Proc. 5th Asian Conference on Computer Vision, January 2002, Melbourne, Australia, Vol. 1, pp. xxi–xxxii.

Model Selection for Geometric Inference

Kenichi Kanatani

[*] Department of Information Technology, Okayama University, Okayama 700-8530 Japan

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

Abstract

Contrasting "geometric fitting", for which the noise level is taken as the asymptotic variable, with "statistical inference", for which the number of observations is taken as the asymptotic variable, we give a new definition of the "geometric AIC" and the "geometric MDL" as the counterparts of Akaike's AIC and Rissanen's MDL. We discuss various theoretical and practical problems that emerge from our analysis. Finally, we experimentally show that the geometric AIC and the geometric MDL have very different characteristics.

1. Introduction

The problem of inferring the geometric structure of the scene from noisy data is one of the central themes of computer vision. This problem has been generalized in abstract terms as *geometric fitting*, for which a general theory of statistical optimization has been developed [8, 9]. Also, the *geometric AIC* has been proposed for model selection [8, 11, 13] and applied to many problems [10, 14, 15, 18, 19, 22, 26, 38]. Similar but different criteria have also been proposed [33, 34, 35, 36, 37].

The geometric AIC was motivated by Akaike's *AIC* (*Akaike information criterion*) [1]. Naturally, interests arose about using Rissanen's *MDL* (*minimum description length*) [28, 29], since it is regarded by many as superior to Akaike's AIC for statistical inference. It is anticipated that a criterion like Rissanen's MDL would outperform the geometric AIC for geometric fitting, too.

In the past, Rissanen's MDL often appeared in the literature of computer vision, but the use was limited to such applications that have the form of standard statistical inference such as linear/nonlinear regression [3, 25]. Also, many MDL-like criteria were introduced, but often the solution having a shorter description length was simply chosen with an arbitrary definition of the complexity [7, 21, 24].

The main contributions of this paper are as follows:

1. We give a new definition of the *geometric AIC* and show that the final form agrees with the already proposed form. This new definition clarifies the ambiguities that existed in the original derivation [8, 11, 13]. At the same time, this definition refutes the existing misunderstanding that the geometric AIC and Akaike's AIC are the same thing.
2. We give a new definition of the *geometric MDL* using the logic introduced by Rissanen [28, 29, 30]. The final form is slightly different from the already proposed form [22]. Again, this new definition makes it clear how the geometric MDL is different from Rissanen's MDL. It also reveals some problems that have been overlooked so far.
3. We experimentally test if the geometric MDL really outperforms the geometric AIC as anticipated. *Our conclusion is negative*. We also show that these two criteria have very different characteristics.

The basic principle behind our approach is that we take the *noise level* as what we call the *asymptotic variable* and define *geometric model selection* that corresponds to *stochastic model selection*, where the number of observations is taken as the asymptotic variable. In this sense these two formalisms are dual to each other, and in this light the geometric MDL is "dual" to Rissanen's MDL. The similarity between the geometric AIC and Akaike's AIC can be viewed as "self-duality".

There have been heated arguments among statisticians and information theorists for and against Akaike's AIC, Rissanen's MLD, and the philosophies behind them. Also, many similar criteria purported to be better than them have been proposed. In this paper, *we neither endorse either of Akaike's AIC and Rissanen's MLD nor justify their derivations and philosophies behind them*. We regard them simply as they are and focus only on the question of *how they should be redefined in the framework of geometric fitting*.

In Sec. 2, we formulate geometric fitting as constraint satisfaction of geometric data in the presence of noise, taking the noise level as the asymptotic variable. In Sec. 3 and 4, we define the geometric AIC and the geometric MDL as counterparts of Akaike's AIC and Rissanen's MDL. We also discuss various theoretical and practical problems that emerge from our analysis. In Sec. 6, we experimentally test if the geometric MDL really outperforms the geometric AIC. In Sec. 7, our conclusion is given.

2. Definitions

2.1 Geometric fitting

Given N data $\boldsymbol{x}_1$, ..., $\boldsymbol{x}_N$, which are m-dimensional vectors, we view each $\boldsymbol{x}_\alpha$ as perturbed from its true value $\bar{\boldsymbol{x}}_\alpha$ by Gaussian noise of mean $\mathbf{0}$ and covariance matrix $V[\boldsymbol{x}_\alpha]$ independently. The true values $\bar{\boldsymbol{x}}_\alpha$ are supposed to satisfy r constraint equations

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

parameterized by a p-dimensional vector $\boldsymbol{u}$. We call the domain $\mathcal{X}$ of the data $\{\boldsymbol{x}_\alpha\}$ the *data space*, and the domain $\mathcal{U}$ of the vector $\boldsymbol{u}$ 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. (1) 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 $\{\boldsymbol{x}_\alpha\}$.

This problem can be extended to the case where the data $\{\boldsymbol{x}_\alpha\}$ are constrained to be in a manifold in the data space $\mathcal{X}$ and the parameter $\boldsymbol{u}$ is constrained to be in a manifold in the parameter space $\mathcal{U}$, enabling us to deal with the situation where $\boldsymbol{x}_\alpha$ and $\boldsymbol{u}$ are, say, normalized to unit vectors. Also, the r equations $F^{(k)}(\boldsymbol{x}, \boldsymbol{u}) = 0$, $k = 1, ..., r$, need not be mutually independent. The subsequent argument still holds if (Moore-Penrose) generalized inverses and projection operations are introduced (see [8] for the details).

We write the covariance matrix $V[\boldsymbol{x}_\alpha]$ in the form

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

and call the constant ϵ the *noise level* and the matrix $V_0[\boldsymbol{x}_\alpha]$ the *normalized covariance matrix*. One reason for this separation is that the absolute magnitude of noise is unknown in many practical problems while its qualitative characteristics can be relatively easily estimated or determined from the gray levels of the input images [20].

Another reason is that the maximum likelihood solution is not affected in geometric fitting if the covariance matrix is multiplied by a positive constant. Hence, it suffices to know only the normalized covariance matrix $V_0[\boldsymbol{x}_\alpha]$. In fact, the geometric fitting problem as defined above can be solved by minimizing the sum of the square *Mahalanobis distances*

$$J = \sum_{\alpha=1}^{N} (\boldsymbol{x}_\alpha - \bar{\boldsymbol{x}}_\alpha, V_0[\boldsymbol{x}_\alpha]^{-1}(\boldsymbol{x}_\alpha - \bar{\boldsymbol{x}}_\alpha)) \tag{3}$$

subject to the constraint (1), where and hereafter we denote the inner product of vectors $\boldsymbol{a}$ and $\boldsymbol{b}$ by $(\boldsymbol{a}, \boldsymbol{b})$.

If we assume that the noise is small, we can eliminate the constraint (1), using first order approximation and Lagrange multipliers, in the following form [8]:

$$J = \sum_{\alpha=1}^{N} \sum_{k,l=1}^{r} W_\alpha^{(kl)} F^{(k)}(\boldsymbol{x}_\alpha, \boldsymbol{u}) F^{(l)}(\boldsymbol{x}_\alpha, \boldsymbol{u}). \tag{4}$$

Here, $W_\alpha^{(kl)}$ is the (kl) element of the inverse of the $r \times r$ matrix whose (kl) element is $(\nabla_{\mathbf{x}} F_\alpha^{(k)}, V[\boldsymbol{x}_\alpha] \nabla_{\mathbf{x}} F_\alpha^{(l)})$, where the subscript α in $\nabla_{\mathbf{x}} F_\alpha^{(k)}$ means $\boldsymbol{x} = \boldsymbol{x}_\alpha$ is substituted.

2.3 Asymptotic variables

It can be shown that the covariance matrix $V[\hat{\boldsymbol{u}}]$ of the maximum likelihood solution $\hat{\boldsymbol{u}}$ that minimizes eq. (4) not only converges to $\boldsymbol{O}$ as $\epsilon \to 0$ but also satisfies the theoretical accuracy bound within terms of $O(\epsilon^4)$ [8, 9].

This corresponds to the fact that in statistical inference the covariance matrix of the maximum likelihood solution not only converges to $\boldsymbol{O}$ as the number n of observations goes to infinity (*consistency*) but also satisfies the Cramer-Rao lower bound within terms of $O(1/n^2)$ (*asymptotic efficiency*).

In general, a complicated problem for which exact analysis is difficult often has a simple form that elucidates the underlying mathematical structure if some variable is very large or very small. Let us tentatively call such a variable an *asymptotic variable*.

In statistical inference, the number n of observations is usually taken as the asymptotic variable. This reflects the fundamental paradigm of statistical inference that the truth that underlies apparent random phenomena can be uncovered by repeated observations. It follows that an estimation method whose performance improves rapidly as $n \to \infty$ is desirable, since *such a method requires a smaller number of observations* to reach acceptable accuracy.

