Engineering

Industrial & Management Engineering fields

Okayama University *Year* 2004

Uncertainty modeling and model selection for geometric inference

Kenichi Kanatani
Okayama University

Uncertainty Modeling and Model Selection for Geometric Inference

Kenichi Kanatani, *Fellow*, *IEEE*

Abstract—We first investigate the meaning of "statistical methods" for geometric inference based on image feature points. Tracing back the origin of feature uncertainty to image processing operations, we discuss the implications of asymptotic analysis in reference to "geometric fitting" and "geometric model selection" and point out that a correspondence exists between the standard statistical analysis and the geometric inference problem. Then, we derive the "geometric AIC" and the "geometric MDL" as counterparts of Akaike's AIC and Rissanen's MDL. We show by experiments that the two criteria have contrasting characteristics in detecting degeneracy.

Index Terms—Statistical method, feature point extraction, asymptotic evaluation, geometric AIC, geometric MDL.

1 Introduction

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 [10]. In the same framework, the *geometric AIC* and the *geometric MDL* have been proposed for model selection [11], [13] and applied to many problems of computer vision [12], [14], [15], [18], [19], [22], [26], [39].

However, the crucial difference of these criteria from other similar criteria [34], [35], [37], [38] have been overlooked, often causing misunderstanding. The difference stems from the interpretation of "statistical methods." The purpose of this paper is twofold:

1. We examine the origin of "uncertainty" in geometric inference and point out that it has a different meaning from that in the traditional statistical problems.
2. We derive the geometric AIC and the geometric MDL in this new light and show by experiments that they have very different characteristics.

In Sections 2 and 3, we focus on the question of why we need a statistical method for computer vision, tracing back the origin of feature uncertainty to image processing operations. In Sections 4, we discuss the implications of asymptotic analysis in reference to "geometric fitting" and "geometric model selection" and point out that a correspondence exists between the standard statistical analysis and geometric inference. In Sections 5, 6, and 7, we derive the "geometric AIC" and the "geometric MDL" as counterparts of Akaike's AIC and Rissanen's MDL. In Section 8, we address related issues. In Section 9, we show by experiments that the two criteria have contrasting characteristics in detecting degeneracy. Section 10 presents our concluding remarks.

2 What Is Geometric Inference?

2.1 Ensembles for Geometric Inference

The goal of statistical methods is not to study the properties of observed data themselves but to infer the properties of the *ensemble* from which we regard the observed data as sampled. The ensemble may be a collection of existing entities (e.g., the entire population), but often it is a hypothetical set of conceivable possibilities. When a statistical method is employed, the underlying ensemble is often taken for granted. However, this issue is very crucial for geometric inference based on feature points.

Suppose, for example, we extract feature points, such as corners of walls and windows, from an image of a building and want to test if they are collinear. The reason why we need a statistical method is that the extracted feature positions have uncertainty. So, we have to judge the extracted feature points as collinear if they are sufficiently aligned. We can also evaluate the degree of uncertainty of the fitted line by propagating the uncertainty of the individual points. What is the ensemble that underlies this type of inference? This question reduces to the question of why the uncertainty of the feature points occurs at all. After all, statistical methods are not necessary if the data are exact. Using a statistical method means regarding the current feature position as sampled from a set of its possible positions. But, where else could it be if not in the current position?

2.2 Uncertainty of Feature Extraction

Many algorithms have been proposed for extracting feature points including the Harris operator [8] and SUSAN [32], and their performance has been extensively compared [3], [27], [31]. However, if we use, for example, the Harris operator to extract a particular corner of a particular building image, the output is unique (Fig. 1). No matter how many times we repeat the extraction, we obtain the same point because no external disturbances exist and the internal parameters (e.g., thresholds for judgment) are unchanged. It follows that the current position is the sole possibility. How can we find it elsewhere?

• The author is with the Department of Information Technology, Okayama University, Okayama 700-8530 Japan.
E-mail: kanatani@suri.it.okayama-u.ac.jp.

Manuscript received 1 June 2003; revised 4 Nov. 2003; accepted 4 Jan. 2004.
Recommended for acceptance by R. Chellappa.
For information on obtaining reprints of this article, please send e-mail to: tpami@computer.org, and reference IEEECS Log Number TPAMI-0116-0603.

0162-8828/04/$20.00 © 2004 IEEE Published by the IEEE Computer Society

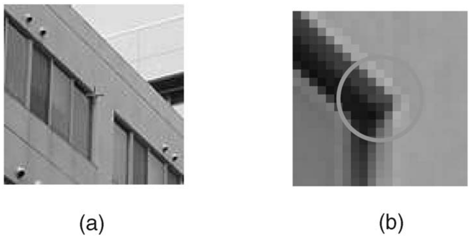


Fig. 1. (a) A feature point in an image of a building. (b) Its enlargement and the uncertainty of the feature location.

If we closely examine the situation, we are compelled to conclude that other possibilities should exist because the extracted position is not necessarily correct. But, if it is not correct, why didn't we extract the correct position in the first place? The answer is: *we cannot*.

2.3 Image Processing for Computer Vision

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

However, what we want is not "image properties" but "3D properties" such as corners of a building, but the way a 3D property is translated into an image property is intrinsically heuristic. As a result, as many algorithms can exist as the number of heuristics for its 2D interpretation.

If we specify a particular 3D feature to extract, say a corner of a window, its appearance in the image is not unique. It is affected by many properties of the scene including the details of its 3D shape, the viewing orientation, the illumination condition, and the light reflectance properties of the material. A slight variation of any of them can result in a substantial difference in the image.

Theoretically, exact extraction would be possible if all the properties of the scene were exactly known, but to infer them from images is the very task of computer vision. It follows that we must make a guess in the image processing stage. For the current image, some guesses may be correct, but others may be wrong. The exact feature position could be found only by an (nonexisting) "ideal" algorithm that could guess everything correctly.

This observation allows us to interpret the "possible feature positions" to be *the positions that would be located by different (nonideal) algorithms based on different guesses*. It follows that the set of hypothetical positions should be associated with *the set of hypothetical algorithms*. The current position is regarded as produced by an algorithm sampled from it. This explains why one always obtains the same position no matter how many times one repeats extraction using that algorithm. To obtain a different position, one has to sample another algorithm.

Remark 1. We may view the statistical ensemble in the following way. If we repeat the *same* experiment, the result should always be the same. But, if we declare that the experiment is the "same" if such and such are the same while other things can vary, those variable conditions define the ensemble. The conventional view is to regard the experiment as the same if the *3D scene* we are viewing is the same while other properties, such as the lighting condition, can vary. The resulting image would be different for each (hypothetical) experiment, so one would obtain a different output using the same image processing algorithm. The expected spread of the outputs measures the robustness of that algorithm.

In this paper, we view the experiment as the same *if the image is the same*. Then, we could obtain different results only by sampling other algorithms. The expected spread of the outputs measures the uncertainty of feature detection from *that image*. We take this view because we want to analyze the reliability of geometric inference from a particular image, while the conventional view is suitable for assessing the robustness of a *particular algorithm*.

3 Statistical Model of Feature Location

3.1 Covariance Matrix of a Feature Point

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

Let $V[p_\alpha]$ be the covariance matrix of the αth feature point p_α. The above argument implies that we can estimate the qualitative characteristics of uncertainty but not its absolute magnitude. So, we write the covariance matrix $V[p_\alpha]$ in the form

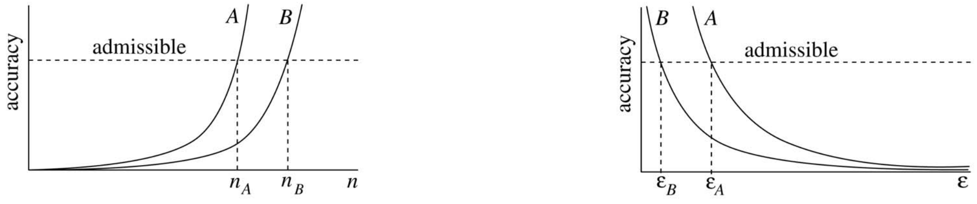
$$V[p_\alpha] = \varepsilon^2 V_0[p_\alpha], \tag{1}$$

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

Remark 2. The decomposition of $V[p_\alpha]$ into ε^2 and $V_0[p_\alpha]$ involves scale ambiguity. In practice, this scale is implicitly determined by the image process operation for estimating the feature uncertainty applied to all the feature points in the same manner (see [20] for the details). The subsequent analysis does not depend on particular normalizations as long as they are done in such a way that ε is much smaller than the data themselves.

3.2 Covariance Matrix Estimation

If the intensity variations around p_α are almost the same in all directions, we can think of the probability distribution as



(a)

(b)

Fig. 2. (a) For the standard statistical analysis, it is desired that the accuracy increases rapidly as the number of experiments $n \to \infty$, because admissible accuracy can be reached with a smaller number of experiments. (b) For geometric inference, it is desired that the accuracy increases rapidly as the noise level $\varepsilon \to 0$, because larger data uncertainty can be tolerated for admissible accuracy.

isotropic, a typical equiprobability line, known as the *uncertainty ellipses*, being a circle (Fig. 1b).

