known fact about MLE in statistical inference: the covariance of ML-estimators not only converges to zero in the asymptotic limit $n \to \infty$ for the number n of observations (*consistency*) but also attains the *Cramer-Rao lower bound* if terms of order $O(1/n^2)$ are ignored (*asymptotic efficiency*).

7. Models and Model Selection

7.1 Stochastic models vs. geometric models

Statisticians are interested in explaining random phenomena, so they describe observed data by a parameterized expression that combines a deterministic part and random noise. The goal of statistical inference is to estimate the model parameter $\boldsymbol{\theta}$, which consists of the coefficients of the equations that describe the deterministic mechanism and *the characteristics of the random noise* such as means and variances. In abstract terms, one wants to estimate the model parameter $\boldsymbol{\theta}$ by observing data that are supposedly sampled from a parameterized probability density $p(\boldsymbol{x}|\boldsymbol{\theta})$. If there exist multiple candidates $p_1(\boldsymbol{x}|\boldsymbol{\theta}_1)$, $p_2(\boldsymbol{x}|\boldsymbol{\theta}_2)$, ..., each candidate $p_k(\boldsymbol{x}|\boldsymbol{\theta}_k)$ is called a (*stochastic*) *model*. Mathematically, it is a point in the function space [2, 13]. Choosing an appropriate (stochastic) model by observing the data is called (*stochastic*) *model selection*.

On the other hand, geometric fitting aims to find *an implicit relationship among observed data*. To this end, we introduce a constraint $\boldsymbol{F}(\boldsymbol{x}, \boldsymbol{u}) = 0$ that should be satisfied in the absence of noise. The goal is to estimate the model parameter $\boldsymbol{u}$ by observing the data. Here, there are two crucial distinctions from the statistical inference problem. First, the constraint equation $\boldsymbol{F}(\boldsymbol{x}, \boldsymbol{u}) = 0$ *cannot be solved for the data variable* $\boldsymbol{x}$. Second, the model parameter $\boldsymbol{u}$ consists only of the coefficients of the equations that describe the geometric structure of the problem; *it does not contain the characteristics of the random noise*. If there exist multiple candidates $\boldsymbol{F}_1(\boldsymbol{x}, \boldsymbol{u}_1) = \boldsymbol{0}$, $\boldsymbol{F}_2(\boldsymbol{x}, \boldsymbol{u}_2) = \boldsymbol{0}$, ... for the constraint, each candidate $\boldsymbol{F}_k(\boldsymbol{x}, \boldsymbol{u}_k) = \boldsymbol{0}$ is called a (*geometric*) *model*. Mathematically, it is a manifold in the data space [8]. Choosing an appropriate (geometric) model by observing the data is called (*geometric*) *model selection* [8, 9].

7.2 Akaike's AIC

A naive idea for stochastic model selection is to fit each candidate model to the data and choose the one that has the smallest residual. This fails, however, because a model that has a larger number of parameters always fits better than a model with a smaller number of parameters. Akaike's idea is to fit a model to the data and validate it by computing the "expected residual" for hypothetical data that have the same statistical properties as the observed data but are independent of them. This idea is similar to the methods known as *cross-validation*, *jack-knife*, and *bootstrap* [6], for which "validation data" are provided independently of the "learning data" to which the model is fitted.

Thus, the basic principle of Akaike essentially does not involve any asymptotic properties. However, the use of asymptotic properties leads to a simple criterion known as (*Akaike's*) *AIC* [1]. Akaike employed the following two asymptotic approximations:

- The ML-estimator converges to its true value as the number of observations increases (the *law of large number*).
- The ML-estimator asymptotically obeys a Gaussian distribution as the number of observations increases (the *central limit theorem*).

Using these approximations and noting that a quadratic form in Gaussian random variables is subject to a χ^2 distribution, whose expectation equals its degree of freedom, Akaike evaluated the expected residual of the logarithmic likelihood and derived his AIC.