In geometric fitting, we take the noise level ϵ as the asymptotic variable. This reflects the requirement that a desirable estimation method should improve its performance rapidly as $\epsilon \to 0$, since *such a method can tolerate a higher noise level* to maintain acceptable accuracy [8, 12].

It is thus anticipated that all the properties of statistical inference in the limit $n \to \infty$ of a large number of observations hold in geometric fitting in the limit $\epsilon \to 0$ of an infinitesimal noise. Indeed, this can be justified by the following thought experiment. Suppose we observe an image many times. In reality, the result is always the same as long as the image and image processing algorithms involved are the same. It is, therefore, impossible to observe a different occurrence of the "noise", by which we mean the inaccuracy due to the limited resolution and imperfection of the processing algorithms. In this sense, the number n of observations is always 1. However, if we hypothetically imagine that noise occurrence changes independently each time we observe the same image. Then, we could obtain more accurate data by taking the average of the n observations. This means that increasing the number n of hypothetical observations is equivalent to effectively reducing the noise level ϵ [12].

2.3 Models and model selection

The goal of statistical inference is to explain the data-generating mechanism of apparent random phenomena. Hence, an observation $\boldsymbol{x}$ is expressed as a composite of deterministic and random parts. In abstract terms, the problem is to estimate from a given sequence of data $\{\boldsymbol{x}_i\}$ the parameter $\boldsymbol{\theta}$ of the density $P(\boldsymbol{x}|\boldsymbol{\theta})$ according to which the data $\{\boldsymbol{x}_i\}$ are assumed to have been sampled. The parameter $\boldsymbol{\theta}$ consists of the deterministic part (e.g., the coefficients of the equation that generates the data in the absence of noise) and the random part (e.g., noise characteristics such as means and variances). If we hypothesize for the density multiple possibilities $P_1(\boldsymbol{x}|\boldsymbol{\theta}_1)$, $P_2(\boldsymbol{x}|\boldsymbol{\theta}_2)$, ..., each is called a (*stochastic*) *model*; the task of choosing an appropriate one is (*stochastic*) *model selection*.

In contrast, the goal of geometric fitting is to estimate the parameter $\boldsymbol{u}$ of the constraint $\boldsymbol{F}(\boldsymbol{x}, \boldsymbol{u}) = \boldsymbol{0}$ that the data $\{\boldsymbol{x}_\alpha\}$ are supposed to satisfy. The parameter $\boldsymbol{u}$ is purely of geometric nature; it does *not* contain any characteristics of random noise. If we hypothesize for the constraint multiple possibilities $\boldsymbol{F}_1(\boldsymbol{x}, \boldsymbol{u}_1) = \boldsymbol{0}$, $\boldsymbol{F}_2(\boldsymbol{x}, \boldsymbol{u}_2) = \boldsymbol{0}$, ..., each is called a (*geometric*) *model*; the task of choosing an appropriate one is (*geometric*) *model selection* [8, 12].

In geometric fitting, the characteristics of the noise are assumed a priori, independently of the constraint (i.e., the model). In particular, the noise level ϵ is a characteristic of the image and image processing algorithms involved, independent of our interpretation of the image.

3. Geometric AIC

The derivation of Akaike's AIC [1], which is tuned to statistical inference, is tailored to geometric fitting as follows.

3.1 Goodness of a model

Under the model (1), the data can be regarded as *one* sample from the following density (we use uppercases for random variables and lowercases for their instances; $|\cdot|$ denotes the determinant):

$$P(\{\boldsymbol{X}_\alpha\}) = \prod_{\alpha=1}^{N} \frac{e^{-(\boldsymbol{X}_\alpha - \bar{\boldsymbol{x}}_\alpha, V[\boldsymbol{x}_\alpha]^{-1}(\boldsymbol{X}_\alpha - \bar{\boldsymbol{x}}_\alpha))/2}}{\sqrt{(2\pi)^m |V[\boldsymbol{x}_\alpha]|}}. \quad (5)$$

The true values $\{\bar{\boldsymbol{x}}_\alpha\}$ are constrained by eq. (1). The measure of the goodness of this model adopted by Akaike is the *Kullback-Leibler distance* (or *divergence*) from this density to the true density $P_T(\{\boldsymbol{X}_\alpha\})$

$$\begin{aligned} D &= \int \cdots \int P_T(\{\boldsymbol{X}_\alpha\}) \log \frac{P_T(\{\boldsymbol{X}_\alpha\})}{P(\{\boldsymbol{X}_\alpha\})} d\boldsymbol{X}_1 \cdots d\boldsymbol{X}_N \\ &= E[\log P_T(\{\boldsymbol{X}_\alpha\})] - E[\log P(\{\boldsymbol{X}_\alpha\})], \quad (6) \end{aligned}$$

where $E[\,\cdot\,]$ denotes expectation with respect to the true density $P_T(\{\boldsymbol{X}_\alpha\})$. The assumed model is regarded as good if D is small. The first term on the last right-hand side does not depend on individual models, so we regard the model as good if

$$\begin{aligned} &-E[\log P(\{\boldsymbol{X}_\alpha\})] \\ &= \frac{1}{2\epsilon^2} E[\sum_{\alpha=1}^{N} (\boldsymbol{X}_\alpha - \bar{\boldsymbol{x}}_\alpha, V_0[\boldsymbol{x}_\alpha]^{-1}(\boldsymbol{X}_\alpha - \bar{\boldsymbol{x}}_\alpha))] \\ &\quad + \frac{mN}{2} \log 2\pi\epsilon^2 + \frac{1}{2} \sum_{\alpha=1}^{N} \log |V_0[\boldsymbol{x}_\alpha]| \quad (7) \end{aligned}$$

is small, where we have substituted eq. (2). The last two terms do not depend on individual models. So, multiplying the first term by $2\epsilon^2$, we seek a model that minimizes the *expected residual*

$$E = E[\sum_{\alpha=1}^{N} (\boldsymbol{X}_\alpha - \bar{\boldsymbol{x}}_\alpha, V_0[\boldsymbol{x}_\alpha]^{-1}(\boldsymbol{X}_\alpha - \bar{\boldsymbol{x}}_\alpha))]. \quad (8)$$

This is the well known *least-squares criterion* under Gaussian noise. Note that ϵ is not a model parameter and hence multiplication of positive quantity that depends only on ϵ does not affect model selection.

3.2 Evaluation of expectation

The difficulty of using eq. (8) as a model selection criterion is that the expectation $E[\,\cdot\,]$ must be evaluated with respect to the *true* density, which we do not know. How to deal with this will lead to the fundamental difference between geometric fitting and statistical inference.

In statistical inference, we can assume that we could, at least in principle, observe as many data as desired. If we are allowed to sample independent instances $\boldsymbol{x}_1$, $\boldsymbol{x}_2$, ..., $\boldsymbol{x}_n$ according to a density $P_T(\boldsymbol{X})$, the expectation $\int Y(\boldsymbol{X}) P_T(\boldsymbol{X}) d\boldsymbol{X}$ of a statistic $Y(\boldsymbol{X})$ can be approximated by $(1/n)\sum_{i=1}^{n} Y(\boldsymbol{x}_i)$ (the *Monte-Carlo method*), which converges to the true expectation in the limit $n \to \infty$ (the *law of large numbers*). Akaike's AIC is based on this principle.

In geometric fitting, in contrast, we can obtain only *one* instance $\{\boldsymbol{x}_\alpha\}$ of $\{\boldsymbol{X}_\alpha\}$, so it is impossible to replace expectation by sample average. But we can assume that we could, at least in principle, use devices of as high resolution as desired. In our framework, therefore, it suffices to find an approximation that holds in the limit $\epsilon \to 0$. Evidently, the expectation $\int \cdots \int Y(\{\boldsymbol{X}_\alpha\}) P_T(\{\boldsymbol{X}_\alpha\}) d\boldsymbol{X}_1 \cdots d\boldsymbol{X}_N$ of $Y(\{\boldsymbol{X}_\alpha\})$ can be approximated by $Y(\{\boldsymbol{x}_\alpha\})$ (note that we do not need $1/N$), because as $\epsilon \to 0$ we have $P_T(\{\boldsymbol{X}_\alpha\}) \to \prod_{\alpha=1}^{N} \delta(\boldsymbol{X}_\alpha - \bar{\boldsymbol{x}}_\alpha)$, where $\delta(\,\cdot\,)$ denotes the Dirac delta function. Thus, $\int \cdots \int Y(\{\boldsymbol{X}_\alpha\}) P_T(\{\boldsymbol{X}_\alpha\}) d\boldsymbol{X}_1 \cdots d\boldsymbol{X}_N$ and $Y(\{\boldsymbol{x}_\alpha\})$ both converge to $Y(\{\bar{\boldsymbol{x}}_\alpha\})$.