On the other hand, if p_α is on an object boundary, distinguishing it from nearby points should be difficult whatever algorithm is used, so its covariance matrix should have an elongated uncertainty ellipse along that boundary.

However, existing feature extraction algorithms are usually designed to output those points that have large image variations around them, so points in a region with an almost homogeneous intensity or on object boundaries are rarely chosen. As a result, the covariance matrix of a feature point extracted by such an algorithm can be regarded as nearly isotropic. This has also been confirmed by experiments [20], justifying the use of the identity as the normalized covariance matrix $V_0[p_\alpha]$.

Remark 3. The intensity variations around different feature points are usually unrelated, so their uncertainty can be regarded as statistically independent. However, if we track feature points over consecutive video frames, it has been observed that the uncertainty has strong correlations over the frames [33].

Remark 4. Many interactive applications require humans to extract feature points by manipulating a mouse. Extraction by a human is also an "algorithm," and it has been shown by experiments that humans are likely to choose "easy-to-see" points such as isolated points and intersections, avoiding points in a region with an almost homogeneous intensity or on object boundaries [20]. In this sense, the statistical characteristics of human extraction are very similar to machine extraction. This is no surprise if we recall that image processing for computer vision is essentially a heuristic that simulates human perception. It has also been reported that strong microscopic correlations exist when humans manually select corresponding feature points over multiple images [25].

3.3 Image Quality and Uncertainty

In the past, the uncertainty of feature points has often been identified with "image noise," giving a misleading impression as if the feature locations were perturbed by random intensity fluctuations. Of course, we may obtain better results using higher-quality images whatever algorithm is used. However, the task of computer vision is not to analyze "image properties" but to study the "3D properties" of the scene. As long as the image properties and the 3D properties do not correspond one to one, any image processing inevitably entails some degree of uncertainty, however high the image quality may be, and the result must be interpreted statistically. The underlying ensemble is the set of hypothetical (inherently imperfect) algorithms of image processing. Yet, it has been customary to evaluate the performance of image processing algorithms by adding *independent Gaussian noise* to individual pixels.

Remark 5. This also applies to *edge detection*, whose goal is to find the boundaries of 3D objects in the scene. In reality, all existing algorithms seek *edges*, i.e., lines and curves across which the intensity changes discontinuously. Yet, this is regarded by many as an objective image processing task, and the detection performance is often evaluated by adding independent Gaussian noise to individual pixels. From the above considerations, we conclude that edge detection is also a heuristic and, hence, no definitive algorithm will ever be found.

4 Asymptotic Analysis

4.1 What Is Asymptotic Analysis?

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

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

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

In contrast, the ensemble for geometric inference is, as we have seen, the set of potential feature positions that could be located if other (hypothetical) algorithms were used. As noted earlier, however, we can choose only *one* sample from the ensemble as long as we use a particular image processing algorithm. In other words, the number n of experiments is 1. Then, how can we evaluate the performance of statistical estimation?

Evidently, we want a method whose accuracy is sufficiently high *even for large data uncertainty*. This implies that we should analyze the growth in accuracy as the noise level ε decreases, because a method whose accuracy increases more rapidly as $\varepsilon \to 0$ can tolerate larger data uncertainty for admissible accuracy (Fig. 2b).

4.2 Geometric Fitting

We illustrate our assertion in more specific terms. Let $\{p_\alpha\}$, $\alpha = 1, \ldots, N$ be the extracted feature points. Suppose each point should satisfy a parameterized constraint

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

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

A typical problem of this form is to fit a line or a curve to given N points in the image, but this can be straightforwardly extended to multiple images. For example, if a point (x_α, y_α) in one image corresponds to a point (x'_α, y'_α) in another, we can regard them as a single point p_α in a 4-dimensional joint space with coordinates $(x_\alpha, y_\alpha, x'_\alpha, y'_\alpha)$. If the camera imaging geometry is modeled as perspective projection, the constraint (2) corresponds to the *epipolar equation*; the parameter $\boldsymbol{u}$ is the *fundamental matrix* [9].

The problem can be stated in abstract terms as *geometric fitting* as follows. We view a feature point in the image plane or a set of feature points in the joint space as an m-dimensional vector $\boldsymbol{x}$; we call it a "datum." Let $\{\boldsymbol{x}_\alpha\}$, $\alpha = 1, \ldots, N$, be the observed data. Their true values $\{\bar{\boldsymbol{x}}_\alpha\}$ are supposed to satisfy r constraint equations

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

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

Let

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

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

$$P(\{X_\alpha\}) = \prod_{\alpha=1}^{N} \frac{e^{-(\boldsymbol{X}_\alpha - \bar{\boldsymbol{x}}_\alpha, V_0[\boldsymbol{x}_\alpha]^{-1}(\boldsymbol{X}_\alpha - \bar{\boldsymbol{x}}_\alpha))/2\varepsilon^2}}{\sqrt{(2\pi\varepsilon^2)^m |V_0[\boldsymbol{x}_\alpha]|}}. \tag{5}$$

Throughout this paper, we use uppercases for random variables and lowercases for their instances; $|\cdot|$ denotes the determinant. The inner product of vectors $\boldsymbol{a}$ and $\boldsymbol{b}$ is denoted by $(\boldsymbol{a}, \boldsymbol{b})$.

Maximum likelihood (*ML*) *estimation* is to find the values of $\{\bar{\boldsymbol{x}}_\alpha\}$ and $\boldsymbol{u}$ that maximize the *likelihood*, i.e., (3) into which the data $\{\boldsymbol{x}_\alpha\}$ are substituted, or equivalently minimize the sum of the squared *Mahalanobis distances* in the form

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

subject to the constraint (3). The solution is called the *maximum likelihood* (*ML*) *estimator*. If the uncertainty is small, which we assume hereafter, the constraint (3) can be eliminated by introducing Lagrange multipliers and applying first order approximation. After some manipulations, we obtain the following form [10]:

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

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

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

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

It can be shown [10] that the covariance matrix of the ML estimator $\hat{\boldsymbol{u}}$ has the form

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

where

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

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

Remark 6. The data $\{\boldsymbol{x}_\alpha\}$ may be subject to some constraints. For example, each $\boldsymbol{x}_\alpha$ may be a unit vector. The above formulation still holds if the inverse $V_0[\boldsymbol{x}_\alpha]^{-1}$ in (6) is replaced by the (Moore-Penrose) generalized (or pseudo) inverse $V_0[\boldsymbol{x}_\alpha]^-$ and if the determinant $|V_0[\boldsymbol{x}_\alpha]|$ is replaced by the product of the positive eigenvalues of $V_0[\boldsymbol{x}_\alpha]$ [10].

Similarly, the r constraints in (3) may be redundant, say only r' ($< r$) of them are independent. The above formulation still holds if the inverse in (8) is replaced by the generalized inverse of rank r' with all but r' largest eigenvalues replaced by zero [10].

Remark 7. It can be proven that no other estimators could reduce the covariance matrix further than (9) except for the higher order term $O(\varepsilon^4)$ [10]. The ML estimator is optimal in this sense. Recall that we are focusing on the asymptotic analysis for $\varepsilon \to 0$. Thus, what we call the "ML estimator" should be understood to be a first approximation to the true ML estimator for small ε.

Remark 8. The p-dimensional parameter vector $\boldsymbol{u}$ may be constrained. For example, it may be a unit vector. If it has only p' ($< p$) degrees of freedom, the parameter space $\mathcal{U}$ is a p'-dimensional manifold in $\mathcal{R}^p$. In this case, the matrix $\boldsymbol{M}(\boldsymbol{u})$ in (9) is replaced by $\boldsymbol{P}_{\boldsymbol{u}} \boldsymbol{M}(\boldsymbol{u}) \boldsymbol{P}_{\boldsymbol{u}}$, where $\boldsymbol{P}_{\boldsymbol{u}}$

is the projection matrix onto the tangent space to $\mathcal{U}$ at $\boldsymbol{u}$ [10]. The inverse $M(\hat{\boldsymbol{u}})^{-1}$ in (9) is replaced by the generalized inverse $M(\hat{\boldsymbol{u}})^{-}$ of rank p' [10].

4.3 Dual Interpretations of Asymptotic Analysis

The above analysis bears a strong resemblance to the standard statistical estimation problem: After observing n data $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n$, we want to estimate the parameter $\boldsymbol{\theta}$ of the probability density $P(\boldsymbol{x}|\boldsymbol{\theta})$, according to which each datum is assumed to be sampled independently. It is known that the covariance matrix $V[\hat{\boldsymbol{\theta}}]$ of the ML estimator $\hat{\boldsymbol{\theta}}$ converges, under a mild condition, to $\boldsymbol{O}$ as the number n of experiments goes to infinity (*consistency*) and that it agrees with the *Cramer-Rao lower bound* expect for $O(1/n^2)$ (*asymptotic efficiency*). If follows that $1/\sqrt{n}$ plays the role of ε for geometric inference.