7.3 Geometric AIC

Since Akaike's AIC does not essentially depend on asymptotic analysis, his idea can immediately be applied to geometric fitting. Kanatani adopted his idea of evaluating the "expected residual" for hypothetical data that have the same statistical properties as the observed data but are independent of them. Let $\{\hat{\boldsymbol{x}}_\alpha\}$ be the ML-estimators of $\{\bar{\boldsymbol{x}}_\alpha\}$ computed from the data $\{\boldsymbol{x}_\alpha\}$. The residual of this estimation is

$$\hat{J} = \sum_{\alpha=1}^{N} (\boldsymbol{x}_\alpha - \hat{\boldsymbol{x}}_\alpha, V_0[\boldsymbol{x}_\alpha]^{-1}(\boldsymbol{x}_\alpha - \hat{\boldsymbol{x}}_\alpha)). \tag{17}$$

Let $\{\boldsymbol{x}^*_\alpha\}$ be hypothetical data that have the same statistical properties but independent of them. The goodness of the model is measured by

$$\hat{J}^* = \sum_{\alpha=1}^{N} (\boldsymbol{x}^*_\alpha - \hat{\boldsymbol{x}}_\alpha, V_0[\boldsymbol{x}_\alpha]^{-1}(\boldsymbol{x}^*_\alpha - \hat{\boldsymbol{x}}_\alpha)). \tag{18}$$

This has the following expectation [8, 9]:

$$E^*[E[\hat{J}^*]] = E[\hat{J}] + 2(Nd + p)\epsilon^2 + O(\epsilon^4). \tag{19}$$

Here, $E[\cdot]$ and $E^*[\cdot]$ denote expectation with respect to $\{\boldsymbol{x}_\alpha\}$ and $\{\boldsymbol{x}^*_\alpha\}$, respectively, and $d = m - r$ is the dimension of the manifold $\mathcal{S}$ defined by the constraint (13).

Eq. (19) implies that fitting the model to the data decreases the residual $\hat{J}$ by $2(Nd + p)\epsilon^2$ on average as compared with the residual $\hat{J}^*$ for independent data. So, we define the *geometric AIC* as follows [8, 9]:

$$\text{G-AIC} = \hat{J} + 2(Nd + p)\epsilon^2. \tag{20}$$

The derivation of eq. (19) is based on infinitesimal perturbation analysis; no asymptotic approximation is involved. Yet, the final form appears very similar to Akaike's AIC in that it essentially consists of the residual term and twice the degree of freedom of the model. This is because the derivation of eq. (19) makes use of the following two *perturbation approximations*:

- The ML-estimator converges to its true value as the noise decreases.
- Since the noise is *assumed* to be Gaussian, ML-estimators subject to a linear constraint are also Gaussian. If the constraint is nonlinear, linear approximation is justified in the neighborhood of the solution in the limit of infinitesimal perturbation.

Using these and noting that a quadratic form in Gaussian random variables is subject to a χ^2 distribution, whose expectation equals its degree of freedom, we can obtain eq. (19) after lengthy manipulations [8, 9].

Thus, the derivations of Akaike's AIC and the geometric AIC are opposite to each other: the former is obtained in the asymptotic limit; the latter in the limit of infinitesimal perturbation. Akaike also gave his AIC an interpretation as the asymptotic *Kullback-Leibler distance* between probability densities. Because this interpretation is inherently asymptotic, it does not apply to geometric fitting.

8. Noise Level Estimation

In statistical inference, a stochastic model describes which part is deterministic and which part is random noise. Hence, the noise variance is also a model parameter to be estimated by MLE. In geometric fitting,

on the other hand, the noise variance is *not* a model parameter, as we have pointed out.

We have already explained how the ML-estimator (5) is wrong and why it is wrong. A correct estimator is obtained from eq. (10) in the form $\hat{\sigma}^2 = \hat{J}/(N-2)$, which is an unbiased estimator. In general, an unbiased estimator of the square noise level ϵ^2 is obtained from eq. (15) in the form

$$\hat{\epsilon}^2 = \frac{\hat{J}}{rN-p}. \tag{21}$$

Geometrically, $\hat{J}$ is the sum of the square Mahalanobis distances from the data points to the manifold defined by eq. (13). Since the codimension (the dimension of the orthogonal directions) is r, the residual $\hat{J}$ of N points would have expectation $rN\epsilon^2$ for the true the manifold. For the optimally fitted manifold with p parameters adjusted, the total degree of freedom decreases to $rN-p$.

However, eq. (21) gives a correct estimation only when the assumed model is correct. What if the validity of the model is questionable? Kanatani [8, 9] resolved this issue by applying model selection to *degeneracy detection*.