It follows that we can approximate E by

$$J = \sum_{\alpha=1}^{N} (\boldsymbol{x}_\alpha - \bar{\boldsymbol{x}}_\alpha, V_0[\boldsymbol{x}_\alpha]^{-1}(\boldsymbol{x}_\alpha - \bar{\boldsymbol{x}}_\alpha)). \quad (9)$$

3.3 Bias removal

There is still a difficulty using eq. (9) as a criterion: the model parameters $\{\bar{\boldsymbol{x}}_\alpha\}$ and $\boldsymbol{u}$ need to be estimated. If we are to view eq. (9) as a measure of the goodness of the model, we should choose for $\{\bar{\boldsymbol{x}}_\alpha\}$ and $\boldsymbol{u}$ their *maximum likelihood estimators* $\{\hat{\boldsymbol{x}}_\alpha\}$ and $\hat{\boldsymbol{u}}$ that minimize eq. (9) subject to the constraint (1). A naive idea is to substitute $\{\hat{\boldsymbol{x}}_\alpha\}$ and $\hat{\boldsymbol{u}}$ for $\{\bar{\boldsymbol{x}}_\alpha\}$ and $\boldsymbol{u}$ in eq. (9) and use as a model selection criterion

$$\hat{J} = \sum_{\alpha=1}^{N} (\boldsymbol{x}_\alpha - \hat{\boldsymbol{x}}_\alpha, V_0[\boldsymbol{x}_\alpha]^{-1}(\boldsymbol{x}_\alpha - \hat{\boldsymbol{x}}_\alpha)), \quad (10)$$

which is called the *residual* (*sum of squares*). However, a logical inconsistency arises.

Eq. (1) does not define a particular model. Rather, it defines a *class* of models parameterized by $\{\bar{\boldsymbol{x}}_\alpha\}$ and $\boldsymbol{u}$. If we choose particular values $\{\hat{\boldsymbol{x}}_\alpha\}$ and $\hat{\boldsymbol{u}}$, we are given a particular model. According to the logic in Sec. 3.1, its goodness should be evaluated by $E[\sum_{\alpha=1}^N(\boldsymbol{X}_\alpha - \hat{\boldsymbol{x}}_\alpha, V_0[\boldsymbol{x}_\alpha]^{-1}(\boldsymbol{X}_\alpha - \hat{\boldsymbol{x}}_\alpha))]$. According to the logic in Sec. 3.2, the expectation can be approximated using a typical instance of $\{\boldsymbol{X}_\alpha\}$.

However, $\{\hat{\boldsymbol{x}}_\alpha\}$ and $\hat{\boldsymbol{u}}$ were *computed* from $\{\boldsymbol{x}_\alpha\}$, so $\{\boldsymbol{x}_\alpha\}$ cannot be a typical instance of $\{\boldsymbol{X}_\alpha\}$ due to the correlation with the assumed model. In fact, $\hat{J}$ is generally smaller than $E[\sum_{\alpha=1}^N(\boldsymbol{X}_\alpha - \hat{\boldsymbol{x}}_\alpha, V_0[\boldsymbol{x}_\alpha]^{-1}(\boldsymbol{X}_\alpha - \hat{\boldsymbol{x}}_\alpha))]$, because $\{\hat{\boldsymbol{x}}_\alpha\}$ and $\hat{\boldsymbol{u}}$ were determined *so as to minimize* $\hat{J}$.

This is the difficulty that Akaike encountered in the derivation of his AIC. His strategy for resolving this can be translated in our setting as follows.

Ideally, we should approximate the expectation using an instance $\{\boldsymbol{x}^*_\alpha\}$ of $\{\boldsymbol{X}_\alpha\}$ generated *independently* of the current data $\{\boldsymbol{x}_\alpha\}$. In other words, we should evaluate

$$J^* = \sum_{\alpha=1}^N(\boldsymbol{x}^*_\alpha - \hat{\boldsymbol{x}}_\alpha, V_0[\boldsymbol{x}_\alpha]^{-1}(\boldsymbol{x}^*_\alpha - \hat{\boldsymbol{x}}_\alpha)). \qquad (11)$$

Let us call $\{\boldsymbol{x}^*_\alpha\}$ the *future data*; they are "another" instance of $\{\boldsymbol{X}_\alpha\}$ that *might* occur if we did a hypothetical experiment. In reality, however, we have the current data $\{\boldsymbol{x}_\alpha\}$ alone at hand[1]. So, we try to compensate for the bias in the form

$$\hat{J}^* = \hat{J} + b\epsilon^2. \qquad (12)$$

It is easily seen that $\hat{J}^*$ and $\hat{J}$ are both $O(\epsilon^2)$ and hence b is $O(1)$. Since $\hat{J}^*$ and $\hat{J}$ are random variables, so is b. It can be proved [8, 11] that

$$E^*[E[b]] = 2(Nd + p) + O(\epsilon^2), \qquad (13)$$

where $E[\,\cdot\,]$ and $E^*[\,\cdot\,]$ denote expectations with respect to $\{\boldsymbol{x}_\alpha\}$ and $\{\boldsymbol{x}^*_\alpha\}$, respectively. From this, we obtain an unbiased estimator of $\hat{J}^*$ in the first order in the form

$$\text{G-AIC} = \hat{J} + 2(Nd + p)\epsilon^2, \qquad (14)$$

where $d = m - r$ is the dimension of the manifold $\mathcal{S}$ defined in the data space $\mathcal{X}$ by the constraint $F^{(k)}(\boldsymbol{x}, \boldsymbol{u}) = 0$, $k = 1, ..., r$. This criterion is the *geometric AIC* proposed by Kanatani [8, 11, 13].

3.4 Duality of asymptotic analysis

Although the final form is the same, Kanatani's original derivation starting from eq. (11) with a heuristic reasoning [8, 11, 13]. This has caused a lot of confusion as to whether the geometric AIC and Akaike's AIC are the same or not. The present formulation makes clear where they are the same and from where they depart.

In Akaike's derivation, the following facts play a fundamental role:

- The maximum likelihood estimator converges to its true value as $n \to \infty$ (the *law of large numbers*).
- The maximum likelihood estimator 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.

In the derivation of eq. (13), the following facts play a crucial role [8, 11]:

- The maximum likelihood estimator converges to its true value as $\epsilon \to 0$.
- The maximum likelihood estimator 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*.
- A quadratic form in standardized Gaussian random variables is subject to a χ^2 distribution, whose expectation equals its degree of freedom.

We observe a kind of "duality" between geometric fitting and statistical inference. In particular, we see that the noise level ϵ in geometric fitting plays the same role as the number n of observations in statistical inference. This is obvious if we recall the thought experiment in Sec. 2.2: reducing the noise is equivalent to increasing the number of hypothetical observations.

The confusion about the relationship between the geometric AIC and Akaike's AIC originates from the apparent similarity of their forms, which is due to the fact that the correction term in Akaike's AIC is independent of the number n of observations. If n were involved, the corresponding form for geometric fitting would have a very different form, as we will show subsequently. In this sense, the similarity, or "self-duality", between the geometric AIC and Akaike's AIC is a mere accident, which has hidden the difference underneath geometric fitting and statistical inference.

4. Geometric MDL

We now derive in our framework the counterpart of Rissanen's *MDL* (*minimum description length*) [28, 29, 30].

4.1 MDL principle

Rissanen's MDL measures the goodness of the model by the information theoretic code length. The basic idea is simple, but the following difficulties must be resolved for applying it in practice:

- Encoding a problem involving real numbers requires an infinitely long code length.

[1] If such data $\{\boldsymbol{x}^*_\alpha\}$ actually exist, the test using them is called *cross-validation*. We can also generate equivalent data by a computer. Such a simulations is called *bootstrap* [5].

- The probability density, from which a minimum length code can be obtained, involves unknown parameters.
- Obtaining an exact form of the minimum length code is very difficult.

Rissanen [28, 29] avoided these difficulties by quantizing the real numbers in a way that does not depend on individual models and substituting the maximum likelihood estimators for the parameters. They, too, are real numbers, so they are also quantized. The quantization width is chosen so as to minimize the total description length (the *two-stage encoding*). The resulting code length is asymptotically evaluated by taking the data length n as the asymptotic variable. This idea of Rissanen can be translated into our framework as follows.

If the data $\{\boldsymbol{x}_\alpha\}$ are sampled according to the probability density (5), they can be encoded, after their domain is quantized, in a shortest prefix code of length

$$-\log P = \frac{J}{2\epsilon^2} + \frac{mN}{2}\log 2\pi\epsilon^2 + \frac{1}{2}\sum_{\alpha=1}^{N}\log|V_0[\boldsymbol{x}_\alpha]|, \tag{15}$$

up to a constant that depends only on the domain and the width of the quantization. Here, J is the sum of the square Mahalanobis distances given by eq. (3). Using the natural logarithm, we take $\log_2 e$ bits as the unit of length.