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

In fact, there were attempts to generate a hypothetical *ensemble of algorithms* by randomly varying the internal parameters (e.g., the thresholds for judgments), not adding random noise to the image [4], [5]. Then, one can compute their means and covariance matrix. Such a process as a whole can be regarded as one operation that effectively achieves higher accuracy.

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

5 Geometric Model Selection

Geometric fitting is to estimate the parameter $\boldsymbol{u}$ of a given model. If we have multiple candidate models

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

from which we are to select an appropriate one for the observed data $\{\boldsymbol{x}_\alpha\}$, the problem is (*geometric*) *model selection* [10].

Suppose, for example, we want to fit a curve to given points in two dimensions. If they are almost collinear, a straight line may fit fairly well, but a quadratic curve may fit better, and a cubic curve even better. Which curve should we fit?

A naive idea is to compare the *residual* (*sum of squares*), i.e., the minimum value $\hat{J}$ of J in (6); we select the one that has the smallest residual $\hat{J}$. This does not work, however, because the ML estimator $\hat{\boldsymbol{u}}$ is so determined as to minimize the residual $\hat{J}$, and the residual $\hat{J}$ can be made arbitrarily smaller if the model is equipped with more parameters to adjust. So, the only conclusion would be to fit a curve of a sufficiently high order passing through all the points.

This observation leads to the idea of compensating for the negative bias of the residual caused by substituting the ML estimator. This is the principle of Akaike's *AIC* (*Akaike information criterion*) [1], which is derived from the asymptotic analysis of the *Kullback-Leibler distance* (or *divergence*) as the number n of experiments goes to infinity.

Another well-known criterion is Rissanen's *MDL* (*Minimum description length*) [28], [29], [30], which measures the goodness of a model by the minimum information theoretic code length of the data and the model. Its form is evaluated asymptotically as the data length n grows to infinity.

In the next two sections, we follow the derivation of Akaike's AIC and Rissanen's MDL and examine the asymptotic behavior *as the noise level ε goes to zero*. We will show that this results in the *geometric AIC* and the *geometric MDL*, which were previously obtained by somewhat an ad hoc manner [10], [22].

6 Geometric AIC

6.1 Goodness of a Model

Akaike [1] adopted as the measure of the goodness of the model given by (5) the *Kullback-Leibler distance* (or *divergence*)

$$\begin{aligned} D &= \int \cdots \int P_T(\{\boldsymbol{X}_\alpha\}) \log \frac{P_T(\{\boldsymbol{X}_\alpha\})}{P(\{\boldsymbol{X}_\alpha\})} d\boldsymbol{X}_1 \cdots d\boldsymbol{X}_N \\ &= E[\log P_T(\{\boldsymbol{X}_\alpha\})] - E[\log P(\{\boldsymbol{X}_\alpha\})], \end{aligned} \tag{12}$$

where $E[\,\cdot\,]$ denotes expectation with respect to the true (unknown) probability density $P_T(\{\boldsymbol{X}_\alpha\})$. The assumed model is regarded as good if D is small.

Substituting (5) and noting that $E[\log P_T(\{\boldsymbol{X}_\alpha\})]$ does not depend on individual models, we regard the model as good if

$$\begin{aligned} &-E[\log P(\{\boldsymbol{X}_\alpha\})] \\ &\quad = \frac{1}{2\varepsilon^2} E\left[\sum_{\alpha=1}^{N} (\boldsymbol{X}_\alpha - \bar{\boldsymbol{x}}_\alpha, V_0[\boldsymbol{x}_\alpha]^{-1}(\boldsymbol{X}_\alpha - \bar{\boldsymbol{x}}_\alpha))\right] \\ &\quad\quad + \frac{mN}{2} \log 2\pi\varepsilon^2 + \frac{1}{2} \sum_{\alpha=1}^{N} \log |V_0[\boldsymbol{x}_\alpha]| \end{aligned} \tag{13}$$

is small. The last two terms on the right-hand side do not depend on individual models. So, multiplying the first term by $2\varepsilon^2$, we seek a model that minimizes the *expected residual*

$$E = E\left[\sum_{\alpha=1}^{N} (\boldsymbol{X}_\alpha - \bar{\boldsymbol{x}}_\alpha, V_0[\boldsymbol{x}_\alpha]^{-1}(\boldsymbol{X}_\alpha - \bar{\boldsymbol{x}}_\alpha))\right]. \tag{14}$$

6.2 Evaluation of Expectation

The difficulty of using (14) as a model selection criterion is that the expectation $E[\,\cdot\,]$ must be evaluated using the *true* density, which we do not know. Here arises a sharp distinction between the standard statistical analysis, in which Akaike was interested, and the geometric inference problem, in which we are interested, as to how to evaluate the expectation.

For the standard statistical analysis, we assume that we could, at least in principle, observe as many data as desired. If we are allowed to sample independent instances $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n$

according to a density $P_T(\boldsymbol{X})$, the expectation $E[Y(\boldsymbol{X})] = \int Y(\boldsymbol{X})P_T(\boldsymbol{X})d\boldsymbol{X}$ of a statistic $Y(\boldsymbol{X})$ can be approximated by the sample mean $(1/n)\sum_{i=1}^{n} Y(\boldsymbol{x}_i)$, 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 contrast, we can obtain only *one* instance $\{\boldsymbol{x}_\alpha\}$ of $\{\boldsymbol{X}_\alpha\}$ for geometric inference, so we cannot replace expectation by the sample mean. However, we are interested only in the limit $\varepsilon \to 0$. So, the expectation

$$E[Y(\{\boldsymbol{X}_\alpha\})] = \int \cdots \int Y(\{\boldsymbol{X}_\alpha\})P_T(\{\boldsymbol{X}_\alpha\})d\boldsymbol{X}_1 \cdots d\boldsymbol{X}_N$$

can be approximated by $Y(\{\boldsymbol{x}_\alpha\})$ because, as $\varepsilon \to 0$, 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. It follows that we can approximate E as follows (note that $1/N$ is not necessary):

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

6.3 Bias Removal

There is still a difficulty using (15) as a criterion: The model parameters $\{\bar{\boldsymbol{x}}_\alpha\}$ and $\boldsymbol{u}$ need to be estimated. If we view (15) as a measure of the goodness of the model, we should compute their ML estimators $\{\hat{\boldsymbol{x}}_\alpha\}$ and $\hat{\boldsymbol{u}}$, minimizing (15) subject to the constraint (1). Substituting $\{\hat{\boldsymbol{x}}_\alpha\}$ and $\hat{\boldsymbol{u}}$ for $\{\bar{\boldsymbol{x}}_\alpha\}$ and $\boldsymbol{u}$ in (15), we obtain the *residual* (*sum of squares*):

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

Here, a logical inconsistency arises. Equation 3 defines not a particular model but a *class* of models parameterized by $\{\bar{\boldsymbol{x}}_\alpha\}$ and $\boldsymbol{u}$. If we choose particular values $\{\hat{\boldsymbol{x}}_\alpha\}$ and $\hat{\boldsymbol{u}}$ (i.e., the ML-estimators), we are given a particular model. According to the logic in Section 6.1, 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 Section 6.2, 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 $\{\boldsymbol{X}_\alpha\}$. 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 so determined 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

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

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, we have the current data $\{\boldsymbol{x}_\alpha\}$ only.[1] So, we try to compensate for the bias in the form

$$\hat{J}^* = \hat{J} + b\varepsilon^2. \tag{18}$$

Both $\hat{J}^*$ and $\hat{J}$ are $O(\varepsilon^2)$, so b is $O(1)$. Since $\hat{J}^*$ and $\hat{J}$ are random variables, so is b. It can be proven [10], [11] that

$$E^*[E[b]] = 2(Nd + p) + O(\varepsilon^2), \tag{19}$$

where $E[\,\cdot\,]$ and $E^*[\,\cdot\,]$ denote expectations for $\{\boldsymbol{x}_\alpha\}$ and $\{\boldsymbol{x}_\alpha^*\}$, respectively, and $d = m - r$ is the dimension of the manifold $\mathcal{S}$ defined the constraint $F^{(k)}(\boldsymbol{x}, \boldsymbol{u}) = 0, k = 1, \ldots, r$ (recall that p is the dimension of the parameter vector $\boldsymbol{u}$).

Thus, we obtain an unbiased estimator of $\hat{J}^*$ in the first order in the form

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

which is the *geometric AIC* of Kanatani [10], [11], who derived (19) directly. Here, we have given a new justification by going back to the Kullback-Leibler distance (12).

7 Geometric MDL

We now turn to Rissanen's MDL [28], [29] and derive the geometric MDL by doing asymptotic analysis as the noise level ε goes to zero.

7.1 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.
- The probability density, from which a minimum length code can be obtained, involves unknown parameters.
- The exact form of the minimum code length is very difficult to compute.

