4

Model Selection Criteria for Geometric Inference

K. Kanatani

ABSTRACT Model selection criteria are designed for selecting an appropriate hypothesis for the phenomenon in question. In traditional statistics, a hypothesis has a parametric expression that combines a deterministic input-output relationship and random fluctuations. However, most of the problems encountered in computer vision do not fit in this framework. In this chapter, we illustrate this difference by taking line fitting as a typical example. First, we discuss the classical regression problem and show how the Akaike Information Criterion (AIC) can be used for model selection. Then, we go on to the geometric fitting problem described in the form that typically appears in computer vision applications. Since the two problems are different, we must modify the AIC; we call the resulting criterion the "geometric AIC." We generalize this idea to an abstract framework and compare it with other criteria such as cross-validation, jackknife, bootstrap, C_P, Bayesian Information Criterion (BIC), and Minimum Description Length (MDL). We conclude by discussing some of the fundamental issues that lie behind all these criteria.

4.1 Introduction

One of the basic procedures of statistical estimation is *model* (or *parametric*) *fitting*. Given noisy data, we estimate the parameters of an assumed functional relationship, called the *model*, of the phenomenon in question. Intensive studies have been done on this subject in the past, and today we can evaluate the theoretical accuracy bound for this type of problem and avail ourselves of various estimation schemes, among which the *maximum likelihood estimation* (*MLE*) is the best known. We also have various means of predicting the accuracy performance of such techniques.

However, all these results are applicable only after a model to be fitted is given. But how can we decide upon which model to fit? In particular, how can we select an appropriate model among a number of possible candidates? In order to decide this, we need a criterion for the "goodness of a model." Evidently, the *residual*, which measures the goodness of the computed fit, cannot serve this purpose, since we can always devise a model that has a sufficient number of parameters to satisfy all the data. Evidently, such a

model is not what we want. What do we want then?

This question was first answered by Akaike [3, 5], who introduced a model selection criterion called *Akaike information criterion* (AIC). Since then similar criteria have been proposed one after another, including *Bayesian information criterion* (BIC) of Schwarz [193] and *minimum description length* (MDL) of Rissanen [175, 176]. Borrowing Akaike's idea, Kanatani tailored it so that it can be applied to a wide range of computer vision problems [111, 113, 114, 116, 165, 230] and called it the *geometric AIC* [112, 115].

However, there is something ambiguous about all the model selection criteria proposed in the past—in particular, their seeming arbitrariness. Which is the best criterion among them? In order to decide this, we need a criterion for the "goodness of a criterion". AIC can be interpreted to be an approximation to the entropy principle (the *Kullback–Leibler information* [125]), but reducing the AIC to the entropy principle does not justify the AIC. After all, how can we justify the entropy principle? Should we reduce it to a more abstract criterion? Or should we introduce a criterion for the "goodness of a criterion for a criterion"?

An inevitable course of events is the Darwinian one: Let people introduce criteria freely, and leave their survival to the "natural selection." This is what has actually taken place in statistics, but it has turned out that almost all criteria ever proposed have survived. The reason for this is simple: All the criteria are somewhat similar, and not a great deal of *practical* differences exist among them.

Today, existing criteria for model selection can be roughly classified into three approaches: *Bayesian*, *non-Bayesian*, and *empirical* (the empirical approach is inherently non-Bayesian). Typical Bayesian criteria are BIC and MDL (see [37, 38, 150] for other criteria), which are known to have nice asymptotic properties for linear regression models [15, 249]. This does not mean that their use in real application is superior to others. Non-Bayesian criteria include AIC and Mallows' C_P [144]. Empirical criteria include *cross-validation*, *jackknife*, and *bootstrap* [55]. A lot of theoretical studies have been done about equivalence relations and interrelationships among them [74, 198, 211], but it has turned out that all behave more or less similarly (and particularly so asymptotically).

In a sense, this is reassuring. At first sight, it appears that one may obtain arbitrarily many criteria by juggling with equations, but after all one can only get more or less similar results. This is analogous to the question: Which distance is the best measure for the separation of two pixels? The Euclidean distance, the city block distance, or the chessboard distance? The answer is that any of them produces almost the same result, except for slight differences depending on the particular application.

The conclusion is that one may pick out the criterion that is the easiest to formulate or compute for a given problem *provided that one is well aware of what one is doing*. There may exist small differences caused by

different choices for the criterion, and a comparative case study may make a good thesis, but such a consideration is only academic. In the following, we mainly focus on the geometric AIC because it is simple to compute and very effective in a wide range of computer vision applications [218–220, 227].

4.2 Classical Regression

4.2.1 Residual of Line Fitting

Suppose we are given N points $(x_1, y_1), \ldots, (x_N, y_N)$ and we want to know whether they are collinear or not. To be precise, we want to test if $\{y_\alpha\}$ can be regarded as samples from independent Gaussian random variables of identical variance with means $\bar{y}_\alpha$ such that

$$\bar{y}_\alpha = A + Bx_\alpha, \qquad \alpha = 1, ..., N \tag{4.1}$$

for *some* (A, B). If we know the noise variance, all we need to do is fit a line to $\{(x_\alpha, y_\alpha)\}$ optimally and test if the discrepancy is compatible with the noise magnitude. This procedure can be formalized as *testing of hypotheses* as follows. We hypothesize that

$$y_\alpha = A + Bx_\alpha + \epsilon_\alpha, \qquad \alpha = 1, ..., N \tag{4.2}$$

for independent Gaussian random variables $\{\epsilon_\alpha\}$ of mean zero and variance σ^2. A hypothesis like this is called a (*statistical*) *model* of the data.