4.2 Two-stage encoding

In order to do encoding using eq. (5), we need the true values $\{\bar{\boldsymbol{x}}_\alpha\}$ and the parameter $\boldsymbol{u}$. Since they are unknown, we use their maximum likelihood estimators that minimize eq. (15) (specifically J). The last two terms of eq. (15) do not depend on individual (geometric) models (recall that ϵ is not a model parameter), so the minimum code length is $\hat{J}/2\epsilon^2$ up to a constant that does not depend on individual models, where $\hat{J}$ is the residual given by eq. (10). For brevity, we hereafter call the code length determined up to a constant that does not depend on individual models simply the *description length*.

Since the maximum likelihood estimators $\{\hat{\boldsymbol{x}}_\alpha\}$ and $\hat{\boldsymbol{u}}$ are real numbers, they must also be quantized. If we use a large quantization width, their code lengths become short, but the description length $\hat{J}/2\epsilon^2$ will increase. So, we take the width that minimizes the total description length. The computation is based on the fact that eq. (4) can be written as follows [8]:

$$J = \hat{J} + \sum_{\alpha=1}^{N}(\boldsymbol{x}_\alpha - \hat{\boldsymbol{x}}_\alpha, V_0[\hat{\boldsymbol{x}}_\alpha]^-(\boldsymbol{x}_\alpha - \hat{\boldsymbol{x}}_\alpha)) \\ +(\boldsymbol{u} - \hat{\boldsymbol{u}}, V_0[\hat{\boldsymbol{u}}]^{-1}(\boldsymbol{u} - \hat{\boldsymbol{u}})) + O(\epsilon^3). \tag{16}$$

Here, the superscript $-$ denotes the (Moore-Penrose) generalized inverse, and $V_0[\hat{\boldsymbol{x}}_\alpha]$ and $V_0[\hat{\boldsymbol{u}}_\alpha]$ are, respectively, the a posteriori covariance matrices of the maximum likelihood estimators $\hat{\boldsymbol{x}}_\alpha$ and $\hat{\boldsymbol{u}}$ given as follows [8]:

$$V_0[\hat{\boldsymbol{x}}_\alpha] = V_0[\boldsymbol{x}_\alpha] \\ -\sum_{k,l=1}^{r} W_\alpha^{(kl)}(V[\boldsymbol{x}_\alpha]\nabla_{\mathbf{x}}F_\alpha^{(k)})(V[\boldsymbol{x}_\alpha]\nabla_{\mathbf{x}}F_\alpha^{(k)})^\top,$$

$$V_0[\hat{\boldsymbol{u}}] = \Big(\sum_{\alpha=1}^{N}\sum_{k,l=1}^{r} W_\alpha^{(kl)}(\nabla_{\mathbf{u}}F_\alpha^{(k)})(\nabla_{\mathbf{u}}F_\alpha^{(l)})^\top\Big)^{-1}. \tag{17}$$

The symbol $W_\alpha^{(kl)}$ has the same meaning as in eq. (4).

4.3 Encoding parameters

In order to quantize $\hat{\boldsymbol{u}}$, we quantize the p-dimensional parameter space $\mathcal{U}$ by introducing appropriate (generally curvilinear) coordinates and defining a grid of width δu_i. Suppose $\hat{\boldsymbol{u}}$ is in a rectangular region of size L_i. There are $\prod_{i=1}^{p}(L_i/\delta u_i)$ grid vertices inside, so specifying one from these requires the code length

$$\log\prod_{i=1}^{p}\frac{L_i}{\delta u_i} = \log V_u - \sum_{i=1}^{p}\log\delta u_i, \tag{18}$$

where $V_u = \prod_{i=1}^{p} L_i$ is the volume of the rectangular region. We could reduce this code length using a large width δu_i, but eq. (16) implies that replacing $\hat{\boldsymbol{u}}$ by the nearest vertex would increase the description length $\hat{J}/2\epsilon^2$ by $(\boldsymbol{\delta u}, V_0[\hat{\boldsymbol{u}}]^{-1}\boldsymbol{\delta u})/2\epsilon^2$ in the first order in ϵ, where we define $\boldsymbol{\delta u} = (\delta u_i)$. Differentiating the sum of this and eq. (18) with respect to δu_i and letting the result be 0, we obtain

$$\frac{1}{\epsilon^2}\Big(V_0[\hat{\boldsymbol{u}}]^{-1}\boldsymbol{\delta u}\Big)_i = \frac{1}{\delta u_i}, \tag{19}$$

where $(\,\cdot\,)_i$ designates the ith component. If the coordinate system of $\mathcal{U}$ is taken in such a way that $V_0[\hat{\boldsymbol{u}}]^{-1}$ is diagonalized, we obtain

$$\delta u_i = \frac{\epsilon}{\sqrt{\lambda_i}}, \tag{20}$$

where λ_i is the ith eigenvalue of $V_0[\hat{\boldsymbol{u}}]^{-1}$. It follows that the volume of one cell of the grid is

$$v_u = \prod_{i=1}^{p}\delta u_i = \frac{\epsilon^p}{\sqrt{|V_0[\hat{\boldsymbol{u}}]^{-1}|}}. \tag{21}$$

Hence, the number of cells inside the region V_u is

$$N_u = \int_{V_u}\frac{d\boldsymbol{u}}{v_u} = \frac{1}{\epsilon^p}\int_{V_u}\sqrt{|V_0[\hat{\boldsymbol{u}}]^{-1}|}d\boldsymbol{u}. \tag{22}$$

Specifying one from these requires the code length

$$\log N_u = \log\int_{V_u}\sqrt{|V_0[\hat{\boldsymbol{u}}]^{-1}|}d\boldsymbol{u} - \frac{p}{2}\log\epsilon^2. \tag{23}$$

4.4 Encoding true values

In order to quantize $\{\hat{\boldsymbol{x}}_\alpha\}$, we need to quantize their domain. Although the domain of the data $\{\boldsymbol{x}_\alpha\}$ is the m-dimensional data space $\mathcal{X}$, their true values are constrained to be in a d-dimensional manifold $\hat{\mathcal{S}}$ parameterized by $\hat{\boldsymbol{u}}$, which we have already encoded. So, we introduce appropriate curvilinear coordinates in $\hat{\mathcal{S}}$ and define a (curvilinear) grid of width $\delta\xi_{i\alpha}$. Since each $\hat{\boldsymbol{x}}_\alpha$ has its own normalized covariance matrix $V_0[\hat{\boldsymbol{x}}_\alpha]$ (see eqs. (17)), we quantize each $\hat{\boldsymbol{x}}_\alpha$ differently, using different curvilinear coordinates for each.

Suppose $\hat{\boldsymbol{x}}_\alpha$ is in a (curvilinear) rectangular region of size $l_{i\alpha}$. There are $\prod_{i=1}^d (l_{i\alpha}/\delta\xi_{i\alpha})$ grid vertices inside, so specifying one from these requires the code length

$$\sum_{i=1}^d \log\frac{l_{i\alpha}}{\delta\xi_{i\alpha}} = \log V_{x\alpha} - \sum_{i=1}^d \log\delta\xi_{i\alpha}, \tag{24}$$

where $V_{x\alpha} = \prod_{i=1}^d l_{i\alpha}$ is the volume of the rectangular region. We could reduce this code length using a large width $\delta\xi_{i\alpha}$, but replacing $\hat{\boldsymbol{x}}_\alpha$ by its nearest vertex would increase the description length $\hat{J}/2\epsilon^2$. Let $\delta\bar{\boldsymbol{x}}_\alpha$ be the m-dimensional vector that expresses the displacement $\{\delta\xi_{i\alpha}\}$ on $\hat{\mathcal{S}}$ with respect to the coordinate system of $\mathcal{X}$. Eq. (16) implies that the increase in $\hat{J}/2\epsilon^2$ is $(\delta\bar{\boldsymbol{x}}_\alpha, V_0[\hat{\boldsymbol{x}}_\alpha]^-\delta\bar{\boldsymbol{x}}_\alpha)/2\epsilon^2$ in the first order in ϵ. Differentiating the sum of this and eq. (24) with respect to $\delta\xi_{i\alpha}$ and letting the result be 0, we obtain

$$\frac{1}{\epsilon^2}\Big(V_0[\hat{\boldsymbol{x}}_\alpha]^-\delta\bar{\boldsymbol{x}}_\alpha\Big)_i = \frac{1}{\delta\xi_{i\alpha}}. \tag{25}$$

