PAPER

Geometric BIC

Kenichi KANATANI[†a)], ***Member***

SUMMARY The "geometric AIC" and the "geometric MDL" have been proposed as model selection criteria for geometric fitting problems. These correspond to Akaike's "AIC" and Rissanen's "BIC" well known in the statistical estimation framework. Another well known criterion is Schwarz' "BIC", but its counterpart for geometric fitting has not been known. This paper introduces the corresponding criterion, which we call the "geometric BIC", and shows that it is of the same form as the geometric MDL. Our result gives a justification to the geometric MDL from the Bayesian principle.

key words: *geometric model selection, AIC, BIC, MDL, Bayesian estimation, degeneracy detection*

1. Introduction

The basic principle of computer vision is to assume a certain structure, or a *model*, in the observed scene, such as certain objects being there, and to do inference by extracting characteristics of the assumed structure from observed images, estimating such properties of the scene as categories, numbers, sizes, shapes, positions, and orientations. However, we sometimes do not know what the model should be. In such a case, selecting an appropriate model from multiple candidates, called *model selection*, is necessary.

For models having a form of standard statistical estimation, such as regression, various types of (*statistical*) *model selection criteria* have been proposed. The best known are Akaike's *AIC* (*Akaike Information Criterion*) [1], Schwarz' *BIC* (*Bayesian Information Criterion*) [19], and Rissanen's *MDL* (*Minimum Description Length*) [18].

However, geometric inference for computer vision, typically structure from motion, does not have the standard form of statistical estimation [4], [11], [12]. For this, the *geometric AIC* [4], [8] and the *geometric MDL* [11] have been introduced, corresponding to Akaike's AIC and Rissanen's MDL.

The main motivation of traditional statistical estimation is the ability to make precise inference using a large but limited number of data, while the main goal of geometric inference for computer vision is to do precise but robust estimation that can tolerate noise [4], [11], [12]. This is a sort of "dual" relationship. Hence, while the AIC and the MDL are derived from asymptotic analysis with respect to the number N of data, the geometric AIC and the geometric MDL are derived from perturbation analysis with respect to the noise level ε [4], [11], [12].

Then, a question arises. What corresponds to Schwarz' BIC? The BIC is also derived from asymptotic analysis with respect to the number N of data. What criterion results if we do perturbation analysis with respect to the noise level ε? It has already been conjectured [11] that because the BIC and the MDL have the same form up to higher-order terms in $1/\sqrt{N}$, the geometric AIC and the "geometric BIC" will have the same form up to higher-order terms in ε. However, the concrete form has not yet been shown.

In this paper, we present a rigorous derivation of the geometric BIC based on Schwarz' BIC principle and confirm that it indeed has the same form as the geometric MDL. This has the following implications. First, our derivation illuminates the Bayesian logic of model selection for geometric estimation. The fact that we arrive at the same form as the geometric MDL is no surprise but rather is a reassuring evidence that the underlying logic and the derivation are correct. At the same time, it also justifies the geometric MDL, whose axiomatic origin has some arbitrariness, from the Bayesian standpoint. Today, the Bayesian principle is thought to be more fundamental that the MDL principle [3].

We begin with a overview of Akaike's AIC, Schwarz' BIC, and Rissanen's MDL (Sect. 2) and a summary of the geometric AIC and the geometric MDL (Sect. 3). Then, we describe the mathematical framework of geometric fitting (Sect. 4). The central part of this paper is Sect. 5, where we derive the geometric BIC. We discuss its applications (Sect. 6) and conclude (Sect. 7).

2. AIC, BIC, and MDL

We first give a brief summary of the AIC, the BIC, and the MDL. A probability density $p(\boldsymbol{x}|\boldsymbol{\theta})$ parameterized by unknown $\boldsymbol{\theta}$ is called a (*statistical*) *model*. The goal of *statistical estimation* is to estimate $\boldsymbol{\theta}$ from multiple data $\boldsymbol{x}_1$, ..., $\boldsymbol{x}_N$ assumed to be independently sampled from $p(\boldsymbol{x}|\boldsymbol{\theta})$. *Maximum likelihood* (*ML*) is to find the value of $\boldsymbol{\theta}$ that maximizes the likelihood $\prod_{\alpha=1}^{N} p(\boldsymbol{x}_\alpha|\boldsymbol{\theta})$. When we have multiple candidate models[*] $p_1(\boldsymbol{x}|\boldsymbol{\theta})$, ..., $p_M(\boldsymbol{x}|\boldsymbol{\theta})$, (*statistical*) *model selection* is to find the most appropriate one from among them. The best known criteria are

Manuscript received March 23, 2009.
Manuscript revised July 23, 2009.

[†] The author is with the Department of Computer Science, Okayama University, Okayama-shi, 700-8530 Japan.
a) E-mail: kanatani@suri.cs.okayama-u.ac.jp

DOI: 10.1587/transinf.E93.D.1

[*] The same symbol $\boldsymbol{\theta}$ is used for the convenience of description, but it may have a different dimension from model to model.

$$\text{AIC} = -2\sum_{\alpha=1}^{N} \log p(\boldsymbol{x}_\alpha|\hat{\boldsymbol{\theta}}) + 2k, \tag{1}$$

$$\text{BIC} = -\sum_{\alpha=1}^{N} \log p(\boldsymbol{x}_\alpha|\hat{\boldsymbol{\theta}}) + \frac{k}{2}\log N, \tag{2}$$

$$\text{MDL} = -\sum_{\alpha=1}^{N} \log p(\boldsymbol{x}_\alpha|\hat{\boldsymbol{\theta}}) + \frac{k}{2}\log N, \tag{3}$$

where k is the degree of freedom of the model (= the dimension of $\boldsymbol{\theta}$) and $\hat{\boldsymbol{\theta}}$ is the ML estimator of $\boldsymbol{\theta}$ obtained by assuming that model. These criteria are computed for each candidate model, and the one that has the smallest value is adopted as the most appropriate.

AIC. Akaike's AIC principle is to choose the model that is the closest to the true model measured in the *Kullback-Leibler* (*KL*) *distance* (or *divergence*) [1]. We hypothesize that the true model is given by $p(\boldsymbol{x}|\boldsymbol{\theta})$ with some unknown θ, to which the ML estimator $\hat{\boldsymbol{\theta}}$ is plugged in. This is justified when the number N of data is large (*consistency* of ML). The KL distance is defined via expectation with respect to the true model. Since it is unknown, the expectation is approximated by summation over the data, which is justified when N is large (the *law of large numbers*). However, if the *same* data are used for computing the ML estimator $\hat{\boldsymbol{\theta}}$ and approximating the expectation, mutual correction gives rise to statistical bias. Akaike [1] estimated the bias by doing asymptotic expansion, assuming that N is large and omitting high order terms in $1/\sqrt{N}$. Subtracting the estimated bias from the KL distance estimate, he obtained his AIC in the form of Eq. (1) up to model-independent terms.