3-D computation from image and sensor data is made possible by our a priori knowledge about the scene, such as the rigidity of the motion and the planarity of the object, which imposes a constraint in the form of eq. (13). However, the computation fails in degenerate cases, e.g., when the amount of motion is zero. In not exactly but nearly degenerate cases, the computation is possible but is sensitively affected by noise and very unstable. In such a case, the computation can be stabilized by automatically detecting degeneracy by model selection and applying an appropriate model that incorporates degeneracy [10, 11, 12, 14, 19].

Degeneracy means that some extra conditions (e.g., some quantity is zero) are added to the constraint (13). Since the constraint (13) itself still holds even in degeneracy, *the general model holds whether the degenerate model holds or not*. This means that ϵ^2 can be estimated from the residual $\hat{J}$ of the general model by using eq. (21) even for testing the validity of the degenerate model.

In statistical inference, the noise variance is estimated for each (stochastic) model, because the noise variance is a model parameter. In geometric fitting, the noise is due to imperfection of our sensing capability. Hence, the noise variance should be the same for all the (geometric) models.

Note that multiplying Akaike's AIC by a positive expression in the number n of data does not affect relative comparison of models, because n is common to all the models. Similarly, multiplying the geometric AIC by a positive expression in the noise level ϵ does not affect relative comparison of models, because ϵ is common to all the models.

9. Concluding Remarks

Taking line fitting to points in two dimensions as a typical example, we have pointed out the inherent difference between statistical inference and geometric fitting: statistical inference is to reason from repeated data; geometric fitting is to reason from a single observation. We have described their duality in the sense that the asymptotic properties of statistical inference in the limit of an infinite number of observations hold for geometric fitting in the limit of infinitesimal perturbations. It follows that noise estimation for the stochastic model, which describes the mechanism of random phenomena, is different from that for the geometric model, which describes the constraint manifold in the data space. We have contrasted stochastic model selection with geometric model selection and described the difference between Akaike's AIC and the geometric AIC in their derivations.

References

[1] H. Akaike, A new look at the statistical model identification, *IEEE Trans. Automatic Control*, **16**-6 (1974), 716–723.

[2] S. Amari, *Differential-Geometric Methods in Statistics*, Springer, Berlin, 1985.

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

[4] P.J. Bickel, C. A. J. Klassen, Y. Ritov and J. A. Wellner, *Efficient and Adaptive Estimation for Semiparametric Models*, Johns Hopkins University Press, Baltimore, MD, 1994.

[5] W. Förstner, Reliability analysis of parameter estimation in linear models with applications to mensuration problems in computer vision, *Comput. Vision Graphics Image Process.*, **40** (1987), 273–310.

[6] B. Efron and R. J. Tibshirani, *An Introduction to Bootstrap*, Chapman-Hall, New York, 1993.

[7] K. Kanatani, *Geometric Computation for Machine Vision*, Oxford University Press, Oxford, U.K., 1993.

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

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

[10] K. Kanatani and Y. Kanazawa, Stabilizing image mosaicing by the geometric AIC, *Proc. 2nd Workshop on Information-Based Induction Sciences*, August 1999, Izu, Japan, pp. 181–186.

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

[12] C. Matsunaga and K. Kanatani, Stabilizing moving camera calibration from images by the geometric AIC, *Proc. 4th Asian Conf. Comput. Vision*, January 2000, Taipei, Taiwan, pp. 1168–1173.

[13] M. K. Murray and J. W. Rice, *Differential Geometry and Statistics*, Chapman-Hall, London, 1993.

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

[15] J. Shi and C. Tomasi, Good features to track, *Proc. IEEE Conf. Comput. Vision Patt. Recogn.*, June 1994, Seattle, WA, pp. 593–600.

[16] A. Singh, An estimation-theoretic framework for image-flow computation, *Proc. 3rd Int. Conf. Comput. Vision*, December 1990, Osaka, Japan, pp.168–177.

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

[18] P. H. S. Torr, A. W. Fitzgibbon and A. Zisserman, The problem of degeneracy in structure and motion recovery from uncalibrated image sequences, *Int. J. Comput. Vision*, **32**-1 (1999), 27–44.

[19] 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.