Rissanen [28], [29] avoided these difficulties by quantizing the real numbers in a way that does not depend on individual models and substituting the ML estimators for the parameters. They, too, are real numbers, so they are also quantized. The quantization width is so chosen as to minimize the total description length (the *two-stage encoding*). The resulting code length is evaluated asymptotically as the data length n goes to infinity. This idea is translated for geometric inference 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

$$-\log P = \frac{J}{2\varepsilon^2} + \frac{mN}{2}\log 2\pi\varepsilon^2 + \frac{1}{2}\sum_{\alpha=1}^{N} \log |V_0[\boldsymbol{x}_\alpha]|, \tag{21}$$

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 in (6). Using the natural logarithm, we take $\log_2 e$ bits as the unit of length.

Note the similarity and contrast to the geometric AIC, which minimizes the *expectation* of (21) (see (13)), while, here, (21) is directly minimized with a different interpretation.

[1] If such data $\{\boldsymbol{x}_\alpha^*\}$ actually exists, the test using them is called *cross-validation*. We can also generate equivalent data by a computer. Such a simulations is called *bootstrap* [6].

7.2 Two-Stage Encoding

In order to do encoding using (5), we need the true values $\{\bar{\boldsymbol{x}}_\alpha\}$ and the parameter $\boldsymbol{u}$. Since they are unknown, we use their ML estimators that minimize (21) (specifically J). The last two terms of (21) do not depend on individual models, so the minimum code length is $\hat{J}/2\varepsilon^2$ up to a constant, where $\hat{J}$ is the residual in (16). 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 ML estimators $\{\hat{\boldsymbol{x}}_\alpha\}$ and $\hat{\boldsymbol{u}}$ are real numbers, they also need to be quantized. If we use a larger quantization width, their code lengths become shorter, but the description length $\hat{J}/2\varepsilon^2$ will increase. So, we take the width that minimizes the total description length. The starting point is the fact that (7) can be written as follows [10]:

$$\begin{aligned} J =& \hat{J} + \sum_{\alpha=1}^{N} (\boldsymbol{x}_\alpha - \hat{\boldsymbol{x}}_\alpha, V_0[\hat{\boldsymbol{x}}_\alpha]^- (\boldsymbol{x}_\alpha - \hat{\boldsymbol{x}}_\alpha)) \\ &+ (\boldsymbol{u} - \hat{\boldsymbol{u}}, V_0[\hat{\boldsymbol{u}}]^{-1} (\boldsymbol{u} - \hat{\boldsymbol{u}})) + O(\varepsilon^3). \end{aligned} \tag{22}$$

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 ML estimators $\hat{\boldsymbol{x}}_\alpha$ and $\hat{\boldsymbol{u}}$ given as follows [10]:

$$\begin{aligned} V_0[\hat{\boldsymbol{x}}_\alpha] &= V_0[\boldsymbol{x}_\alpha] \\ &\quad - \sum_{k,l=1}^{r} W_\alpha^{(kl)} (V[\boldsymbol{x}_\alpha] \nabla_{\mathbf{x}} F_\alpha^{(k)}) (V[\boldsymbol{x}_\alpha] \nabla_{\mathbf{x}} F_\alpha^{(k)})^\top, \\ V_0[\hat{\boldsymbol{u}}] &= \left(\sum_{\alpha=1}^{N} \sum_{k,l=1}^{r} W_\alpha^{(kl)} (\nabla_{\mathbf{u}} F_\alpha^{(k)}) (\nabla_{\mathbf{u}} F_\alpha^{(l)})^\top \right)^{-1}. \end{aligned} \tag{23}$$

The symbol $W_\alpha^{(kl)}$ has the same meaning as in (7). It is easily seen that $V_0[\hat{\boldsymbol{x}}_\alpha]^-$ is a singular matrix of rank d whose domain is the tangent space to the optimally fitted manifold $\hat{\mathcal{S}}$ at $\hat{\boldsymbol{x}}_\alpha$.

7.3 Encoding Parameters

In order to quantize $\hat{\boldsymbol{u}}$, we introduce appropriate (generally curvilinear) coordinates $(u_i), i = 1, \ldots, p$ into the p-dimensional parameter space $\mathcal{U}$ and quantize it into a grid of width δu_i. Suppose $\hat{\boldsymbol{u}}$ is in a (curvilinear) rectangular region of sides 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

$$\log \prod_{i=1}^{p} \frac{L_i}{\delta u_i} = \log V_u - \sum_{i=1}^{p} \log \delta u_i, \tag{24}$$

where $V_u = \prod_{i=1}^{p} L_i$ is the volume of the rectangular region. We could reduce (24) using a large width δu_i, but (22) implies that replacing $\hat{\boldsymbol{u}}$ by the nearest vertex would increase the description length $\hat{J}/2\varepsilon^2$ by $(\delta \boldsymbol{u}, V_0[\hat{\boldsymbol{u}}]^{-1} \delta \boldsymbol{u})/2\varepsilon^2$ in the first order in ε, where we define $\delta \boldsymbol{u} = (\delta u_i)$. So, we choose such $\delta \boldsymbol{u}$ that minimizes the sum of $(\delta \boldsymbol{u}, V_0[\hat{\boldsymbol{u}}]^{-1} \delta \boldsymbol{u})/2\varepsilon^2$ and (24). Differentiating this sum with respect to δu_i and letting the result be 0, we obtain

$$\frac{1}{\varepsilon^2} \left(V_0[\hat{\boldsymbol{u}}]^{-1} \delta \boldsymbol{u} \right)_i = \frac{1}{\delta u_i}, \tag{25}$$

where $(\,\cdot\,)_i$ designates the ith component. If the coordinate system of $\mathcal{U}$ is so taken that $V_0[\hat{\boldsymbol{u}}]^{-1}$ is diagonalized, (25) reduces to

$$\delta u_i = \frac{\varepsilon}{\sqrt{\lambda_i}}, \tag{26}$$

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

$$v_u = \prod_{i=1}^{p} \delta u_i = \frac{\varepsilon^p}{\sqrt{|V_0[\hat{\boldsymbol{u}}]^{-1}|}}. \tag{27}$$

Hence, the number of cells inside the region V_u is

$$N_u = \int_{V_u} \frac{d\boldsymbol{u}}{v_u} = \frac{1}{\varepsilon^p} \int_{V_u} \sqrt{|V_0[\hat{\boldsymbol{u}}]^{-1}|} d\boldsymbol{u}. \tag{28}$$

Specifying one from these requires the code length

$$\log N_u = \log \int_{V_u} \sqrt{|V_0[\hat{\boldsymbol{u}}]^{-1}|} d\boldsymbol{u} - \frac{p}{2} \log \varepsilon^2. \tag{29}$$

7.4 Encoding True Values

For quantizing the ML-estimators $\{\hat{\boldsymbol{x}}_\alpha\}$, we need not quantize the entire m-dimensional data space $\mathcal{X}$ because they are constrained to be in the optimally fitted d-dimensional manifold $\hat{\mathcal{S}}$ ($\subset \mathcal{X}$) specified by $\hat{\boldsymbol{u}}$, which we have already encoded. So, we only need to quantize $\hat{\mathcal{S}}$. To this end, we introduce appropriate curvilinear coordinates in it. Since each $\hat{\boldsymbol{x}}_\alpha$ has its own normalized covariance matrix $V_0[\hat{\boldsymbol{x}}_\alpha]$ (see (23)), we introduce different coordinates $(\xi_{i\alpha}), i = 1, \ldots, d$ for each α. Then, they are quantized into a (curvilinear) grid of width $\delta \xi_{i\alpha}$.

Suppose $\hat{\boldsymbol{x}}_\alpha$ is in a (curvilinear) rectangular region of sides $l_{i\alpha}$. 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

$$\log \prod_{i=1}^{d} \frac{l_{i\alpha}}{\delta \xi_{i\alpha}} = \log V_{x\alpha} - \sum_{i=1}^{d} \log \delta \xi_{i\alpha}, \tag{30}$$

where $V_{x\alpha} = \prod_{i=1}^{d} l_{i\alpha}$ is the volume of the rectangular region. We could reduce (30) using a large width $\delta \xi_{i\alpha}$, but replacing $\hat{\boldsymbol{x}}_\alpha$ by its nearest vertex would increase the description length $\hat{J}/2\varepsilon^2$. Let $\delta \bar{\boldsymbol{x}}_\alpha$ be the m-dimensional vector that expresses the displacement $\{\delta \xi_{i\alpha}\}$ on $\hat{\mathcal{S}}$ in the (original) coordinates of $\mathcal{X}$. Equation (22) implies that the increase in $\hat{J}/2\varepsilon^2$ is $(\delta \bar{\boldsymbol{x}}_\alpha, V_0[\hat{\boldsymbol{x}}_\alpha]^- \delta \bar{\boldsymbol{x}}_\alpha)/2\varepsilon^2$ in the first order in ε, so we choose such $\{\delta \xi_{i\alpha}\}$ that minimize the sum of $(\delta \bar{\boldsymbol{x}}_\alpha, V_0[\hat{\boldsymbol{x}}_\alpha]^- \delta \bar{\boldsymbol{x}}_\alpha)/2\varepsilon^2$ and (30). Differentiating this sum with respect to $\delta \xi_{i\alpha}$ and letting the result be 0, we obtain

$$\frac{1}{\varepsilon^2} (V_0[\hat{\boldsymbol{x}}_\alpha]^- \delta \bar{\boldsymbol{x}}_\alpha)_i = \frac{1}{\delta \xi_{i\alpha}}. \tag{31}$$