BIC. Schwarz' BIC principle is to assume an a prior probability of the model, evaluate the a posteriori probability using the Bayes theorem, and choose the model that has the largest value of it. Schwarz [19] assumed equal priors for the candidate models and analyzed asymptotic expansion of the (logarithmic) posterior (the *Laplace expansion*), noting that the distribution of $\boldsymbol{\theta}$ concentrates on a small neighborhood of the ML estimator $\hat{\boldsymbol{\theta}}$ when N is large (the *central limit theorem*). Omitting higher-order terms in $1/\sqrt{N}$ and excluding model-independent terms, he obtained his BIC in the form of Eq. (2) independent of the a priori probability of $\boldsymbol{\theta}$.

MDL. Rissanen's MDL principle is to choose the model that gives the shortest description when it is optimally encoded along with the data [18]. According to information theory, the data $\{\boldsymbol{x}_\alpha\}$ are optimally encoded using its occurrence probability $p(\boldsymbol{x}|\boldsymbol{\theta})$, but since the true value of $\boldsymbol{\theta}$ is unknown, the ML estimator $\hat{\boldsymbol{\theta}}$ is substituted. However, the data $\{\boldsymbol{x}_\alpha\}$ and the ML estimator $\hat{\boldsymbol{\theta}}$ are both real numbers, which require an infinite description length. So, they are quantized into discrete values, and the quantization width is determined so that the resulting code length is the shortest. As the model is better approximated (i.e., $\hat{\boldsymbol{\theta}}$ is approximated to higher accuracy), the code length of the data $\{\boldsymbol{x}_\alpha\}$ becomes shorter and approaches the information theoretical limit (*Shannon's theorem*). At the same time, good description of the model (i.e., high accuracy approximation of $\hat{\boldsymbol{\theta}}$) requires a larger code length. Rissanen [18] evaluated their optimal balance, analyzed its asymptotic expansion, omitting higher-order terms in $1/\sqrt{N}$, and obtained his MDL in the form of Eq. (3) up to model-independent terms[†].

3. Geometric AIC and Geometric MDL

Next, we briefly summarize the geometric AIC and the geometric MDL. Given N data $\{\boldsymbol{x}_\alpha\}$, *geometric fitting* is the problem of estimating the law (or the *constraint*) that governs their true values $\{\bar{\boldsymbol{x}}_\alpha\}$ in the form of an "implicit" equation

$$\boldsymbol{F}(\boldsymbol{x};\boldsymbol{u}) = \boldsymbol{0}, \tag{4}$$

parameterized by unknown $\boldsymbol{u}$. This equation is called the (*geometric*) *model*. Many computer vision problems fall in this category. For example, we may want to fit a parametric curve to a point sequence (x_α, y_α), $\alpha = 1, ..., N$. Or we may want to compute the fundamental matrix or the homography from point correspondences (x_α, y_α), (x'_α, y'_α), $\alpha = 1, ..., N$, over two views [2]. By estimating the parameter $\boldsymbol{u}$ (e.g., coefficients of the curve equations, the fundamental matrix, or the homography) so that Eq. (4) fits the data $\{\boldsymbol{x}_\alpha\}$ well, the structure of the scene or its motion can be inferred [2].

When we have multiple candidate models[††] $\boldsymbol{F}_1(\boldsymbol{x},\boldsymbol{u}) = \boldsymbol{0}$, ..., $\boldsymbol{F}_M(\boldsymbol{x},\boldsymbol{u}) = \boldsymbol{0}$, (*geometric*) *model selection* is to find the most appropriate one from among them. For this, the following geometric AIC and the geometric MDL have been proposed [4], [8], [11]:

$$\text{G-AIC} = \hat{J} + 2(Nd+p)\varepsilon^2, \tag{5}$$

$$\text{G-MDL} = \hat{J} - (Nd+p)\varepsilon^2 \log \varepsilon^2. \tag{6}$$

Here, $\hat{J}$ is the residual (the sum of squares of the Mahalanobis distances) of the fitted model from the data $\{\boldsymbol{x}_\alpha\}$, d is the dimension of the manifold defined by the model, p is the degree of freedom of the model (= the dimension of $\boldsymbol{u}$), and ε is the noise level (their precise definitions are given later).

Geometric AIC. The geometric AIC is derived from Akaike's AIC principle by assuming Gaussian noise and minimizing the KL distance of the candidate model from the true model. Since the true model is unknown, we replace the true values $\{\bar{\boldsymbol{x}}_\alpha\}$ and the unknown $\boldsymbol{u}$ by their ML estimators $\{\hat{\boldsymbol{x}}_\alpha\}$ and $\hat{\boldsymbol{u}}$. We then evaluate the bias and subtract it. The difference from the AIC is that while the AIC is based on the asymptotic expansion in $1/\sqrt{N}$, the geometric AIC is obtained by perturbation expansion in the noise level ε. The integration for evaluating the KL distance is approximated

[†] Equation (3) is a crude approximation. A more detailed form involves the Fisher information matrix $\boldsymbol{I}(\boldsymbol{\theta})$ [18].

[††] As before, the same symbol $\boldsymbol{u}$ is used for convenience, but it may have a different dimension from model to model.

by summation over data. This is justified because the true values $\{\bar{\boldsymbol{x}}_\alpha\}$ are very close to their ML estimators $\{\hat{\boldsymbol{x}}_\alpha\}$ when ε is small. Omitting higher-order terms in ε and excluding model-independent terms, we obtain the geometric AIC in the form of Eq. (5) [4], [8].

Geometric MDL. The geometric MDL is derived from Rissanen's MDL principle by assuming Gaussian noise and minimizing the description length of both the data and the model when optimally encoded. The data $\{\boldsymbol{x}_\alpha\}$ and the ML estimators $\{\hat{\boldsymbol{x}}_\alpha\}$ and $\hat{\boldsymbol{u}}$ are quantized, and the quantization width is determined so that the resulting code length is the shortest. The difference from the MDL is that while the MDL is based on the asymptotic expansion in $1/\sqrt{N}$, the geometric MDL is obtained by perturbation expansion in the noise level ε. Omitting higher-order terms in ε and excluding model-independent terms, we obtain the geometric MDL in the form of Eq. (3) [11].

4. Geometric Fitting

We now describe the mathematical framework in which the geometric BIC is to be derived.

4.1 Noise Modeling

Let $\{\boldsymbol{x}_\alpha\}$, $\alpha = 1, ..., N$, be m-dimensional vector data[†], which are assumed to be purturbed from their true values $\{\bar{\boldsymbol{x}}_\alpha\}$ by independent Gaussian noise of mean $\mathbf{0}$ and covariance matrix

$$V[\boldsymbol{x}_\alpha] = \varepsilon^2 V_0[\boldsymbol{x}_\alpha], \tag{7}$$

where ε, which we call the *noise level*, is an average uncertainty of data observation independent of individual $\boldsymbol{x}_\alpha$. The matrix $V_0[\boldsymbol{x}_\alpha]$, which we call the *normalized covariance matrix*, describes the relative uncertainty of observing that particular $\boldsymbol{x}_\alpha$. The goal of geometric fitting is to find a method whose accuracy qickly increases as $\varepsilon \to 0$, because such a method can torelate larger uncertainty than others.

Note that we mean by "noise" the uncertainty of detecting $\boldsymbol{x}_\alpha$, say using an image processing operator [11], [12]. The uncertainty is not "random" in the usual sense[††]; it is merely unknown. Also note that $V_0[\boldsymbol{x}_\alpha]$ is not a "function" of $\boldsymbol{x}_\alpha$. It is the uncertainty of the "observation process" for finding $\boldsymbol{x}_\alpha$; it does not depend on what actual value we find for $\boldsymbol{x}_\alpha$.