Note that $V_0[\hat{\boldsymbol{x}}_\alpha]^-$ is a singular matrix of rank d whose domain is the tangent space to $\hat{\mathcal{S}}$ at $\hat{\boldsymbol{x}}_\alpha$. We define the curvilinear coordinates in $\hat{\mathcal{S}}$ in such a way that the p basis vectors at $\hat{\boldsymbol{x}}_\alpha$ form an orthonormal system. Suppose the coordinate system of $\mathcal{X}$ is defined in such a way that its basis vectors consist of the p basis vectors of $\hat{\mathcal{S}}$ at $\hat{\boldsymbol{x}}_\alpha$ plus an orthonormal system of $m-p$ vectors orthogonal to them. If, moreover, we choose the curvilinear coordinates of $\hat{\mathcal{S}}$ in such a way that $V_0[\hat{\boldsymbol{x}}_\alpha]^-$ is diagonalized, we obtain the solution $\delta\xi_{i\alpha}$ of eq. (25) in the form

$$\delta\xi_{i\alpha} = \left\{\begin{array}{ll} \epsilon/\sqrt{\lambda_{i\alpha}} & i=1,...,d \\ 0 & i=d+1,...,m \end{array}\right., \tag{26}$$

where $\lambda_{1\alpha}$, ..., $\lambda_{d\alpha}$ are the d positive eigenvalues of $V_0[\hat{\boldsymbol{x}}_\alpha]^-$. It follows that the volume of one cell of the grid is

$$v_{x\alpha} = \prod_{i=1}^d \delta\xi_{i\alpha} = \frac{\epsilon^d}{\sqrt{|V_0[\hat{\boldsymbol{x}}_\alpha]^-|_d}}, \tag{27}$$

where $|V_0[\hat{\boldsymbol{x}}_\alpha]^-|_d$ denotes the product of its d positive eigenvalues. Hence, the number of cells inside the region $V_{x\alpha}$ is

$$N_\alpha = \int_{V_{x\alpha}} \frac{d\boldsymbol{x}}{v_{x\alpha}} = \frac{1}{\epsilon^d}\int_{V_{x\alpha}} \sqrt{|V_0[\hat{\boldsymbol{x}}_\alpha]^-|_d}\, d\boldsymbol{x}. \tag{28}$$

Specifying one from these requires the code length

$$\log N_\alpha = \log\int_{V_{x\alpha}} \sqrt{|V_0[\hat{\boldsymbol{x}}_\alpha]^-|_d}\, d\boldsymbol{x} - \frac{d}{2}\log\epsilon^2. \tag{29}$$

4.5 Geometric MDL

From eqs. (23) and (29), the total code length for $\{\hat{\boldsymbol{x}}_\alpha\}$ and $\hat{\boldsymbol{u}}$ becomes

$$\sum_{\alpha=1}^N \log\int_{V_{x\alpha}} \sqrt{|V_0[\hat{\boldsymbol{x}}_\alpha]^-|_d}\, d\boldsymbol{x} + \log\int_{V_u}\sqrt{|V_0[\hat{\boldsymbol{u}}]^{-1}|}d\boldsymbol{u}. \\ -\frac{Nd+p}{2}\log\epsilon^2 \tag{30}$$

The resulting increase in the description length $\hat{J}/2\epsilon^2$ is $(\delta\bar{\boldsymbol{x}}_\alpha, V_0[\hat{\boldsymbol{x}}_\alpha]^-\delta\bar{\boldsymbol{x}}_\alpha)/2\epsilon^2 + (\delta\boldsymbol{u}, V_0[\hat{\boldsymbol{u}}]^{-1}\delta\boldsymbol{u})/2\epsilon^2$ in the first order in ϵ. If we substitute eqs. (20) and (26) together with $V_0[\hat{\boldsymbol{x}}_\alpha]^- = \mathrm{diag}(1/\lambda_{1\alpha}, ..., 1/\lambda_{d\alpha}, 0, ..., 0)$ and $V_0[\hat{\boldsymbol{u}}]^{-1} = \mathrm{diag}(1/\lambda_1, ..., 1/\lambda_p)$, this increase in the description length is

$$\frac{(\delta\bar{\boldsymbol{x}}_\alpha, V_0[\hat{\boldsymbol{x}}_\alpha]^-\delta\bar{\boldsymbol{x}}_\alpha)}{2\epsilon^2} + \frac{(\delta\boldsymbol{u}, V_0[\hat{\boldsymbol{u}}]^{-1}\delta\boldsymbol{u})}{2\epsilon^2} = \frac{Nd+p}{2} \tag{31}$$

if higher order terms in ϵ are omitted. Since eqs. (20) and (26) are obtained by omitting terms of $o(\epsilon)$, the omitted terms in eq. (31) are $o(1)$. It follows that the total description length is

$$\frac{\hat{J}}{2\epsilon^2} - \frac{Nd+p}{2}\log\epsilon^2 + \sum_{\alpha=1}^N \log\int_{V_{x\alpha}}\sqrt{|V_0[\hat{\boldsymbol{x}}_\alpha]^-|_d}\, d\boldsymbol{x} \\ + \log\int_{V_u}\sqrt{|V_0[\hat{\boldsymbol{u}}]^{-1}|}d\boldsymbol{u} + \frac{Nd+p}{2} + o(1). \tag{32}$$

Since ϵ is not a model parameter, multiplication by a positive quantity that depends only on ϵ does not affect model selection. So, we multiply the above expression by $2\epsilon^2$ and write

$$\begin{aligned}\text{G-MDL} = {} & \hat{J} - (Nd+p)\epsilon^2\log\epsilon^2 \\ & +2\epsilon^2\Big(\sum_{\alpha=1}^N \log\int_{V_{x\alpha}}\sqrt{|V_0[\hat{\boldsymbol{x}}_\alpha]^-|_d}\, d\boldsymbol{x} \\ & +\log\int_{V_u}\sqrt{|V_0[\hat{\boldsymbol{u}}]^{-1}|}d\boldsymbol{u}\Big) \\ & +(Nd+p)\epsilon^2 + o(\epsilon^2),\end{aligned} \tag{33}$$

which we call the *geometric MDL*.

4.6 Scale choice

In practice, it is difficult to use eq. (33) as a criterion because of the difficulty in evaluating the third term on the right-hand side. First, the matrices $V_0[\hat{\boldsymbol{x}}_\alpha]$ and $V_0[\hat{\boldsymbol{u}}]$ given by eqs. (17) have complicated forms, so it is difficult to integrate them. But a more serious problem is that the regions $V_{x\alpha}$ and V_u must be finite so that the integrations exist. If the data space $\mathcal{X}$ and the parameter space $\mathcal{U}$ are unbounded, we must specify in them finite regions in which the true values are likely to exist. This is nothing but the Bayesian standpoint that requires prior distributions for parameters to be estimated.

After all, reducing model selection to code length requires the Bayesian standpoint, because if the parameters can be anywhere in unbounded regions, it is impossible to obtain a code of finite length unless some information about their likely locations is given. An expedient for this is to omit diverging quantities and higher order terms as long as they do not affect the model selection very much, so that the final form is independent of prior distributions.

If we note that $-\log \epsilon^2 \gg 1$ as $\epsilon \to 0$, we may omit terms of $O(\epsilon^2)$ in eq. (33). Then, we obtain

$$\text{G-MDL} = \hat{J} - (Nd + p)\epsilon^2 \log \epsilon^2. \qquad (34)$$

This is the form proposed by Matsunaga and Kanatani [22] by a heuristic reasoning. One need not worry about integration in this form, but instead the problem of scale arises. If we multiply the unit of length by, say, 10, both ϵ^2 and $\hat{J}$ are multiplied by 1/100. Since N, d, and p are nondimensional constants, G-MDL should also be multiplied by 1/100. But $\log \epsilon^2$ reduces by $\log 100$, meaning that model selection could be affected by the unit of length we use. In eq. (33), in contrast, the influence of scale is canceled between the second and third terms on the right-hand side.

The inconsistency in eq. (34) comes from the term $\log \epsilon^2$. Since the logarithm can be defined only for a nondimensional quantity, eq. (34) should have the form

$$\text{G-MDL} = \hat{J} - (Nd + p)\epsilon^2 \log\Big(\frac{\epsilon}{L}\Big)^2, \qquad (35)$$

where L is a reference length. In theory, it can be determined from the third term on the right-hand side of eq. (33), but its evaluation is difficult. So, we adopt a practical compromise, choosing a scale L such that $\boldsymbol{x}_\alpha/L$ is $O(1)$. This can be roughly interpreted as giving a prior distribution in a region of volume L^m in the data space $\mathcal{X}$. For example, if $\{\boldsymbol{x}_\alpha\}$ are image pixel data, we can take L to be the image size.

Since we are assuming that the noise is much smaller than the data, we have $-\log(\epsilon/L)^2 \gg 1$. Hence, if we use a different scale $L' = \gamma L$, we have $\log \gamma^2 \approx 0$ as long as $\gamma \approx 1$. Hence, $-\log(\epsilon/L')^2 = -\log(\epsilon/L)^2 + \log \gamma^2 \approx -\log(\epsilon/L)^2$, so the model selection is not affected very much as long as we use the scale of the same order of magnitude.