Let the coordinates $(\xi_{i\alpha})$ be such that the d basis vectors at $\hat{\boldsymbol{x}}_\alpha$ form an orthonormal system. Also, let the coordinates of $\mathcal{X}$ be such that at $\hat{\boldsymbol{x}}_\alpha \in \hat{\mathcal{S}}$ the m basis vectors consist of the d basis vectors of $\hat{\mathcal{S}}$ plus $m - d$ additional basis vectors orthogonal to $\hat{\mathcal{S}}$. Then, the first d components of $\delta \bar{\boldsymbol{x}}_\alpha$ coincide with $\{\delta \xi_{i\alpha}\}, i = 1, \ldots, d$; the remaining components are 0. If, furthermore, the coordinates $(\xi_{i\alpha})$ are so defined that $V_0[\hat{\boldsymbol{x}}_\alpha]^-$ is diagonalized, the solution $\delta \xi_{i\alpha}$ of (31) is given by

$$\delta\xi_{i\alpha} = \frac{\varepsilon}{\sqrt{\lambda_{i\alpha}}}, \tag{32}$$

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

$$v_{x\alpha} = \prod_{i=1}^{d} \delta\xi_{i\alpha} = \frac{\varepsilon^d}{\sqrt{|V_0[\hat{\boldsymbol{x}}_\alpha]^-|_d}}, \tag{33}$$

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

$$N_\alpha = \int_{V_{x\alpha}} \frac{d\boldsymbol{x}}{v_{x\alpha}} = \frac{1}{\varepsilon^d}\int_{V_{x\alpha}} \sqrt{|V_0[\hat{\boldsymbol{x}}_\alpha]^-|_d}\, d\boldsymbol{x}. \tag{34}$$

Specifying one from these requires the code length

$$\log N_\alpha = \log \int_{V_{x\alpha}} \sqrt{|V_0[\hat{\boldsymbol{x}}_\alpha]^-|_d}\, d\boldsymbol{x} - \frac{d}{2}\log\varepsilon^2. \tag{35}$$

7.5 Geometric MDL

From (29) and (35), the total code length for $\{\hat{\boldsymbol{x}}_\alpha\}$ and $\hat{\boldsymbol{u}}$ becomes

$$\sum_{\alpha=1}^{N} \log \int_{V_{x\alpha}} \sqrt{|V_0[\hat{\boldsymbol{x}}_\alpha]^-|_d}\, d\boldsymbol{x} + \log \int_{V_u} \sqrt{|V_0[\hat{\boldsymbol{u}}]^{-1}|} d\boldsymbol{u} - \frac{Nd+p}{2}\log\varepsilon^2. \tag{36}$$

The accompanying increase in the description length $\hat{J}/2\varepsilon^2$ is

$$(\delta\bar{\boldsymbol{x}}_\alpha, V_0[\hat{\boldsymbol{x}}_\alpha]^-\delta\bar{\boldsymbol{x}}_\alpha)/2\varepsilon^2 + (\delta\boldsymbol{u}, V_0[\hat{\boldsymbol{u}}]^{-1}\delta\boldsymbol{u})/2\varepsilon^2$$

in the first order in ε. If we substitute (26) and (32) together with $V_0[\hat{\boldsymbol{x}}_\alpha]^- = \mathrm{diag}(1/\lambda_{1\alpha}, \ldots, 1/\lambda_{d\alpha}, 0, \ldots, 0)$ and $V_0[\hat{\boldsymbol{u}}]^{-1} = \mathrm{diag}(1/\lambda_1, \ldots, 1/\lambda_p)$, this increase is

$$\frac{(\delta\bar{\boldsymbol{x}}_\alpha, V_0[\hat{\boldsymbol{x}}_\alpha]^-\delta\bar{\boldsymbol{x}}_\alpha)}{2\varepsilon^2} + \frac{(\delta\boldsymbol{u}, V_0[\hat{\boldsymbol{u}}]^{-1}\delta\boldsymbol{u})}{2\varepsilon^2} = \frac{Nd+p}{2}. \tag{37}$$

Since (26) and (32) are obtained by omitting terms of $o(\varepsilon)$, the omitted terms in (37) are $o(1)$. It follows that the total description length is

$$\frac{\hat{J}}{2\varepsilon^2} - \frac{Nd+p}{2}\log\varepsilon^2 + \sum_{\alpha=1}^{N}\log\int_{V_{x\alpha}}\sqrt{|V_0[\hat{\boldsymbol{x}}_\alpha]^-|_d}\, d\boldsymbol{x} + \log\int_{V_u}\sqrt{|V_0[\hat{\boldsymbol{u}}]^{-1}|} d\boldsymbol{u} + \frac{Nd+p}{2} + o(1). \tag{38}$$

Multiplying this by $2\varepsilon^2$, which does not affect model selection, we obtain

$$\hat{J} - (Nd+p)\varepsilon^2\log\varepsilon^2 + 2\varepsilon^2\left(\sum_{\alpha=1}^{N}\log\int_{V_{x\alpha}}\sqrt{|V_0[\hat{\boldsymbol{x}}_\alpha]^-|_d}\, d\boldsymbol{x} + \log\int_{V_u}\sqrt{|V_0[\hat{\boldsymbol{u}}]^{-1}|} d\boldsymbol{u}\right) + (Nd+p)\varepsilon^2 + o(\varepsilon^2). \tag{39}$$

7.6 Scale Choice

In practice, it is difficult to use (39) as a criterion because of the difficulty in evaluating the third term. If we note that $-\log\varepsilon^2 \gg 1$ as $\varepsilon \to 0$, we may omit terms of $O(\varepsilon^2)$ and define

$$\text{G-MDL} = \hat{J} - (Nd+p)\varepsilon^2\log\varepsilon^2. \tag{40}$$

This is the form suggested by Matsunaga and Kanatani [22]. However, 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, $\log\varepsilon^2$ reduces by $\log 100$, which could affect model selection.[2] In (39), in contrast, the influence of scale is canceled between the second and third terms.

To begin with, the logarithm can be defined only for a nondimensional quantity, so (40) should have the form

$$\text{G-MDL} = \hat{J} - (Nd+p)\varepsilon^2\log\left(\frac{\varepsilon}{L}\right)^2, \tag{41}$$

where L is a reference length. In theory, it can be determined from the third term of (39), but its evaluation is difficult. So, we adopt a practical compromise, choosing a scale L such that x_α/L is $O(1)$. We may interpret this as introducing a prior distribution in a region of volume L^m in the data space $\mathcal{X}$. For example, if $\{x_\alpha\}$ are image coordinate data, we can take L to be the image size. We call (41) the *geometric MDL*.

Remark 9. Recall that, for asymptotic analysis as $\varepsilon \to 0$, it is essential to fix the scale of the normalized covariance matrix $V_0[\boldsymbol{x}_\alpha]$ in (4) in such a way that the noise level ε is much smaller than the data themselves (Remark 2). So, we have $-\log(\varepsilon/L)^2 \gg 1$. If we use a different scale $L' = \gamma L$, we have $-\log(\varepsilon/L')^2 = -\log(\varepsilon/L)^2 + \log\gamma^2 \approx -\log(\varepsilon/L)^2$ as long as the scale is of the same order of magnitude. It has been confirmed that the scale choice does not practically affect model selection in most applications. Nonetheless, the introduction of the scale is a heuristic compromise, and more studies about this will be necessary.

8 Some Issues of the Geometric AIC/MDL

8.1 Dual Interpretations of Model Selection

We have observed in Section 4.3 that the standard statistical analysis and the geometric inference problem have a duality in the sense that $1/\sqrt{n}$ for the former plays the role of ε for geometric inference. The same holds for model selection, too. Akaike's AIC is

$$\text{AIC} = -2\log\prod_{i=1}^{n} P(\boldsymbol{x}_i|\hat{\boldsymbol{\theta}}) + 2k, \tag{42}$$

where $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n$ are n samples from the density $P(\boldsymbol{x}|\boldsymbol{\theta})$ parameterized by a k-dimensional vector $\boldsymbol{\theta}$, and $\hat{\boldsymbol{\theta}}$ is its ML estimator.

For the geometric fitting problem, on the other hand, the unknowns are the p parameters of the constraint plus the N true positions specified by the d coordinates of the d-dimensional manifold $\mathcal{S}$. If (20) is divided by ε^2, we have $\hat{J}/\varepsilon^2 + 2(Nd+p)$, which is ($-2$ times the logarithmic

[2] The preference is unchanged if the candidate models have the same d and p, but we usually compare models of different d and p.

likelihood)+2(the number of unknowns), the same form as (42). The same holds for (41), which corresponds to Rissanen's MDL (see (43) and (44) to be explained below) if ε is replaced by $1/\sqrt{n}$.

This correspondence can be understood if we recall our observation that the limit $\varepsilon \to 0$ is mathematically equivalent to sampling a large number n of potential algorithms (Section 4.3).