The important thing is that ε and $V_0[\boldsymbol{x}_\alpha]$ are properties of our data observation process *independent* of which model we assume. If ε and $V_0[\boldsymbol{x}_\alpha]$ are unknown, they need to be estimated from some general knowledge that apply to all the candidate models. This is one of the differences from the traditional statistical estimation process.

4.2 Maximum Likelihood

Suppose Eq. (4) is an r-dimensional equation. We write it componentwise as

$$F^{(k)}(\boldsymbol{x}; \boldsymbol{u}) = 0, \qquad k = 1, ..., r. \tag{8}$$

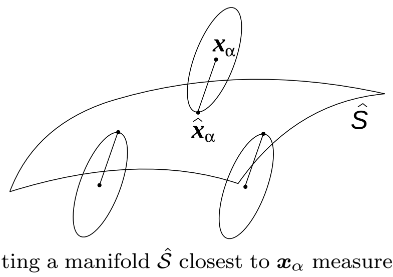

Fig. 1 Fitting a manifold $\hat{\mathcal{S}}$ closest to $\boldsymbol{x}_\alpha$ measured in the Mahalanobis distance. The point $\hat{\boldsymbol{x}}_\alpha$ on it closest to $\boldsymbol{x}_\alpha$ in the Mahalanobis distance is its ML estimator. The ellipsoids represent equal probability surfaces $(\boldsymbol{x}_\alpha - \hat{\boldsymbol{x}}_\alpha, V_0[\boldsymbol{x}_\alpha]^{-1}(\boldsymbol{x}_\alpha - \hat{\boldsymbol{x}}_\alpha)) =$ constant.

For a given $\boldsymbol{u}$, these r equations define a manifold (an algebraic variety) $\mathcal{S}$ in the m-dimensional space $\mathcal{X}$ of the variable $\boldsymbol{x}$, which we call the *data space*. If the r equations in Eq. (8) are algebraically independent[†††], the manifold $\mathcal{S}$ has dimension $d = m - r$. Geometric fitting is regarded as the problem of adjusting $\boldsymbol{u}$ so that $\mathcal{S}$ becomes close to the data $\{\boldsymbol{x}_\alpha\}$.

From our Gaussian noise assumption, the probability density of the data $\{\boldsymbol{x}_\alpha\}$ given their true values $\{\bar{\boldsymbol{x}}_\alpha\}$ and the parameter $\boldsymbol{u}$ is

$$p(\{\boldsymbol{x}_\alpha\}|\{\bar{\boldsymbol{x}}_\alpha\}, \boldsymbol{u}) = \frac{e^{-J/2\varepsilon^2}}{\sqrt{(2\pi)^{Nm}\varepsilon^{2Nm}|V_0[\boldsymbol{x}_\alpha]|^N}}, \tag{9}$$

where we define

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

Throughout this paper, we denote the inner product of vectors $\boldsymbol{a}$ and $\boldsymbol{b}$ by $(\boldsymbol{a}, \boldsymbol{b})$. If regarded as a function of $\{\bar{\boldsymbol{x}}_\alpha\}$ and $\boldsymbol{u}$, Eq. (9) is the likelihood of the data $\{\boldsymbol{x}_\alpha\}$. The values $\{\bar{\boldsymbol{x}}_\alpha\}$ and $\boldsymbol{u}$ that maximize Eq. (9) are their ML estimators; they are the minimizer of the function J in Eq. (10) subject to

$$F^{(k)}(\bar{\boldsymbol{x}}_\alpha; \boldsymbol{u}) = 0, \qquad k = 1, ..., r, \quad \alpha = 1, ..., N. \tag{11}$$

Let $\hat{J}$ be the resulting minimum of J.

Geometrically, ML is to adjust $\boldsymbol{u}$ so that the manifold $\mathcal{S}$ associated with the model is closest to points $\boldsymbol{x}_\alpha$ in the data space measured in the sum of the square Mahalanobis distances in Eq. (10) (Fig. 1). The resulting value $\hat{\boldsymbol{u}}$ is the ML estimator of $\boldsymbol{u}$, and the points $\hat{\boldsymbol{x}}_\alpha$ on the fitted manifold $\hat{\mathcal{S}}$ closest to $\boldsymbol{x}_\alpha$ measured in Eq. (10) are their ML estimators.

[†] The following argument holds if each $\boldsymbol{x}_\alpha$ is constrained to have a smaller degree m' $(< m)$ of freedom, e.g., being a unit vector. We only need to introduce degenerate covariance matrices, pseudoinverse, and projection on to the constrained space [4]. Here, for simplicity, we assume that no such intrinsic constraints exist.

[††] For example, if we repeat the observation, we always find the same value. This is the fundamental difference from the traditional statistical estimation [11], [12].

[†††] The following argument holds if the r equations in Eq. (8) has redundancies with only r' $(< r)$ being independent, as long as the r equations define hypersurfaces with nonsingular (or *transversal*) intersections. [4]. We only need to introduce pseudoinverse and projection operation operators [4]. Here, for simplicity, we assume that the r equations are independent.

4.3 Two-Stage Estimation

We compute ML in two stages. First, we fix $\boldsymbol{u}$ and minimize Eq. (10) with respect to $\{\bar{\boldsymbol{x}}_\alpha\}$ subject to Eq. (11). Let $\{\tilde{\boldsymbol{x}}_\alpha(\boldsymbol{u})\}$ be the solution, and $\tilde{J}(\boldsymbol{u})$ the resulting minimum of Eq. (10). Next, we minimize

$$\tilde{J}(\boldsymbol{u})=\sum_{\alpha=1}^{N}(\boldsymbol{x}_\alpha-\tilde{\boldsymbol{x}}_\alpha(\boldsymbol{u}),V_0[\boldsymbol{x}_\alpha]^{-1}(\boldsymbol{x}_\alpha-\tilde{\boldsymbol{x}}_\alpha(\boldsymbol{u}))), \quad (12)$$

with respect to $\boldsymbol{u}$; we no longer need to consider Eq. (11), which is identically satisfied by $\{\tilde{\boldsymbol{x}}_\alpha(\boldsymbol{u})\}$. The value $\hat{\boldsymbol{u}}$ that minimizes Eq. (12) is the ML estimator of $\boldsymbol{u}$. The corresponding $\{\tilde{\boldsymbol{x}}_\alpha(\hat{\boldsymbol{u}})\}$ are the ML estimators of $\{\bar{\boldsymbol{x}}_\alpha\}$, and $\tilde{J}(\hat{\boldsymbol{u}})$ equals the minimum $\hat{J}$ of J.