Nevertheless, the need of such a reference length is certainly a handicap as compared with the geometric AIC. However, this is unavoidable, because it originates from the very MDL principle of Rissanen, as we now argue.

4.7 MDL in statistical inference

The difficulties and expedients described above may appear to be peculiar to the geometric MDL and may cast doubt on its legitimacy. In truth, however, the same situation arises for Rissanen's MDL, for which the data length n is the asymptotic variable. Originally, Rissanen presented his MDL in the following form [28]:

$$\text{MDL} = -\log \prod_{\alpha=1}^{n} P(\boldsymbol{x}_\alpha|\hat{\boldsymbol{\theta}}) + \frac{k}{2}\log n + O(1). \qquad (36)$$

Here, the data $\{\boldsymbol{x}_\alpha\}$ are assumed to be sampled independently from a (stochastic) model (i.e., the probability density) $P(\boldsymbol{x}|\boldsymbol{\theta})$ parameterized by a k-dimensional vector $\boldsymbol{\theta}$; $\hat{\boldsymbol{\theta}}$ is its maximum likelihood estimator. The symbol $O(1)$ denotes terms of order 0 in the limit $n \to \infty$. The geometric MDL (34) of Matsunaga and Kanatani [22] was inspired by this form.

This form evokes the problem of the unit of the data $\{\boldsymbol{x}_\alpha\}$. If we regard a pair of data as "one" datum, viewing $(\boldsymbol{x}_1, \boldsymbol{x}_2)$, $(\boldsymbol{x}_3, \boldsymbol{x}_4)$, ... as sampled from $P(\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{\theta}) = P(\boldsymbol{x}|\boldsymbol{\theta})P(\boldsymbol{y}|\boldsymbol{\theta})$, the data length is apparently halved though the problem is the same. As a result, the second term on the right-hand side of eq. (36) reduces by $(k/2)\log 2$, and this could affect the model selection. Rissanen's MDL was criticicized for this defect (and others). Later, Rissanen presented the following form [30]:

$$\text{MDL} = -\log \prod_{\alpha=1}^{n} P(\boldsymbol{x}_\alpha|\hat{\boldsymbol{\theta}}) + \frac{k}{2}\log\frac{n}{2\pi} + \log \int_{V_\theta} \sqrt{|\boldsymbol{I}(\boldsymbol{\theta})|}d\boldsymbol{\theta} + o(1). \qquad (37)$$

Here, $\boldsymbol{I}(\boldsymbol{\theta})$ is the Fisher information matrix of $P(\boldsymbol{x}_\alpha|\boldsymbol{\theta})$. In this form, the effect of scale change is canceled by the corresponding change in the Fisher information matrix. However, the problem of integration arises if the domain V_θ of the parameter $\boldsymbol{\theta}$ is unbounded, just as in the case of eq. (33), so an appropriate expedient such as assuming a prior distribution becomes necessary. This has been criticized by some as a handicap over Akaike's AIC while welcomed by others as giving extra freedoms to adjust.

Thus, the geometric MDL and Rissanen's MDL share the same problem whether the asymptotic variable is the noise level ϵ or the data length n; the properties of the one are faithfully mirrored in the other. This is obvious if we recall that ϵ is effectively related to the number n of hypothetical observations (Sec. 2.2).

5. 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. Since ϵ is a constant predetermined by the image and the image processing algorithms independently of our interpretation, 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 [8]. This can be intuitively understood as follows. Recall that $\hat{J}$ can be viewed as 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 $(rN - p)\epsilon^2$. Thus, we obtain an unbiased estimator of ϵ^2 in the form

$$\hat{\epsilon}^2 = \frac{\hat{J}}{rN - p}. \tag{38}$$

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

One may wonder if model selection is necessary at all when the true model is known. In practice, however, the situation that requires model selection most is *degeneracy detection*. In 3-D analysis from images, for instance, the constraint (1) 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 when the condition nearly degenerates. 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 [12, 18, 19, 22, 26, 38].

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 degeneracy (i.e., $\mathcal{S}' \subset \mathcal{S}$), we can estimate the noise level $\hat{\epsilon}$ from the residual $\hat{J}$ of the general model $\mathcal{S}$ by eq. (38).

In statistical inference, on the other hand, the noise variance *is* a model parameter, because by "noise" we mean the random effects that account for the discrepancy between the assumed model and the actual observation. Hence, the noise variance must be estimated, if it is not known, according to the assumed model. This is one of the most different aspects between statistical inference and geometric fitting.

6. Is the Geometric MDL Really Better?

We experimentally test if the geometric MDL really outperforms the geometric AIC as anticipated. Our conclusion is negative. We also show that these two criteria have very different characteristics.

6.1 Rank estimation

Given a sequence of images of points from multiple objects independently moving in the scene, we can estimate the number of objects by computing the rank of a matrix consisting of the image coordinates of these points if there is no image noise [4, 14, 15]. In the presence of noise, we can estimate the rank by truncating smaller singular values, but it is difficult to set an appropriate threshold.

The rank r of an $n \times m$ matrix is the dimension of the subspace spanned by its m columns in $\mathcal{R}^n$, or of the subspace spanned by its n rows in $\mathcal{R}^m$. In the presence of image noise, each matrix element undergoes Gaussian noise of mean 0 and a constant variance ϵ^2 [14, 15]. The degree of freedom of an r-dimensional subspace in $\mathcal{R}^n$ is[2] $r(n-r)$. Hence, the geometric AIC and the geometric MDL are respectively given by

$$\text{G-AIC} = \hat{J}_r + 2r(m+n-r)\epsilon^2,$$
$$\text{G-MDL} = \hat{J}_r - r(m+n-r)\epsilon^2 \log\Big(\frac{\epsilon}{L}\Big)^2. \tag{39}$$

The same form is obtained if we calculate the degree of freedom of an r-dimensional subspace in $\mathcal{R}^m$. Thus, the expressions are symmetric with respect to n and m, as they should be.

Let $\nu = \min(n, m)$. The residual $\hat{J}_r$ is given by

$$\hat{J}_r = \sum_{i=r+1}^{\nu} \sigma_i^2, \tag{40}$$

where $\{\sigma_i\}$ are the singular values, in descending order, of the matrix. Evaluating eqs. (39) for $r = 1, 2, ...$, we choose the value r that minimizes them.

If the noise variance ϵ^2 is not known, we need to estimate it. It can be estimated if the rank r is known to be less than an upper bound $r_{\max}$. From (38), we obtain

$$\hat{\epsilon}^2 = \frac{\hat{J}_{r_{\max}}}{(n - r_{\max})(N - r_{\max})}. \tag{41}$$

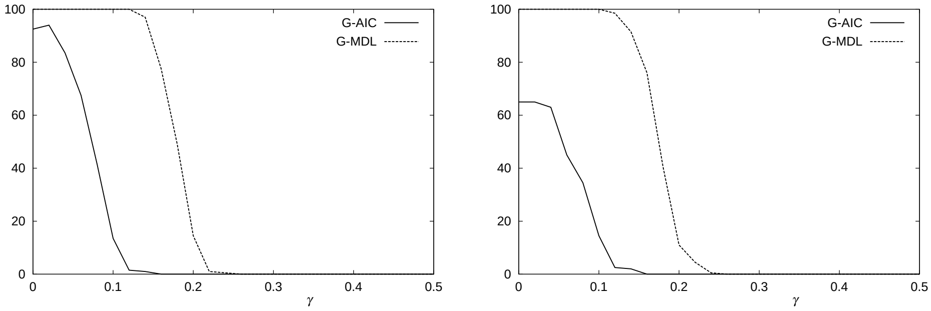
We defined a 10×20 matrix whose elements were randomly generated uniformly over $[-1, 1]$. We computed its singular value decomposition in the form $\boldsymbol{V}\text{diag}(\sigma_1, ..., \sigma_{10})\boldsymbol{U}^\top$, the singular values $\sigma_1, ..., \sigma_5$ being, respectively, 3.81, 3.58, 3.09, 2.98, 2.75. Then, we defined the matrix

$$\boldsymbol{A} = \boldsymbol{V}\text{diag}(\sigma_1, ..., \sigma_5, \gamma\sigma_5, 0, ..., 0)\boldsymbol{U}^\top. \tag{42}$$

We added Gaussian noise of mean 0 and variance ϵ^2 to each element of $\boldsymbol{A}$ independently and estimated its rank with $r_{\max} = 6$. Fig. 1 plots the ratio of the number of times the rank was estimated to be 5 over 200 trials for each γ. We used the reference length $L = 1$. Fig. 1(a) shows the case where ϵ is known; Fig. 1(b) shows the case where it is estimated.