8.2 Priors for the Geometric MDL

For the geometric MDL, one can notice that the coding scheme described in Section 7 cannot apply if the manifold $\mathcal{S}$ has a complicated shape. In fact, our derivation went as if the manifold $\mathcal{S}$ were flat and compact. This is justified only when the data $\{\boldsymbol{x}_i\}$ and the parameter $\boldsymbol{u}$ are found in fairly small regions.

Take (39) for example. The regions V_{xi} and V_u must be compact for the integrations to 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 finite length code unless some information about their likely locations is given. The expedient for deriving (41) is in a sense reducing the Bayesian prior to a single scale L.

This type of implicit reduction is also present in Rissanen's MDL, for which the data length n is the asymptotic variable. Originally, Rissanen presented his MDL in the following form [28]:

$$\mathrm{MDL} = -\log \prod_{i=1}^{n} P(\boldsymbol{x}_i|\hat{\boldsymbol{\theta}}) + \frac{k}{2}\log n. \tag{43}$$

As in the case of Akaike's AIC, $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n$ are n samples from the density $P(\boldsymbol{x}|\boldsymbol{\theta})$ parameterized by a k-dimensional vector $\boldsymbol{\theta}$, and $\hat{\boldsymbol{\theta}}$ is its ML estimator.

This form evoked the problem of the "unit." If we regard a pair of data as one datum, viewing $(\boldsymbol{x}_1, \boldsymbol{x}_2), (\boldsymbol{x}_3, \boldsymbol{x}_4), \ldots$ as samples from $P(\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{\theta}) = P(\boldsymbol{x}|\boldsymbol{\theta})P(\boldsymbol{y}|\boldsymbol{\theta})$, the data length is halved, though the problem is the same. Later, Rissanen presented the following form [30]:

$$\mathrm{MDL} = -\log \prod_{i=1}^{n} P(\boldsymbol{x}_i|\hat{\boldsymbol{\theta}}) + \frac{k}{2}\log\frac{n}{2\pi} + \log\int_{V_\theta}\sqrt{|\boldsymbol{I}(\boldsymbol{\theta})|}d\boldsymbol{\theta}. \tag{44}$$

Here, $\boldsymbol{I}(\boldsymbol{\theta})$ is the Fisher information of $P(\boldsymbol{x}|\boldsymbol{\theta})$. In this form, the unit change is canceled by the corresponding change in the Fisher information. However, the problem of integration arises if the domain V_θ is unbounded, so an appropriate prior is necessary.

Thus, Rissanen's MDL and the geometric MDL share the same properties whether we focus on the limit $n \to \infty$ or the limit $\varepsilon \to 0$, confirming our previous observation about the dual interpretation.

8.3 Noise-Level Estimation

In order to use the geometric AIC or the geometric MDL, we need to know the noise level ε. If not known, it must be estimated. Here arises a sharp contrast between the standard statistical analysis and our geometric inference.

For the standard statistical analysis, the noise magnitude is a *model parameter* because "noise" is defined to be *the random effects that cannot be accounted for by the assumed model*. Hence, the noise magnitude should be estimated, if not known, *according to the assumed model*. For geometric inference, on the other hand, the noise level ε is *a constant that reflects the uncertainty of feature detection*. So, it should 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 $\hat{J}/\varepsilon^2$ is subject to a χ^2 distribution with $rN - p$ degrees of freedom in the first order [10]. Specifically, we obtain an unbiased estimator of ε^2 in the form

$$\hat{\varepsilon}^2 = \frac{\hat{J}}{rN - p}. \tag{45}$$

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, a typical situation where model selection is called for is *degeneracy detection*. In 3D analysis from images, for example, the constraint (3) corresponds to our knowledge about the scene such as rigidity of motion. However, the computation fails if degeneracy occurs (e.g., the motion is zero). Even if exact degeneracy does not occur, the computation may become numerically unstable in near degeneracy conditions. In such a case, the computation can be stabilized by switching to a model that describes the degeneracy [18], [19], [22], [26], [39].

Degeneracy means *addition* of new constraints, such as some quantity being zero. It follows that the manifold $\mathcal{S}$ degenerates into a submanifold $\mathcal{S}'$ of it. Since the general model still holds irrespective of the degeneracy, i.e., $\mathcal{S}' \subset \mathcal{S}$, we can estimate the noise level ε from the residual $\hat{J}$ of the general model $\mathcal{S}$, which we know is true, using (45).

Remark 10. Equation (45) can be intuitively understood as follows. Recall that $\hat{J}$ is the sum of the square distances from $\{\boldsymbol{x}_\alpha\}$ to the manifold $\hat{\mathcal{S}}$ defined by the constraint $F^{(k)}(\boldsymbol{x}, \boldsymbol{u}) = 0, k = 1, \ldots, r$. Since $\hat{\mathcal{S}}$ has codimension r (the dimension of the orthogonal directions to it), the residual $\hat{J}$ should have expectation $rN\varepsilon^2$. However, $\hat{\mathcal{S}}$ is fitted by adjusting its p-dimensional parameter $\boldsymbol{u}$, so the expectation of $\hat{J}$ reduces to $(rN - p)\varepsilon^2$.

Note that we need more than $\lfloor p/r \rfloor$ data for this estimation. For example, if we know that the true model is a planar surface, we need to observe more than three points for degeneracy detection.

Remark 11. It may appear that the residual $\hat{J}$ of the general model cannot be stably computed in the presence of degeneracy. However, what is unstable is *model specification*, not the residual. For example, if we fit a planar surface to almost collinear points in 3D, it is difficult to specify the fitted plane stably; the solution is very susceptible to noise. Yet, the residual is stably computed since unique specification of the fit is difficult *because all the candidates have almost the same residual*.

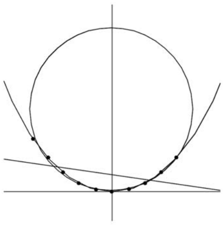


Fig. 3. Fitting a line, a circle, and an ellipse.

Note that the noise-level estimation from the general model $\mathcal{S}$ by (45) is still valid even if degeneracy occurs, because degeneracy means shrinkage of the model manifold $\mathcal{S}'$ *within* $\mathcal{S}$, which does not affect the data deviations in the "orthogonal" directions (in the Mahalanobis sense) to $\mathcal{S}$ that account for the residual $\hat{J}$.

9 Deneracy Detection Experiments

We now illustrate the different characteristics of the geometric AIC and the geometric MDL for degeneracy detection.

9.1 Detection of Circles and Lines

Consider an ellipse that is tangent to the x-axis at the origin O with the minor radius 50 in the y direction and eccentricity $1/\beta$. On it, we take 11 points with equally spaced x coordinates. Adding 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[3] [16], [17]. Fig. 3 shows one instance for $\beta = 2.5$ and $\varepsilon = 0.1$. Note that a line and a circle are degeneracies of an ellipse.

Lines, circles, and ellipses define 1-dimensional (geometric) models with 2, 3, and 5 degrees of freedom, respectively. Their geometric AIC and the geometric MDL for N points are

$$\begin{aligned} \text{G-AIC}_l &= \hat{J}_l + 2(N+2)\varepsilon^2, \\ \text{G-AIC}_c &= \hat{J}_c + 2(N+3)\varepsilon^2, \\ \text{G-AIC}_e &= \hat{J}_e + 2(N+5)\varepsilon^2, \\ \text{G-MDL}_l &= \hat{J}_l - (N+2)\varepsilon^2 \log\left(\frac{\varepsilon}{L}\right)^2, \\ \text{G-MDL}_c &= \hat{J}_c - (N+3)\varepsilon^2 \log\left(\frac{\varepsilon}{L}\right)^2, \\ \text{G-MDL}_e &= \hat{J}_e - (N+5)\varepsilon^2 \log\left(\frac{\varepsilon}{L}\right)^2, \end{aligned} \tag{46}$$

where 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 value. We used the reference length $L = 1$.

Fig. 4a shows the percentage of choosing a line for $\varepsilon = 0.01$ after 1,000 independent trials for each β. If there was no noise, it should be 0 percent for $\beta \neq 0$ and 100 percent for $\beta = 0$. In the presence of noise, the geometric AIC produces 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 percent, but it judges an ellipse to be a line over a wide range of β.

In Fig. 4a, we used the true value of ε^2. If it is unknown, it can be estimated from the residual of the general ellipse model by (45). Fig. 4b shows the result using its estimate. Although the sharpness is somewhat lost, similar performance characteristics are observed.

Fig. 5 shows the percentage of choosing a circle for $\varepsilon = 0.01$. If there were no noise, it should be 0 percent for $\beta \neq 1$ and 100 percent for $\beta = 1$. In the presence of noise, as we see, it is difficult to distinguish a circular arc from an 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 $\beta < 1.1$.

9.2 Detection of Space Lines

Consider a rectangular region $[0, 10] \times [-1, 1]$ on the xy plane in the xyz space. We randomly take 11 points in it and magnify the region A times in the y direction. Adding Gaussian noise of mean 0 and variance ε^2 to the x, y, and z coordinates of each point independently, we fit a space line and a plane in a statistically optimal manner (Fig. 6a). The rectangular region degenerates into a line segment as $A \to 0$.