4.4 A Posteriori Covariance Matrices

Since $\{\tilde{\boldsymbol{x}}_\alpha(\boldsymbol{u})\}$ identically satisfy Eq. (11), they are constrained to be in the d-dimensional manifold $\mathcal{S}$ in the data space $\mathcal{X}$. Hence, although the original data $\{\boldsymbol{x}_\alpha\}$ have m degrees of freedom, the points $\{\tilde{\boldsymbol{x}}_\alpha(\boldsymbol{u})\}$ have only d degrees of freedom. They are projections of $\{\boldsymbol{x}_\alpha\}$ onto $\mathcal{S}$ through the Mahalanobis distance minimization. The normalized covariance matrix $V_0[\tilde{\boldsymbol{x}}_\alpha]$ of $\tilde{\boldsymbol{x}}_\alpha(\boldsymbol{u})$ is defined by the associated projection of $\boldsymbol{x}_\alpha(\boldsymbol{u})$ onto the tangent space $T_{\tilde{\boldsymbol{x}}_\alpha}(\mathcal{S})$ to $\mathcal{S}$ at $\tilde{\boldsymbol{x}}_\alpha$ and has the following form[†][4], [8]:

$$V_0[\tilde{\boldsymbol{x}}_\alpha]=V_0[\boldsymbol{x}_\alpha] - \sum_{k,l=1}^{r}W_\alpha^{(kl)}(V_0[\boldsymbol{x}_\alpha]\nabla_{\boldsymbol{x}}F_\alpha^{(k)})(V_0[\boldsymbol{x}_\alpha]\nabla_{\boldsymbol{x}}F_\alpha^{(l)})^\top. \quad (13)$$

Here, $\nabla_{\boldsymbol{x}}F^{(k)}$ denotes the gradient of $F^{(k)}$ with respect to $\boldsymbol{x}$. The subscript α means its evaluation at $\boldsymbol{x}=\boldsymbol{x}_\alpha$, and $W_\alpha^{(kl)}$ is the (kl) element of the inverse of the $r\times r$ matrix whose (kl) element is $(\nabla_{\boldsymbol{x}}F_\alpha^{(k)},V_0[\boldsymbol{x}_\alpha]\nabla_{\boldsymbol{x}}F_\alpha^{(l)})$: symbolically, we write

$$\left(W_\alpha^{(kl)}\right)=\left((\nabla_{\boldsymbol{x}}F_\alpha^{(k)},V_0[\boldsymbol{x}_\alpha]\nabla_{\boldsymbol{x}}F_\alpha^{(l)})\right)^{-1}. \quad (14)$$

Note that $V_0[\tilde{\boldsymbol{x}}_\alpha]$ in Eq. (13) is an $m\times m$ matrix but has rank d $(<m)$, because it is the projection of $V_0[\boldsymbol{x}_\alpha]$ onto the d-dimensional tangent space $T_{\tilde{\boldsymbol{x}}_\alpha}(\mathcal{S})$ to $\mathcal{S}$.

The posterior covariance matrix of the ML estimator $\hat{\boldsymbol{u}}$ of $\boldsymbol{u}$ is evaluated as follows [4], [8]:

$$V[\hat{\boldsymbol{u}}]=\varepsilon^2\hat{\boldsymbol{M}}^{-1}+O(\varepsilon^4), \quad (15)$$

$$\hat{\boldsymbol{M}}=\sum_{\alpha=1}^{N}\sum_{k,l=1}^{r}\hat{W}_\alpha^{(kl)}\nabla_{\boldsymbol{u}}\hat{F}_\alpha^{(k)}\nabla_{\boldsymbol{u}}\hat{F}_\alpha^{(l)\top}. \quad (16)$$

Here, $\nabla_{\boldsymbol{u}}F^{(k)}$ denotes the gradient of $F^{(k)}$ with respect to $\boldsymbol{u}$, and the subscript α means evaluation at $\boldsymbol{x}=\boldsymbol{x}_\alpha$. The hats on $W_\alpha^{(kl)}$ and $F_\alpha^{(k)}$ mean substitution of $\hat{\boldsymbol{u}}$ for $\boldsymbol{u}$.

If $\boldsymbol{x}_\alpha$ and $\hat{\boldsymbol{u}}$ in the expression of $\hat{\boldsymbol{M}}$ in Eq. (16) are replaced by their true values $\bar{\boldsymbol{x}}_\alpha$ and $\boldsymbol{u}$, respectively, the first term on the right-hand side of Eq. (15) gives the *KCR lower bound* on the covariance matrix of any unbiased estimator of $\boldsymbol{u}$ [4], [9], [12].

5. Geometric BIC

The above mathematical framework is the same as that used to derive the geometric AIC and the geometric MDL [4], [8], [11]. We newly derive the geometric BIC in the same framework.

5.1 A Priori and A Posteriori Probabilities

Suppose we have M models $\mathcal{M}_1$, ..., $\mathcal{M}_M$. Let $p(\mathcal{M}_i)$ be the a priori probability for the model $\mathcal{M}_i$, $p(\boldsymbol{u}|\mathcal{M}_i)$ the a priori probability density of its parameter $\boldsymbol{u}$, and $p(\{\bar{\boldsymbol{x}}_\alpha\}|\boldsymbol{u},\mathcal{M}_i)$ the a priori probability density[††] of the true values $\{\bar{\boldsymbol{x}}_\alpha\}$ given $\boldsymbol{u}$.

The likelihood $p(\{\boldsymbol{x}_\alpha\}|\{\bar{\boldsymbol{x}}_\alpha\},\boldsymbol{u},\mathcal{M}_i)$ of the data $\{\boldsymbol{x}_\alpha\}$ given the parameter $\boldsymbol{u}$ and the true values $\{\bar{\boldsymbol{x}}_\alpha\}$ for model $\mathcal{M}_i$ is Eq. (9). According to the Bayes theorem, the a posteriori probability $p(\mathcal{M}_i|\{\boldsymbol{x}_\alpha\})$ of model $\mathcal{M}_i$ given the data $\{\boldsymbol{x}_\alpha\}$ is

$$\begin{aligned}p(\mathcal{M}_i|\{\boldsymbol{x}_\alpha\})=&\iint\cdots\int p(\{\boldsymbol{x}_\alpha\}|\{\bar{\boldsymbol{x}}_\alpha\},\boldsymbol{u},\mathcal{M}_i)\\ &\times p(\{\bar{\boldsymbol{x}}_\alpha\}|\boldsymbol{u},\mathcal{M}_i)p(\boldsymbol{u}|\mathcal{M}_i)d\bar{\boldsymbol{x}}_1\cdots d\bar{\boldsymbol{x}}_N d\boldsymbol{u}\,p(\mathcal{M}_i)\\ &\Big/\sum_{i=1}^{M}p(\{\boldsymbol{x}_\alpha\},\mathcal{M}_i), \end{aligned} \quad (17)$$

where $p(\{\boldsymbol{x}_\alpha\},\mathcal{M}_i)$ in the denominator is the expression in the numerator. Following Schwarz [19], we assume that each model has the same a priori probability and choose the model that maximizes Eq. (17). Since the denominator in Eq. (17) does not depend on individual models, we choose the model that maximizes

$$L=\int e^{-J}p(\{\bar{\boldsymbol{x}}_\alpha\}|\boldsymbol{u})p(\boldsymbol{u})d\bar{\boldsymbol{x}}^N d\boldsymbol{u}, \quad (18)$$

where and hereafter we denote $\iint\cdots\int d\bar{\boldsymbol{x}}_1\cdots d\bar{\boldsymbol{x}}_N d\boldsymbol{u}$ by $\int d\bar{\boldsymbol{x}}^N d\boldsymbol{u}$ and omit $\mathcal{M}_i$ to avoid notational clutter.