The geometric AIC predicts the rank to be 6 with some probability even when the true rank is 5 ($\gamma = 0$). It predicts the rank to be definitely 6 even for a small value of γ. The geometric MDL almost always guesses

[2] An r-dimensional subspace of $\mathcal{R}^n$ is specified by r points in $\mathcal{R}^n$, but the r points can move freely within that subspace. So, the degree of freedom is $rn - r^2$.



(a) (b)

Figure 1: The ratio (%) of estimating the rank to be 5 by the geometric AIC (solid line) and the geometric MDL (dashed line) using (a) the true noise level and (b) the estimated noise level.

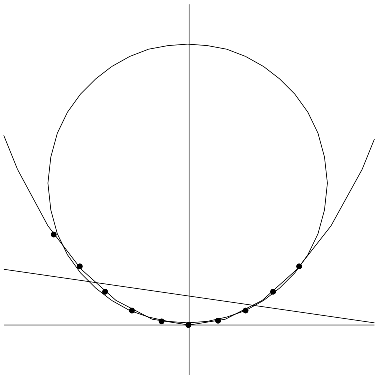


Figure 2: Fitting a line, a circle, and an ellipse.

the rank to be 5 when the true rank is 5 ($\gamma = 0$), but it keeps guessing the rank to be 5 for a wide range of γ for which the true rank is 6.

6.2 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* [8, 16, 17]. Fig. 2 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 N points are given as follows:

$$\begin{aligned}
\text{G-AIC}_l &= \hat{J}_l + 2(N+2)\epsilon^2, \\
\text{G-AIC}_c &= \hat{J}_c + 2(N+3)\epsilon^2, \\
\text{G-AIC}_e &= \hat{J}_e + 2(N+5)\epsilon^2, \\
\text{G-MDL}_l &= \hat{J}_l - (N+2)\epsilon^2 \log\Big(\frac{\epsilon}{L}\Big)^2, \\
\text{G-MDL}_c &= \hat{J}_c - (N+3)\epsilon^2 \log\Big(\frac{\epsilon}{L}\Big)^2, \\
\text{G-MDL}_e &= \hat{J}_e - (N+5)\epsilon^2 \log\Big(\frac{\epsilon}{L}\Big)^2. \qquad (43)
\end{aligned}$$

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. 3(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 judges an ellipse to be a line over a wide range of β.

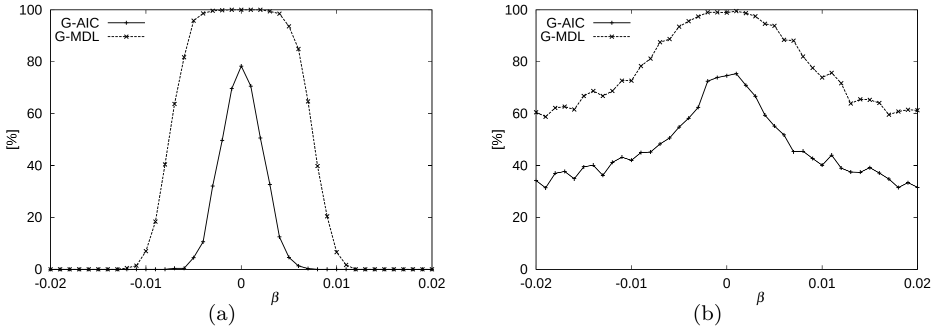
In Fig. 3(a), we used the true value of the noise variance ϵ^2. If it is unknown, it can be estimated from the residual of the general ellipse model. Fig. 3(b) shows the result using its estimate. Although the sharpness is somewhat lost, we observe similar performance characteristics of the geometric AIC and the geometric MDL.

Fig. 4 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 $\beta < 1.1$.

6.4 Virtual studio

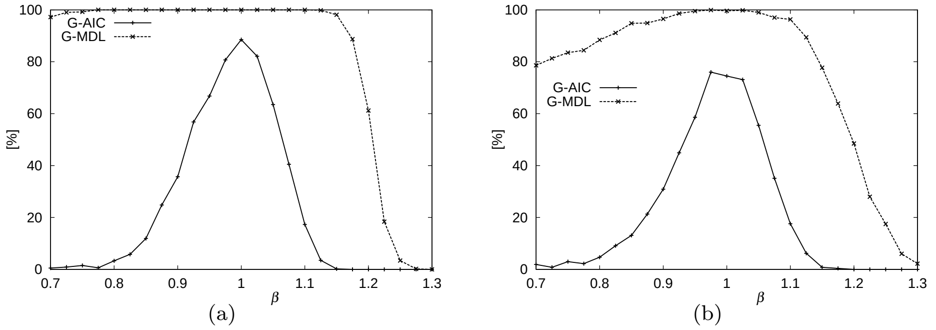
We now do *virtual studio* experiments, taking images of moving objects such as persons by a moving camera and superimposing them in a graphics-generated background [6, 27, 32].

In order to generate a background image that is compatible to the moving viewpoint, we need to compute, at each frame in real time, the position and zooming of the camera, which can arbitrarily change in the course



(a) (b)

Figure 3: 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.



(a) (b)

Figure 4: 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.

of the shooting. A typical technique for this is to place a grid pattern colored in light and dark blue behind the object and separate the object image from the pattern image by a chromakey technique. With the true geometry of the grid pattern known, the position and the focal length of the camera can be determined if four or more grid points are detected in the grid pattern image [22, 27, 31]. However, the following two problems must be resolved:

1. When the camera optical axis is perpendicular to the pattern, the 3-D position and focal length of the camera are indeterminate because zooming out and moving the camera forward cause the same visual effect.
2. Some unoccluded grid points become occluded while some occluded points become unoccluded as the object moves in the scene. As a result, the computed camera position fluctuates even if the camera is stationary or moving very slowly.

These problems, often dealt with by ad hoc measures in the past, can be resolved by model selection: we model various modes of camera motion and zooming that are likely to occur and choose at each frame the most appropriate one by model selection [22].

Fig. 5 shows five sampled frames from a real image sequence. Unoccluded grid points in the image were matched to their true positions in the pattern by observing the cross ratio of adjacent points. This pattern is so designed that the cross ratio is different everywhere in such a way that matching can be done in a statistically optimal way in the presence of image noise [23].

Fig. 6 shows the estimated focal lengths and the estimated camera trajectory viewed from above. Here, we used the following models (see [22] for the details of the computation):

- The camera is stationary with fixed zooming.
- The camera rotates with fixed zooming.
- The camera linearly moves with fixed zooming.
- The camera moves arbitrarily with fixed zooming.
- The camera moves arbitrarily with linearly changing zooming.
- Everything changes.

Degeneracy was detected in the 15th frame, and the frame-wise estimation failed thereafter. We can observe that the geometric MDL tends to select a simpler model and define a more rigid trajectory than the geometric AIC, which tends select a general model and define a more flexible trajectory.

6.5 Observations

From the experiments we have done, 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 probability 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*

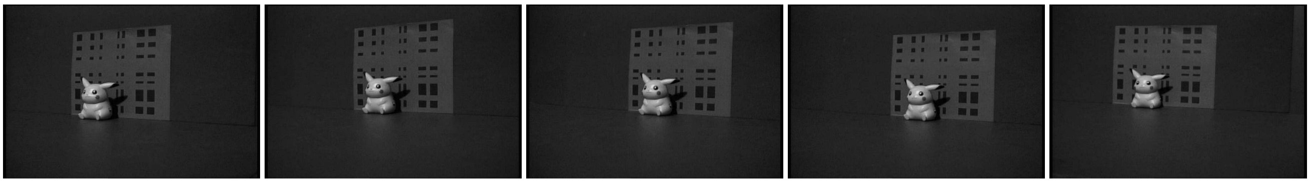
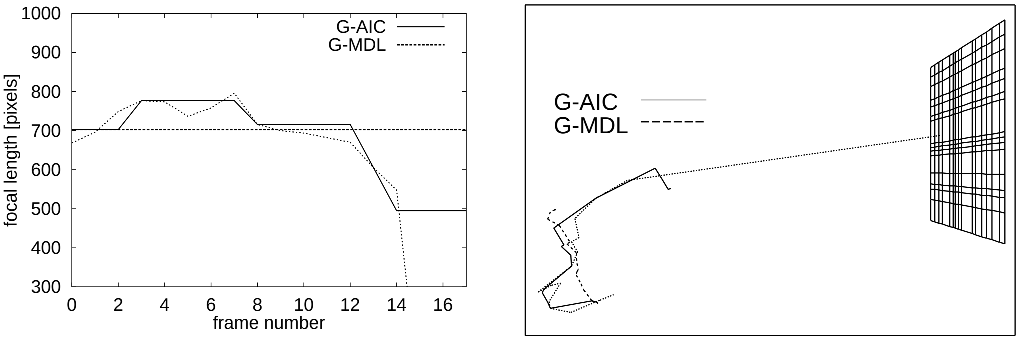


Figure 5: Sampled frames from a real image sequence.