A space line is a 1-dimensional model with four degrees of freedom; a plane is a 2-dimensional model with three degrees of freedom. Their geometric AIC and geometric MDL are

$$\begin{aligned} \text{G-AIC}_l &= \hat{J}_l + 2(N+4)\varepsilon^2, \\ \text{G-AIC}_p &= \hat{J}_p + 2(2N+3)\varepsilon^2, \\ \text{G-MDL}_l &= \hat{J}_l - (N+4)\varepsilon^2 \log\left(\frac{\varepsilon}{L}\right)^2, \\ \text{G-MDL}_p &= \hat{J}_p - (2N+3)\varepsilon^2 \log\left(\frac{\varepsilon}{L}\right)^2, \end{aligned} \tag{47}$$

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

Fig. 6b shows the percentage of choosing a line for $\varepsilon = 0.01$ after 1,000 independent trials for each A. If there were no noise, it should be 0 percent for $A \neq 0$ and 100 percent for $A = 0$. In the presence of noise, the geometric AIC has a high capability of distinguishing a line from a plane, but it judges a line to be a plane with some probability. In contrast, the geometric MDL judges a line to be a line almost 100 percent, but it judges a plane to be a line over a wide range of A.

In Fig. 6b, we used the true value of ε^2. Fig. 6c shows the corresponding result using its estimate obtained from the general plane model by (45). We observe somewhat degraded but similar performance characteristics.

9.3 Observations

We have observed that the geometric AIC has a higher capability for detecting degeneracy than the geometric MDL, but the general model is chosen with some

[3] We used by a technique called *renormalization* [10].

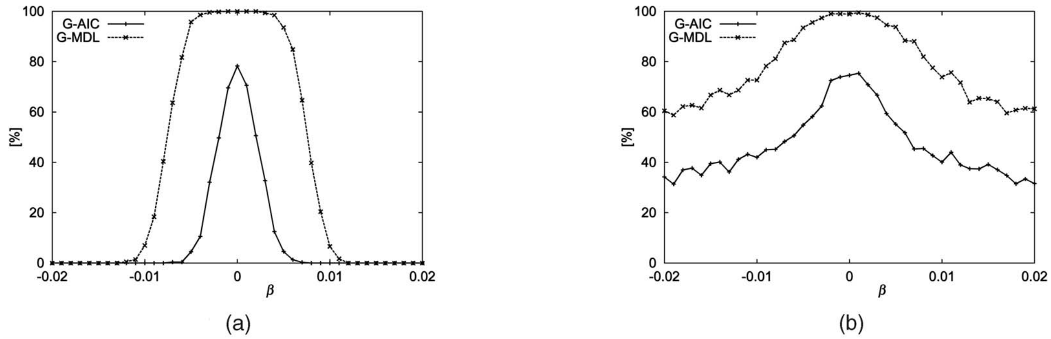


Fig. 4. The ratio (%) of detecting a line by the geometric AIC (solid lines with +) and the geometric MDL (dotted lines with ×) using (a) the true noise level and (b) the estimated noise level.

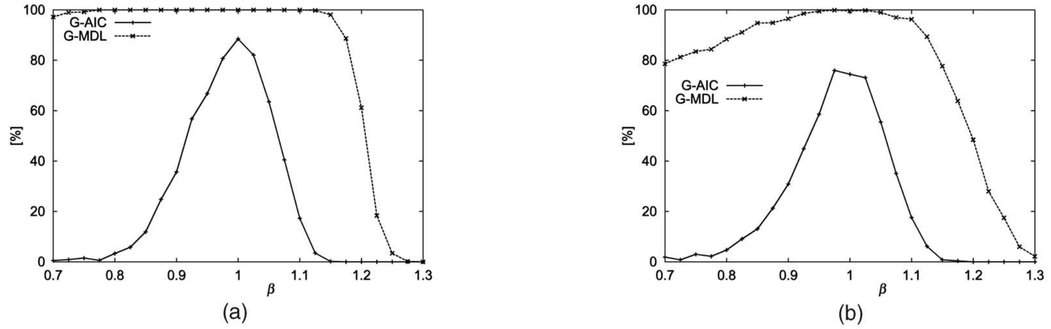


Fig. 5. The ratio (%) of detecting a circle by the geometric AIC (solid lines with +) and the geometric MDL (dotted lines with ×) using (a) the true noise level and (b) the estimated noise level.

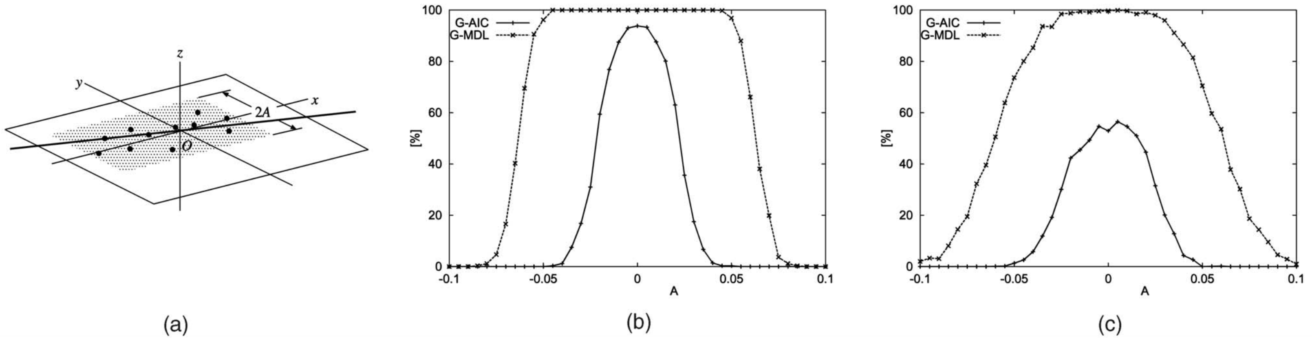


Fig. 6. (a) Fitting a space line and a plane to points in space. (b) The rate (%) of detecting a space line by the geometric AIC (solid lines with +) and the geometric MDL (dotted lines with ×) with the true noise level. (c) The corresponding results using the estimated noise level.

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

At the cost of this consistency, however, the geometric MDL regards a wide range of nondegenerate models as degenerate. This is no surprise, since the penalty $-(Nd+p)\varepsilon^2\log(\varepsilon/L)^2$ for the geometric MDL in (41) is heavier than the penalty $2(Nd+p)\varepsilon^2$ for the geometric AIC in (20). As a result, the geometric AIC is more faithful to the data than the geometric MDL, which is more likely to choose a degenerate model. This contrast has also been observed in many applications [15], [22].

Remark 12. Despite the fundamental difference of geometric model selection from the standard (stochastic) model selection, many attempts have been made in the past to apply Akaike's AIC and their variants to computer vision problems based on the asymptotic analysis of $n \to \infty$, where the interpretation of n is different from problem to problem [34], [35], [36], [37], [38]. Rissanen's MDL is also used in computer vision applications. Its use may be justified if the problem has the standard form of linear/nonlinear regression [2], [23]. Often, however, the solution having a shorter description length was chosen with a rather arbitrary definition of the complexity [7], [21], [24].

Remark 13. One may wonder why we are forced to choose one from the two asymptotic analyses, $n \to \infty$ or $\varepsilon \to 0$.

Why don't we use the general form of the AIC or the MDL rather than worrying about their asymptotic expressions? The answer is that we *cannot*.

The starting principle of the AIC is the Kullback-Leibler distance of the assumed probability distribution from the true distribution (Section 6.1). We cannot compute it exactly because we do not know the true distribution. So, Akaike approximated it, invoking the law of large numbers and the central limit theorem, thereby estimating the true distribution from a large number of observations, while the geometric AIC is obtained by assuming that the noise is very small, thereby identifying the data as their true values to a first approximation (Section 6.2).

Similarly, the exactly shortest code length is difficult to compute if real numbers are involved, so Rissanen approximated it by omitting higher order terms in the data length n. The geometric MDL is obtained by omitting higher order terms in the noise level ε (Section 7).

Thus, analysis of asymptotic expressions in one form or another is inevitable if the principle of the AIC or the MDL is to be applied in practice.

Remark 14. Note that one cannot say one model selection criteria is superior to another, because each is based on its own logic. Also, if we want to compare the performance of two criteria in practice, we must formulate them in such a way that they conform to a common assumption. In this sense, one cannot compare Akaike's AIC with the geometric AIC or Rissanen's MDL with the geometric MDL, because the underlying asymptotic limits are different. Similarly, if we want to compare the geometric AIC or the geometric MDL with other existing criteria, e.g., Schwarz' BIC, derived in the asymptotic limit $n \to \infty$, they must be formulated in the asymptotic limit $\varepsilon \to 0$.