5.2 Expansion Around ML Estimators

In order to simplify the notation, we introduce the following metric associated with the Mahalanobis distance:

$$(\boldsymbol{a},\boldsymbol{b})_\alpha\equiv(\boldsymbol{a},V_0[\boldsymbol{x}_\alpha]^{-1}\boldsymbol{b}),\quad \|\boldsymbol{a}\|_\alpha\equiv\sqrt{(\boldsymbol{a},\boldsymbol{a})_\alpha}. \quad (19)$$

Then, we see from Eq. (10) that

$$J=\sum_{\alpha=1}^{N}\|\boldsymbol{x}_\alpha-\bar{\boldsymbol{x}}_\alpha\|_\alpha^2$$

[†] Note that $V_0[\tilde{\boldsymbol{x}}_\alpha]$ is *defined* by the right-hand side of Eq. (13); *not* that $\tilde{\boldsymbol{x}}_\alpha$ is "substituted" into $V_0[\boldsymbol{x}_\alpha]$. Since $V_0[\boldsymbol{x}_\alpha]$ is a symbol, not a function, nothing can be substituted. However, $V_0[\tilde{\boldsymbol{x}}_\alpha]$ *is* a "function" of $\boldsymbol{x}_\alpha$, $\boldsymbol{u}$, and $V_0[\boldsymbol{x}_\alpha]$.

[††] Strictly, we need a subscript i for the parameter $\boldsymbol{u}$ and the functions $p(\,\cdot\,)$ and $p(\,\cdot\,|\,\cdots)$, because they are different from model to model. To avoid notational complications, however, we omit such a subscript i.

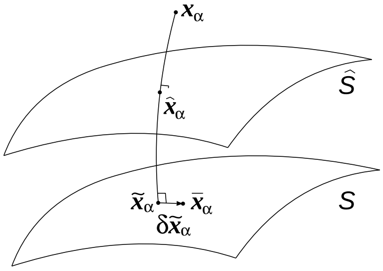
$\boldsymbol{x}_\alpha$
$\hat{\boldsymbol{x}}_\alpha$
$\hat{S}$
$\tilde{\boldsymbol{x}}_\alpha$
$\bar{\boldsymbol{x}}_\alpha$
$\delta\tilde{\boldsymbol{x}}_\alpha$
S

Fig. 2 Measured in the Mahalanobis distance, $\tilde{\boldsymbol{x}}_\alpha$ and $\hat{\boldsymbol{x}}_\alpha$ are the closest points in the true manifold $\mathcal{S}$ and the fitted manifold $\hat{\mathcal{S}}$, respectively, from the data point $\boldsymbol{x}_\alpha$. The true position $\bar{\boldsymbol{x}}_\alpha$ is in $\mathcal{S}$.

$$
\begin{aligned}
&=\sum_{\alpha=1}^{N}\|(\boldsymbol{x}_\alpha-\tilde{\boldsymbol{x}}_\alpha)+(\tilde{\boldsymbol{x}}_\alpha-\bar{\boldsymbol{x}}_\alpha)\|_\alpha^2\\
&=\sum_{\alpha=1}^{N}\|\boldsymbol{x}_\alpha-\tilde{\boldsymbol{x}}_\alpha\|_\alpha^2+\sum_{\alpha=1}^{N}\|\tilde{\boldsymbol{x}}_\alpha-\bar{\boldsymbol{x}}_\alpha\|_\alpha^2+\cdots, \qquad (20)
\end{aligned}
$$

where $\tilde{\boldsymbol{x}}_\alpha$ is a shorthand of $\tilde{\boldsymbol{x}}_\alpha(\boldsymbol{u})$, and "$\cdots$" denotes omitted higher-order terms in ε. The reasoning behind Eq. (20) is as follows. Since $\tilde{\boldsymbol{x}}_\alpha$ is, by definition, the point in $\mathcal{S}$ "closest" to $\boldsymbol{x}_\alpha$ measured in the norm $\|\cdot\|_\alpha$, the displacement $\boldsymbol{x}_\alpha-\tilde{\boldsymbol{x}}_\alpha$ is "orthogonal" to $\mathcal{S}$ in the metric $(\,\cdot\,,\cdot\,)_\alpha$. The ML estimator $\tilde{\boldsymbol{x}}_\alpha\in\mathcal{S}$ is in the $O(\varepsilon)$ neighborhood of its true position $\bar{\boldsymbol{x}}_\alpha\in\mathcal{S}$, so the deviation $\tilde{\boldsymbol{x}}_\alpha-\bar{\boldsymbol{x}}_\alpha$ is in higher-order contact with $\mathcal{S}$ at $\tilde{\boldsymbol{x}}_\alpha$ (Fig. 2). If the manifold $\mathcal{S}$ is flat, the terms "$\cdots$" vanish[†], and Eq. (20) is regarded as the "Pythagorean theorem" of a right triangle with respect to the metric in Eq. (19).

Consider the term $\sum_{\alpha=1}^{N}\|\boldsymbol{x}_\alpha-\tilde{\boldsymbol{x}}_\alpha\|_\alpha^2$ $(=\sum_{\alpha=1}^{N}\|\boldsymbol{x}_\alpha-\tilde{\boldsymbol{x}}_\alpha(\boldsymbol{u})\|_\alpha^2)$ in Eq. (20). Letting $\boldsymbol{u}=\hat{\boldsymbol{u}}+\delta\boldsymbol{u}$, we expand it in $\delta\boldsymbol{u}$ around $\hat{\boldsymbol{u}}$. Since by definition $\hat{\boldsymbol{u}}$ minimizes this term, the first order term in $\delta\boldsymbol{u}$ vanishes. From Eq. (15), we obtain

$$
\sum_{\alpha=1}^{N}\|\boldsymbol{x}_\alpha-\tilde{\boldsymbol{x}}_\alpha\|_\alpha^2=\hat{J}+(\delta\boldsymbol{u},\hat{\boldsymbol{M}}\delta\boldsymbol{u})+\cdots, \qquad (21)
$$

where "$\cdots$" denotes omitted higher-order terms in $\delta\boldsymbol{u}$. This is obtained by noting that Eq. (15) implies that the a posteriori probability density of $\boldsymbol{u}$ should be proportional to $e^{-(\delta\boldsymbol{u},V[\hat{\boldsymbol{u}}]^{-1}\delta\boldsymbol{u})}$ $(=e^{-(\delta\boldsymbol{u},\hat{\boldsymbol{M}}\delta\boldsymbol{u})/2\varepsilon^2})$ except for higher-order terms in $\delta\boldsymbol{u}$ (see [4] for the details). Since the posterior probability of the parameter $\boldsymbol{u}$ should be proportional to the likelihood in Eq. (9), they should have the same logarithmic expansion form.