(a) (b)

Figure 6: (a) Estimated focal lengths. (b) Estimated camera trajectory. In (a) and (b), the solid lines are for model selection by the geometric AIC; the thick dashed lines are for model selection by the geometric MDL; the thin dotted lines are for estimation without model selection.

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 non-degenerate models as degenerate. This is natural, since the penalty for one degree of freedom is heavier in the geometric MDL than in the geometric AIC (see. eqs. (14) and (35)). Thus, the geometric AIC is more faithful to the data than the geometric MDL, which is more likely to choose a degenerate model.

In the virtual studio example, the estimation by the geometric MDL appears more consistent with the actual camera motion than the geometric AIC. But this is because we fixed the zooming and moved the camera smoothly. If we added variations to the zooming and the camera motion, the geometric MDL would still prefer a smooth motion. So, we cannot say which solution should be *closer* to the the true solution; it depends on what kind of solution *we expect* is desirable for the application in question.

7. Concluding Remarks

In this paper, we formulated geometric fitting as constraint satisfaction of geometric data in the presence of noise, taking the noise level as the asymptotic variable, in contrast to statistical inference whose aim is to give a good description of random phenomena in terms of deterministic mechanisms and random noise with the number of observations taken as the asymptotic variable. Then, we gave a new definition of the geometric AIC and the geometric MDL as counterparts of Akaike's AIC and Rissanen's MDL. We discussed various problems in using them in practical situations. Finally, we experimentally showed that the geometric MDL does not necessarily outperform the geometric AIC and that the two criteria have very different characteristics.

If we take in geometric fitting the number N of data $\{\boldsymbol{x}_\alpha\}$ as the asymptotic variable, the number of unknown parameters $\{\bar{\boldsymbol{x}}_\alpha\}$ (true values of the data) increases as N increases. In other words, if we add one datum $\boldsymbol{x}_{N+1}$, we have a new problem with a new set of unknowns, whose instance we observe *once*. As a result, the asymptotic behavior of estimation in the limit $N \to \infty$ becomes very anomalous. For this reason, $\{\bar{\boldsymbol{x}}_\alpha\}$ are often called the *nuisance parameters*. One way to avoid such an anomaly is to view $\{\bar{\boldsymbol{x}}_\alpha\}$ as "random samples" from a yet unknown distribution and regard, instead of $\{\bar{\boldsymbol{x}}_\alpha\}$ themselves, the (hyper)parameters of that distribution as the unknowns to be estimated. Such a description is called a *semiparametric model* [2].

Such an approach is effective for dealing with problems where one can observe, at least in principle, as many data as possible, a typical situation being time series analysis. For example, the problem of estimating the number of signal sources from time series data can be reduced to estimating the rank of a matrix determined from the time series, and one can use Akaike's AIC or Rissanen's MDL for that purpose [39]. In such a problem, the goal is to estimate something with maximum accuracy using a minimum number of data. In many computer vision problems, in contrast, the goal is to estimate something with maximum accuracy using devices with minimum resolution. The theory in this paper is intended to such applications as illustrated in the examples we have given.

Acknowledgement. The author thanks David Suter of Monash University, Australia, Pierre-Louis Bazin of Brown University, U.S.A, Shun-ichi Amari of Riken Brain Science Institute, Japan, Te-Sun Han of the University of Electro-Communications, Japan, Yoichi Seki of Gunma University, Japan, Jun-ichi Takeuchu of NEC, and Chikara Matsunaga of For-A, Japan for helpful discussions. 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. 13680432).

References

[1] H. Akaike, "A new look at the statistical model identification," *IEEE Trans. Automatic Control*, vol. 16, no. 6, pp. 716–723, 1974.

[2] 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.

[3] K. Bubna and C. V. Stewart, "Model selection techniques and merging rules for range data segmentation algorithms," *Comput. Vision Image Understand.*, vol. 80, no. 2, pp. 215–245, 2000.

[4] J. P. Costeira and T. Kanade, "A multibody factorization method for independently moving objects," *Int. J. Comput. Vision*, vol. 29, no. 3, pp. 159–179, 1998.

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

[6] S. Gibbs, C. Arapis, C. Breiteneder, V. Lalioti, S. Mostafawy and J. Speier, "Virtual studios: An overview," *IEEE Multimedia*, vol. 5, no. 1, pp. 24–35, 1998.

[7] 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.*, vol. 18, no. 1, pp. 58–64, 1996.

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

[9] —, "Cramer-Rao lower bounds for curve fitting," *Graphical Models Image Process.*, vol. 60, no. 2, pp. 93–99, 1998.

[10] —, "Self-evaluation for active vision by the geometric information criterion," in *Proc. 7th Int. Conf. Computer Analysis of Images and Patterns*, Kiel, Germany, September 1997, pp. 247–254.

[11] —, "Geometric information criterion for model selection," *Int. J. Comput. Vision*, vol. 26, no. 3, pp. 171–189, 1998.

[12] —, "Statistical optimization and geometric inference in computer vision," *Phil. Trans. Roy. Soc. Lond.*, ser. A, vol. 356, pp. 1303–1320, 1998.

[13] —, "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.

[14] —, "Motion segmentation by subspace separation and model selection," in *Proc. 8th Int. Conf. Comput. Vision*, Vancouver, Canada, vol. 2, July 2001, pp. 301–306.

[15] K. Kanatani and C. Matsunaga, "Estimating the number of independent motions for multibody motion segmentation," in *Proc. 5h Asian Conf. Comput. Vision*, Melbourne, Australia, January 2002.

[16] Y. Kanazawa and K. Kanatani, "Optimal line fitting and reliability evaluation," *IEICE Trans. Inf. & Syst.*, vol. E79-D, no. 9, pp. 1317–1322, 1996.

[17] —, "Optimal conic fitting and reliability evaluation," *IEICE Trans. Inf. & Syst.*, vol. E79-D, no. 9, pp. 1323–1328, 1996.

[18] —, "Infinity and planarity test for stereo vision," *IEICE Trans. Inf. & Syst.*, vol. E80-D, no. 8, pp. 774–779, 1997.

[19] —, "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.

[20] —, "Do we really have to consider covariance matrices for image features?" in *Proc. 8th Int. Conf. Computer Vision* , Vancouver, Canada, July 2001, vol. 2, pp. 586–591.

[21] Y. G. Leclerc, "Constructing simple stable descriptions for image partitioning," *Int. J. Comput. Vision*, vol. 3, no. 1, pp. 73–102, 1989.

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

[23] C. Matsunaga, Y. Kanazawa and K. Kanatani, "Optimal grid pattern for automated camera calibration using cross ratio," *IEICE Trans. Inf. & Syst.*, vol. E83-A, no. 10, pp. 1921–1928, 2000.

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

[25] B. A. Maxwell, "Segmentation and interpretation of multicolored objects with highlights," *Comput. Vision Image Understand.*, vol. 77, no. 1, pp. 1–24, 2000.

[26] N. Ohta and K. Kanatani, "Moving object detection from optical flow without empirical thresholds," *IEICE Trans. Inf. & Syst.*, vol. E81-D, no. 2, pp. 243–245, 1998.

[27] S.-W. Park, Y. Seo and K.-S. Hong, "Real-time camera calibration for virtual studio," *Real-Time Imaging*, vol. 6, no. 6, pp. 433–448, 2000.

[28] J. Rissanen, "Universal coding, information, prediction and estimation," *IEEE Trans. Inform. Theory*, vol. 30, no. 4, pp. 629–636, 1984.

[29] —, *Stochastic Complexity in Statistical Inquiry*, World Scientific, Singapore, 1989.

[30] —, "Fisher information and stochastic complexity," IEEE Trans. Inform. Theory, vol. 42, no. 1, pp. 40–47, 1996.

[31] Y. Seo, M.-H. Ahn and K.-S. Hong, "A multiple view approach for auto-calibration of a rotating and zooming camera," *IEICE Trans. Inf. & Syst.*, vol. E83-D, no. 7, pp. 1375–1385, 2000.

[32] M. Tamir, "The Orad virtual set," *Int. Broadcast Eng.*, pp. 16–18, March 1996.

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

[34] —, "Geometric motion segmentation and model selection," *Phil. Trans. Roy. Soc. Lond.*, ser. A, vol. 356, pp. 1321–1340, 1998.

[35] —, "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.

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

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

[38] Iman Triono, N. Ohta and K. Kanatani, "Automatic recognition of regular figures by geometric AIC," *IEICE Trans. Inf. & Syst.*, vol. E81-D, no. 2, pp. 246–248, 1998.

[39] M. Wax and T. Kailath, "Detection of signals by information theoretic criteria," *IEEE Trans. Acoust. Speech, Signal Process.*, vol. 33, no. 2, pp. 387–392, 1985.