Note also that one cannot prove that a particular criterion works at all. In fact, although Akaike's AIC and Rissanen's MDL are based on rigorous mathematics, there is no guarantee that they work well in practice. The mathematical rigor is in their *reduction* from their starting principles (the Kullback-Leibler distance and the minimum description length principle), which are beyond proof. What one can tell is which criterion is more suitable for a particular application when used in a particular manner. The geometric AIC and the geometric MDL have shown to be effective in many computer vision applications [12], [14], [15], [18], [19], [22], [26], [39], but other criteria may be better in other applications.

10 Conclusions

We have investigated the meaning of "statistical methods" for geometric inference based on image feature points. Tracing back the origin of feature uncertainty to image processing operations, we discussed the implications of asymptotic analysis in reference to "geometric fitting" and "geometric model selection." Then, we derived the "geometric AIC" and the "geometric MDL" in this new light. We showed by experiments that the two criteria have contrasting characteristics for degeneracy detection.

The main emphasis of this paper is on the correspondence between the asymptotic analysis for $\varepsilon \to 0$ for geometric inference and the asymptotic analysis for $n \to \infty$ for the standard statistical analysis, based on our interpretation of the uncertainty of feature detection.

However, there are many issues yet to be resolved, in particular, the choice of the scale length L for the geometric MDL and the effect of using the estimate $\hat{\varepsilon}$ given by (45) for its true value ε. The results in this paper are only a first attempt, and further analysis is expected in the future.

Acknowledgments

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

References

[1] H. Akaike, "A New Look at the Statistical Model Identification," *IEEE Trans. Automatic Control,* vol. 16, no. 6, pp. 716-723, 1974.

[2] K. Bubna and C.V. Stewart, "Model Selection Techniques and Merging Rules for Range Data Segmentation Algorithms," *Computer Vision Image Understanding,* vol. 80, no. 2, pp. 215-245, 2000.

[3] F. Chabat, G.Z. Yang, and D.M. Hansell, "A Corner Orientation Detector," *Image Vision Computer,* vol. 17, no. 10, pp. 761-769, 1999.

[4] K. Cho and P. Meer, "Image Segmentation Form Consensus Information," *Computer Vision Image Understanding,* vol. 68, no. 1, pp. 72-89, 1997.

[5] K. Cho, P. Meer, and J. Cabrera, "Performance Assessment through Bootstrap," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 19, no. 11, pp. 1185-1198, Nov. 1997.

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

[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 Pattern Analysis and Machine Intelligence,* vol. 18, no. 1, pp. 58-64, Jan. 1996.

[8] C. Harris and M. Stephens, "A Combined Corner and Edge Detector, *Proc. Fourth Alvey Vision Conf.,* pp. 147-151, Aug. 1988.

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

[10] K. Kanatani, *Statistical Optimization for Geometric Computation: Theory and Practice,* Elsevier Science, 1996.

[11] K. Kanatani, "Geometric Information Criterion for Model Selection," *Int'l J. Computer Vision,* vol. 26, no. 3, pp. 171-189, 1998.

[12] K. Kanatani, "Motion Segmentation by Subspace Separation and Model Selection," *Proc. Eighth Int'l Conf. Computer Vision,* vol. 2, pp. 301-306, July 2001.

[13] K. Kanatani, "Model Selection for Geometric Inference, Plenary Talk," *Proc. Fifth Asian Conf. Computer Vision,* vol. 1, pp. xxi-xxxii, Jan. 2002.

[14] K. Kanatani, "Evaluation and Selection of Models for Motion Segmentation," *Proc. Seventh European Conf. Computer Vision,* vol. 3, pp. 335-349, May 2002.

[15] K. Kanatani and C. Matsunaga, "Estimating the Number of Independent Motions for Multibody Motion Segmentation," *Proc. Fifth Asian Conf. Computer Vision,* vol. 1, pp. 7-12, Jan. 2002.

[16] Y. Kanazawa and K. Kanatani, "Optimal Line Fitting and Reliability Evaluation," *IEICE Trans. Information and Systems,* vol. E79-D, no. 9, pp. 1317-1322, 1996.

[17] Y. Kanazawa and K. Kanatani, "Optimal Conic Fitting and Reliability Evaluation," *IEICE Trans. Information and Systems,* vol. E79-D, no. 9, pp. 1323-1328, 1996.

[18] Y. Kanazawa and K. Kanatani, "Infinity and Planarity Test for Stereo Vision," *IEICE Trans. Information and Systems,* vol. E80-D, no. 8, pp. 774-779, 1997.

[19] Y. Kanazawa and K. Kanatani, "Stabilizing Image Mosaicing by Model Selection," *3D Structure from Images—SMILE 2000,* M. Pollefeys, L. Van Gool, A. Zisserman, and A. Fitzgibbon, eds., pp. 35-51, 2001.

[20] Y. Kanazawa and K. Kanatani, "Do We Really Have to Consider Covariance Matrices for Image Features?" *Proc. Eighth Int'l Conf. Computer Vision,* vol. 2, pp. 586-591, July 2001.

[21] Y.G. Leclerc, "Constructing Simple Stable Descriptions for Image Partitioning," *Int'l J. Computer 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," *Proc. Sixth European Conf. Computer Vision,* vol. 2, pp. 595-609, June/July 2000.

[23] B.A. Maxwell, "Segmentation and Interpretation of Multicolored Objects with Highlights," *Computer Vision Image Understanding,* vol. 77, no. 1, pp. 1-24, 2000.

[24] S.J. Maybank and P.F. Sturm, "MDL, Collineations and the Fundamental Matrix," *Proc. 10th British Machine Vision Conf.,* pp. 53-62, Sept. 1999.

[25] D.D. Morris, K. Kanatani, and T. Kanade, "Gauge Fixing for Accurate 3D Estimation," *Proc. IEEE Conf. Computer Vision Pattern Recognition,* vol. 2, pp. 343-350, Dec. 2001.

[26] N. Ohta and K. Kanatani, "Moving Object Detection from Optical Flow without Empirical Thresholds," *IEICE Trans. Information and Systems,* vol. E81-D, no. 2, pp. 243-245, 1998.

[27] D. Reisfeld, H. Wolfson, and Y. Yeshurun, "Context-Free Attentional Operators: The Generalized Symmetry Transform," *Int'l J. Computer Vision,* vol. 14, no. 2, pp. 119-130, 1995.

[28] J. Rissanen, "Universal Coding, Information, Prediction and Estimation," *IEEE Trans. Information Theory,* vol. 30, no. 4, pp. 629-636, 1984.

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

[30] J. Rissanen, "Fisher Information and Stochastic Complexity," *IEEE Trans. Information Theory,* vol. 42, no. 1, pp. 40-47, 1996.

[31] C. Schmid, R. Mohr, and C. Bauckhage, "Evaluation of Interest Point Detectors," *Int'l J. Computer Vision,* vol. 37, no. 2, pp. 151-172, 2000.

[32] S.M. Smith and J.M. Brady, "SUSAN—A New Approach to Low Level Image Processing," *Int'l J. Computer Vision,* vol. 23, no. 1, pp. 45-78, May 1997.

[33] Y. Sugaya and K. Kanatani, "Outlier Removal for Feature Tracking by Subspace Separation," *IEICE Trans. Information and Systems,* vol. E86-D, no. 6, pp. 1095-1102, June 2003.

[34] P.H.S. Torr, "An Assessment of Information Criteria for Motion Model Selection," *Proc. IEEE Conf. Computer Vision Pattern Recognition,* pp. 47-53, June 1997.

[35] P.H.S. Torr, "Geometric Motion Segmentation and Model Selection," *Philosophical Trans. Royal Soc. A,* vol. 356, pp. 1321-1340, 1998.

[36] P.H.S. Torr, "Bayesian Model Estimation and Selection for Epipolar Geometry and Generic Manifold Fitting," *Int'l J. Computer Vision,* vol. 50, no. 1, pp. 35-61, 2002.

[37] P.H.S. Torr, A. FitzGibbon, and A. Zisserman, "Maintaining Multiple Motion Model Hypotheses through Many Views to Recover Matching and Structure," *Proc. Sixth Int'l Conf. Computer Vision,* pp. 485-492, Jan. 1998.

[38] P.H.S. Torr and A. Zisserman, "Concerning Bayesian Motion Segmentation, Model Averaging, Matching and the Trifocal Tensor," *Proc. Sixth European Conf. Computer Vision,* vol. 1, pp. 511-528, June/July 2000.

[39] I. Triono, N. Ohta, and K. Kanatani, "Automatic Recognition of Regular Figures by Geometric AIC," *IEICE Trans. Information and Systems,* vol. E81-D, no. 2, pp. 246-248, 1998.

Kenichi Kanatani received the MS and PhD degrees in applied mathematics from the University of Tokyo in 1974 and 1979, respectively. After serving as professor of computer science at Gunma University, Japan, he became a professor of information technology at Okayama University, Japan. He is the author of *Group-Theoretical Methods in Image Understanding* (Springer, 1990), *Geometric Computation for Machine Vision* (Oxford, 1993), and *Statistical Optimization for Geometric Computation: Theory and Practice* (Elsevier, 1996). He is a fellow of the IEEE.