Similarly, the term $\|\tilde{\boldsymbol{x}}_\alpha-\bar{\boldsymbol{x}}_\alpha\|_\alpha^2$ in Eq. (20) is written from Eq. (13) as

$$
\|\tilde{\boldsymbol{x}}_\alpha-\bar{\boldsymbol{x}}_\alpha\|_\alpha^2=(\delta\boldsymbol{x}_\alpha,V_0[\tilde{\boldsymbol{x}}_\alpha]^-\delta\boldsymbol{x}_\alpha)+\cdots, \qquad (22)
$$

where we put $\delta\boldsymbol{x}_\alpha=\tilde{\boldsymbol{x}}_\alpha-\bar{\boldsymbol{x}}_\alpha$, and $V_0[\tilde{\boldsymbol{x}}_\alpha]^-$ is the pseudoinverse[††]of $V_0[\tilde{\boldsymbol{x}}_\alpha]$.

Thus, Eq. (20) has the following expansion:

$$
J=\hat{J}+(\delta\boldsymbol{u},\hat{\boldsymbol{M}}\delta\boldsymbol{u})+\sum_{\alpha=1}^{N}(\delta\tilde{\boldsymbol{x}}_\alpha,V_0[\tilde{\boldsymbol{x}}_\alpha]^-\delta\tilde{\boldsymbol{x}}_\alpha)+\ldots.(23)
$$

The reasoning invoked here is essentially the same as that used for deriving the geometric AIC and the geometric MDL [4], [8], [11].

5.3 Expansion of A Posteriori Probability

Substituting Eq. (23) and omitting higher-order terms in ε, we can write Eq. (18) as follows:

$$
\begin{aligned}
L&=e^{-\hat{J}/2\varepsilon^2}\!\int\! e^{-(\delta\boldsymbol{u},\hat{\boldsymbol{M}}\delta\boldsymbol{u})/2\varepsilon^2}\Big(\int\! e^{-\sum_{\alpha=1}^{N}(\delta\tilde{\boldsymbol{x}}_\alpha,V_0[\tilde{\boldsymbol{x}}_\alpha]^-\delta\tilde{\boldsymbol{x}}_\alpha)/2\varepsilon^2}\\
&\quad\times p(\{\bar{\boldsymbol{x}}_\alpha\}|\boldsymbol{u})d\bar{\boldsymbol{x}}^N\Big)p(\boldsymbol{u})d\boldsymbol{u}\\
&=e^{-\hat{J}/2\varepsilon^2}\int e^{-(\boldsymbol{u}-\hat{\boldsymbol{u}},\hat{\boldsymbol{M}}(\boldsymbol{u}-\hat{\boldsymbol{u}}))/2\varepsilon^2}\\
&\quad\times\prod_{\alpha=1}^{N}\Big(\int e^{-(\bar{\boldsymbol{x}}_\alpha-\tilde{\boldsymbol{x}}_\alpha,V_0[\tilde{\boldsymbol{x}}_\alpha]^-(\bar{\boldsymbol{x}}_\alpha-\tilde{\boldsymbol{x}}_\alpha))/2\varepsilon^2}\\
&\quad\times p(\{\bar{\boldsymbol{x}}_\alpha\}|\boldsymbol{u})d\bar{\boldsymbol{x}}_\alpha\Big)p(\boldsymbol{u})d\boldsymbol{u}. \qquad (24)
\end{aligned}
$$

The expression $e^{-(\bar{\boldsymbol{x}}_\alpha-\tilde{\boldsymbol{x}}_\alpha,V_0[\tilde{\boldsymbol{x}}_\alpha]^-(\bar{\boldsymbol{x}}_\alpha-\tilde{\boldsymbol{x}}_\alpha))/2\varepsilon^2}$ in $\bar{\boldsymbol{x}}_\alpha$ takes a value close to 1 only when $\bar{\boldsymbol{x}}_\alpha$ is in the $O(\varepsilon)$ neighborhood of $\tilde{\boldsymbol{x}}_\alpha$, exponentially decaying to 0 around it. The a priori probability $p(\{\bar{\boldsymbol{x}}_\alpha\}|\boldsymbol{u})$ represents the state of our knowledge about the true position of $\boldsymbol{x}_\alpha$ given $\boldsymbol{u}$ [3], e.g., that the feature point we seek may be detected around a certain region in the image if the scene has a certain structure specified by $\boldsymbol{u}$. Hence, we may assume that $p(\{\bar{\boldsymbol{x}}_\alpha\}|\boldsymbol{u})$ varies smoothly around $\tilde{\boldsymbol{x}}_\alpha$ unless a specific evidence for otherwise exists. Thus, if $p(\{\bar{\boldsymbol{x}}_\alpha\}|\boldsymbol{u})$ is expanded around $\tilde{\boldsymbol{x}}_\alpha$ into $p(\{\tilde{\boldsymbol{x}}_\alpha\}|\boldsymbol{u})+(\nabla_{\boldsymbol{x}}p(\{\tilde{\boldsymbol{x}}_\alpha\}|\boldsymbol{u}),\bar{\boldsymbol{x}}_\alpha-\tilde{\boldsymbol{x}}_\alpha)+\cdots$, the remaining terms "$\cdots$", which are second (quadratic) or higher orders in ε, can be ignored. However, the first order (linear) term is an odd function around $\tilde{\boldsymbol{x}}_\alpha$, so integration of it after multiplication by $e^{-(\bar{\boldsymbol{x}}_\alpha-\tilde{\boldsymbol{x}}_\alpha,V_0[\tilde{\boldsymbol{x}}_\alpha]^-(\bar{\boldsymbol{x}}_\alpha-\tilde{\boldsymbol{x}}_\alpha))/2\varepsilon^2}$ vanishes. The integration of the remaining zeroth order (constant) term $p(\{\tilde{\boldsymbol{x}}_\alpha\}|\boldsymbol{u})$ is evaluated from the normalization relation of the Gaussian distribution in the form

$$
\begin{aligned}
&\int e^{-(\bar{\boldsymbol{x}}_\alpha-\tilde{\boldsymbol{x}}_\alpha,V_0[\tilde{\boldsymbol{x}}_\alpha]^-(\bar{\boldsymbol{x}}_\alpha-\tilde{\boldsymbol{x}}_\alpha))/2\varepsilon^2}p(\{\tilde{\boldsymbol{x}}_\alpha\}|\boldsymbol{u})d\bar{\boldsymbol{x}}_\alpha\\
&\qquad=\sqrt{(2\pi)^d\varepsilon^{2d}|V_0[\tilde{\boldsymbol{x}}_\alpha]|_+}\;p(\{\tilde{\boldsymbol{x}}_\alpha\}|\boldsymbol{u}), \qquad (25)
\end{aligned}
$$

where $|V_0[\tilde{\boldsymbol{x}}_\alpha]|_+$ denotes the product of positive eigenvalues of $V_0[\tilde{\boldsymbol{x}}_\alpha]$, i.e., its determinant restricted to its domain $T_{\tilde{\boldsymbol{x}}_\alpha}(\mathcal{S})$.

Next, we compute the product $\prod_{\alpha=1}^{N}$ of Eq. (25), multiply it by $e^{-(\delta\boldsymbol{u},\hat{\boldsymbol{M}}\delta\boldsymbol{u})/2\varepsilon^2}p(\boldsymbol{u})$, and integrate the resulting expression with respect to $\boldsymbol{u}$. Again, the expression $e^{-(\boldsymbol{u}-\hat{\boldsymbol{u}},\hat{\boldsymbol{M}}(\boldsymbol{u}-\hat{\boldsymbol{u}}))/2\varepsilon^2}$ in $\boldsymbol{u}$ has a value close to 1 only when $\boldsymbol{u}$ is in the $O(\varepsilon)$ neighborhood of $\hat{\boldsymbol{u}}$, expo-

[†] Thus, the terms "$\cdots$" in Eq. (20) depend not only on ε but also the "curvature" of the manifold $\mathcal{S}$. Rigorous order analysis would require more precise analysis, but what we need later is only the leading terms.

[††] The matrix $V_0[\tilde{\boldsymbol{x}}_\alpha]$ in Eq. (13) is singular and has rank d. Its domain is the d-dimensional tangent space $T_{\tilde{\boldsymbol{x}}_\alpha}(\mathcal{S})$ to $\mathcal{S}$, whose orthogonal complement is the null space of $V_0[\tilde{\boldsymbol{x}}_\alpha]$; no deviations are allowed in it. The pseudoinverse $V_0[\tilde{\boldsymbol{x}}_\alpha]^-$ means the inverse operation within $T_{\tilde{\boldsymbol{x}}_\alpha}(\mathcal{S})$, preserving the same null space.

nentially decaying to 0 around it. As before, $|V_0[\tilde{\boldsymbol{x}}_\alpha]|_+$, $p(\{\tilde{\boldsymbol{x}}_\alpha\}|\boldsymbol{u})$, and $p(\boldsymbol{u})$ can be regarded as smooth functions of $\boldsymbol{u}$ around $\hat{\boldsymbol{u}}$, so their second (quadratic) and higher-order expansion terms can be ignored. The first order (linear) terms are odd functions of $\boldsymbol{u}$ around $\hat{\boldsymbol{u}}$, so their integration after multiplication by $e^{-(\delta\boldsymbol{u},\hat{\boldsymbol{M}}\delta\boldsymbol{u})/2\varepsilon^2}$ vanishes. Hence, we only need to integrate the zeroth order (constant) terms of $|V_0[\hat{\boldsymbol{x}}_\alpha]|_+$, $p(\{\hat{\boldsymbol{x}}_\alpha\}|\boldsymbol{u})$, and $p(\boldsymbol{u})$ multiplied by $e^{-(\delta\boldsymbol{u},\hat{\boldsymbol{M}}\delta\boldsymbol{u})/2\varepsilon^2}$, where $|V_0[\hat{\boldsymbol{x}}_\alpha]|_+$ is a shorthand of $|V_0[\tilde{\boldsymbol{x}}_\alpha(\hat{\boldsymbol{u}})]|_+$. Using the normalization relation of the Gaussian distribution, we obtain

$$\begin{aligned}&\int e^{-(\delta\boldsymbol{u},\hat{\boldsymbol{M}}\delta\boldsymbol{u})/2\varepsilon^2}\prod_{\alpha=1}^{N}\sqrt{(2\pi)^d\varepsilon^{2d}|V_0[\hat{\boldsymbol{x}}_\alpha]|_+}\\&\quad\times p(\{\hat{\boldsymbol{x}}_\alpha\}|\boldsymbol{u})p(\boldsymbol{u})d\boldsymbol{u}\\&=\sqrt{(2\pi)^p\varepsilon^{2p}|\hat{\boldsymbol{M}}|_+^{-p}}\prod_{\alpha=1}^{N}\sqrt{(2\pi)^d\varepsilon^{2d}|V_0[\hat{\boldsymbol{x}}_\alpha]|_+}\\&\quad\times p(\{\hat{\boldsymbol{x}}_\alpha\}|\hat{\boldsymbol{u}})p(\hat{\boldsymbol{u}})+\cdots.\end{aligned}\tag{26}$$

The reasoning invoked here is essentially the same as that of Schwarz [19] used for deriving his BIC.

5.4 Geometric BIC

Thus, Eq. (24) is evaluated except for higher-order terms in ε in the following form:

$$\begin{aligned}L=&e^{-\hat{J}/2\varepsilon^2}\sqrt{(2\pi)^p\varepsilon^{2p}|\hat{\boldsymbol{M}}|_+^{-p}}\ p(\hat{\boldsymbol{u}})\\&\times\prod_{\alpha=1}^{N}\sqrt{(2\pi)^d\varepsilon^{2d}|V_0[\hat{\boldsymbol{x}}_\alpha]|_+}\ p(\{\hat{\boldsymbol{x}}_\alpha\}|\hat{\boldsymbol{u}}).\end{aligned}\tag{27}$$

Its logarithm takes the form

$$\begin{aligned}\log L=&-\frac{\hat{J}}{2\varepsilon^2}+\frac{p}{2}\log 2\pi+\frac{p}{2}\log\varepsilon^2-\frac{p}{2}\log|\hat{\boldsymbol{M}}|_+\\&+\log p(\hat{\boldsymbol{u}})+\frac{Nd}{2}\log 2\pi+\frac{Nd}{2}\log\varepsilon^2\\&+\frac{1}{2}\sum_{\alpha=1}^{N}\log|V_0[\hat{\boldsymbol{x}}_\alpha]|_++\sum_{\alpha=1}^{N}\log p(\{\hat{\boldsymbol{x}}_\alpha\}|\hat{\boldsymbol{u}}).\end{aligned}\tag{28}$$

The model that has the largest value of $\log L$ can be regarded as the most appropriate. Multiplying Eq. (28) by $-2\varepsilon^2$, we obtain

$$\begin{aligned}&\hat{J}-(Nd+p)\varepsilon^2\log\varepsilon^2+\varepsilon^2\Big(p\log|\hat{\boldsymbol{M}}|_+\\&-\sum_{\alpha=1}^{N}\log|V_0[\hat{\boldsymbol{x}}_\alpha]|_+-(Nd+p)\log 2\pi-2\log p(\hat{\boldsymbol{u}})\\&-2\sum_{\alpha=1}^{N}\log p(\{\hat{\boldsymbol{x}}_\alpha\}|\hat{\boldsymbol{u}})\Big).\end{aligned}\tag{29}$$

The terms of $O(\varepsilon^2)$ approach 0 more quickly than the terms of $O(\varepsilon^2\log\varepsilon^2)$ as $\varepsilon\to 0$. Omitting the last term, we obtain the *geometric BIC*

$$\text{G-BIC}=\hat{J}-(Nd+p)\varepsilon^2\log\varepsilon^2,\tag{30}$$

which has the same form as the geometric MDL in Eq. (6).

This situation corresponds to the fact that Rissanen's MDL has the same form as Schwarz' BIC as far as the leading terms are concerned, in spite of the fact that they are derived from quite different reasonings: one from information theory, the other from the Bayes theorem. Thus, our result is a reassuring evidence that the logic and the derivation we used are correct. Today, however, the view that the Bayesian principle is more fundamental than the MDL principle is becoming more and more dominant [3]. From this viewpoint, our result can also justify the geometric MDL, whose axiomatic origin has some arbitrariness, from the Bayesian standpoint.

5.5 Dimensional anomaly

Both Eqs. (6) and (30) involve logarithm of ε, which has the dimension of length (in pixels). This anomaly is caused by our crude order comparison: the scale factor to divide ε for canceling the dimensionality is separated away due to the additivity of the logarithm and discarded[†], because it increases less rapidly than $O(\log\varepsilon)$ as $\varepsilon\to 0$.

This anomaly could be compensated for if higher-order terms in ε were included, but that would cause much complication. A realistic compromise is, as suggested in [11], to introduce a typical reference length L, such as the image size, and replace $\log\varepsilon^2$ by $\log(\varepsilon/L)^2$ ($=\log\varepsilon^2-2\log L$); the term $-2\log L$ has little effect on model selection when ε is sufficiently small [11].

6. Applications

The geometric AIC and the geometric MDL have been used for model selection of various problems for computer vision, including fitting lines, curves, planes, and surfaces to 2-D and 3-D points [11], reliability evaluation of 3-D computation using a moving camera [6], detecting symmetry of 2-D shapes [5], segmenting a curve into line segments [7], inferring object shapes by stereo vision [13], moving object detection from optical flow [17], camera motion estimation for virtual studio systems [16], correspondence detection between images [15], automatic regularity enforcement on 2-D figures [21], automatic image mosaicing [14], and multibody motion segmentation [10], [20].

Almost all these applications are for *degeneracy detection*. For particular parameter values, the model degenerates and has a lower degree of freedom, or the manifold it defines has a lower dimension. For example, curves and surfaces degenerate into lines and planes if some of the coefficients vanish. Depending

[†] The same problem also arises to Rissanen's MDL: if multiple data are combined, e.g., a consecutive pair, into one, the apparent number N of the data decreases, so the MDL changes its value. This effect is compensated for by higher-order terms in $1/\sqrt{N}$ involving the Fisher information matrix $\boldsymbol{I}(\theta)$. See the footnote † in page 145.

on the parameter values, rigid motions may degenerate into pure rotations, and projective transformations into affine transformations. If such degeneracies occur, the computation based on a nondegenerate model may fail. For example, 3-D reconstruction fails if the assumed rigid camera motion degenerates into a pure rotation. In such a case, one needs to switch to the degenerate model. To this, geometric model selection is called for, because the nondegenerate model always has a smaller residual than the degenerate one and hence they cannot be compared by the residual alone. For such applications, the following has been known [11].

- The geometric AIC tends to select a model that is faithful to the data. It almost always judges a nondegenerate model to be nondegenerate but sometimes judges a degenerate model to be nondegenerate.
- The geometric MDL prefers the simplicity of the model to the faithfulness to the data. It almost always judges a degenerate model to be degenerate but very often judges a nondegenerate model to be degenerate.

Hence, the choice between the geometric AIC and the geometric MDL should be based on which the user gives preference, detecting degeneracy or nondegeneracy.

Since the geometric BIC introduced in this paper has the same form as the geometric MDL, no essentially new results are obtained by it. However, we obtain a new "interpretation" that the use of the geometric MDL can also be justified from the Bayesian principle.

7. Conclusions

Noting that Akaike's AIC and Rissanen's MDL in the traditional statistical framework correspond to the geometric AIC and the geometric MDL, respectively, in the geometric fitting framework, we have answered the question as to what Schwarz' BIC in the statistical framework corresponds to in the geometric framework.

We first described the difference between the traditional statistical estimation framework and the geometric fitting framework, pointing out that the asymptotic analysis as the number N of data goes to ∞ in the former corresponds to the perturbation analysis as the noise level ε goes to 0 in the latter. Then, we introduced the Bayesian logic for geometric model selection and derived the geometric BIC. We found that it has the same form as the geometric MDL. Our result justifies the geometric MDL from the Bayesian principle. We also discussed applications of geometric model selection.

Acknowledgments

This work was supported in part by the Ministry of Education, Culture, Sports, Science and Technology, Japan, under a Grant in Aid for Scientific Research C (No. 21500172).

References

[1] H. Akaike, "A new look at the statistical model identification," IEEE Trans. Autom. Control, vol.26, no.6, pp.716–723, Dec. 1974.

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

[3] E.T. Jaynes, Probability Theory: The Logic of Science, Cambridge University Press, Cambridge, U.K., 2003.

[4] K. Kanatani, Statistical Optimization for Geometric Computation: Theory and Practice, Elsevier Science, Amsterdam, the Netherlands, 1996; Reprinted, Dover, New York, NY, U.S.A., 2005.

[5] K. Kanatani, "Comments on 'Symmetry as a Continuous Feature'," IEEE Trans. Pattern Anal. Mach. Intell., vol.19, no.3, pp.246–247, March 1997.

[6] K. Kanatani, "Self-evaluation for active vision by the geometric information criterion," Proc. 7th Int. Conf. Comput. Anal. Images Patterns, pp.247–254, Kiel, Germany, Sept. 1997.

[7] K. Kanatani, "Comments on 'Nonparametric Segmentation of Curves into Various Representations'," IEEE Trans. Pattern Anal. Mach. Intel., vol.19, no.12, pp.1391–1392, Dec. 1997.

[8] K. Kanatani, "Geometric information criterion for model selection," Int. J. Comput. Vis., vol.26, no.3, pp.171–189, Feb./March 1998.

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

[10] K. Kanatani, "Motion segmentation by subspace separation: Model selection and reliability evaluation," Int. J. Image Graphics, vol.2, no.2, pp.179–197, April 2002.

[11] K. Kanatani, "Uncertainty modeling and model selection for geometric inference," IEEE Trans. Pattern Anal. Mach. Intell., vol.26, no.10, pp.1307–1319, Oct. 2004.

[12] K. Kanatani, "Statistical optimization for geometric fitting: Theoretical accuracy analysis and high order error analysis," Int. J. Comput. Vis., vol.80, no.2, pp.167–188, Nov. 2008.

[13] Y. Kanazawa and K. Kanatani, "Infinity and planarity test for stereo vision," IEICE Trans. Inf. & Syst., vol.E80-D, no.8, pp.774–779, Aug. 1997.

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

[15] Y. Kanazawa and K. Kanatani "Robust image matching preserving global consistency," Proc. 6th Asian Conf. Comput. Vis., vol.2, pp.1128–1133, Jeju, Korea, Jan. 2004.

[16] C. Matsunaga and K. Kanatani, "Calibration of a moving camera using a planar pattern: Optimal computation, reliability evaluation and stabilization by model selection," Proc. 7th Euro. Conf. Computer Vision, vol.2, pp.595–609, Dublin, Ireland, July 2000.

[17] 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, Feb. 1998.

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

[19] G. Schwarz, "Estimating the dimension of a model," Annals Statis.,, vol.6, no.2, pp.461–464, March 1978.

[20] Y. Sugaya and K. Kanatani, "Multi-stage optimization for multi-body motion segmentation," IEICE Trans. Inf. & Syst., vol.E87-D, no.7, pp.1935–1942, July 2004.

[21] I. 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, Feb. 1998.

Kenichi Kanatani received his B.E., M.S., and Ph.D. in applied mathematics from the University of Tokyo in 1972, 1974 and 1979, respectively. After serving as Professor of computer science at Gunma University, Gunma, Japan, he is currently Professor of computer science at Okayama University, Okayama, Japan. He is the author of many books on computer vision and received many awards including the best paper awards from IPSJ (1987) and IEICE (2005). He is an IEEE Fellow.