Under this model, an optimal fit is obtained by *maximum likelihood estimation* (*MLE*). We minimize

$$J = \sum_{\alpha=1}^{N} (y_\alpha - A - Bx_\alpha)^2. \tag{4.3}$$

Let $(\hat{A}, \hat{B})$ be the resulting estimate, which is called the *maximum likelihood estimator* (*ML-estimator*). The minimum value of the function J is then

$$\hat{J} = \sum_{\alpha=1}^{N} (y_\alpha - \hat{A} - \hat{B}x_\alpha)^2, \tag{4.4}$$

which is called the *residual*. The probability distribution of $\hat{J}$ can be computed as follows. Writing

$$\hat{A} = A + \Delta\hat{A}, \qquad \hat{B} = B + \Delta\hat{B}, \tag{4.5}$$

and substituting these together with equation (4.2) into equation (4.4), we obtain

$$\hat{J} = \sum_{\alpha=1}^{N} \epsilon_\alpha^2 - 2\sum_{\alpha=1}^{N} \epsilon_\alpha(\Delta\hat{A} + x_\alpha\Delta\hat{B}) + \sum_{\alpha=1}^{N} (\Delta\hat{A} + x_\alpha\Delta\hat{B})^2. \tag{4.6}$$

Since the ML-estimator $(\hat{A}, \hat{B})$ is determined so as to minimize $\hat{J}$, we have

$$\frac{\partial \hat{J}}{\partial \Delta \hat{A}} = 0, \qquad \frac{\partial \hat{J}}{\partial \Delta \hat{B}} = 0. \tag{4.7}$$

These two equations are written in the following form:

$$\begin{pmatrix} N & \sum_{\alpha=1}^{N} x_\alpha \\ \sum_{\alpha=1}^{N} x_\alpha & \sum_{\alpha=1}^{N} x_\alpha^2 \end{pmatrix} \begin{pmatrix} \Delta\hat{A} \\ \Delta\hat{B} \end{pmatrix} = \begin{pmatrix} \sum_{\alpha=1}^{N} \epsilon_\alpha \\ \sum_{\alpha=1}^{N} \epsilon_\alpha x_\alpha \end{pmatrix}. \tag{4.8}$$

Using this, we can rewrite equation (4.6) in the form

$$\hat{J} = \sum_{\alpha=1}^{N} \epsilon_\alpha^2 - \sigma^2 \begin{pmatrix} \Delta\hat{A} \\ \Delta\hat{B} \end{pmatrix}^\top \boldsymbol{J} \begin{pmatrix} \Delta\hat{A} \\ \Delta\hat{B} \end{pmatrix}, \tag{4.9}$$

where $\boldsymbol{J}$ is the following *Fisher information matrix*

$$\boldsymbol{J} = \frac{1}{\sigma^2} \begin{pmatrix} N & \sum_{\alpha=1}^{N} x_\alpha \\ \sum_{\alpha=1}^{N} x_\alpha & \sum_{\alpha=1}^{N} x_\alpha^2 \end{pmatrix}. \tag{4.10}$$

From equation (4.8), we see that $(\Delta\hat{A}, \Delta\hat{B})$ are *linearly* related to the noise $\{\epsilon_\alpha\}$. Hence, they are Gaussian random variables. In particular, $E[\Delta\hat{A}] = E[\Delta\hat{B}] = 0$. It is easily seen from equation (4.8) that the covariance matrix

$$V[\hat{A}, \hat{B}] = E[\begin{pmatrix} \Delta\hat{A}^2 & \Delta\hat{A}\Delta\hat{B} \\ \Delta\hat{B}\Delta\hat{A} & \Delta\hat{B}^2 \end{pmatrix}] \tag{4.11}$$

has the following form:

$$V[\hat{A}, \hat{B}] = \boldsymbol{J}^{-1}. \tag{4.12}$$

This corresponds to the well known fact that the ML-estimator for linear regression under Gaussian noise attains the *Cramer–Rao lower bound* given by $\boldsymbol{J}^{-1}$ [112]. Now equation (4.9) can be written as

$$\hat{J} = \sum_{\alpha=1}^{N} \epsilon_\alpha^2 - \sigma^2 \begin{pmatrix} \Delta\hat{A} \\ \Delta\hat{B} \end{pmatrix}^\top V[\hat{A}, \hat{B}]^{-1} \begin{pmatrix} \Delta\hat{A} \\ \Delta\hat{B} \end{pmatrix}. \tag{4.13}$$

Since $(\Delta\hat{A}, \Delta\hat{B})^\top$ is a two-dimensional Gaussian random variable of mean $\mathbf{0}$ and covariance matrix $V[\hat{A}, \hat{B}]$, the quadratic form

$$(\Delta\hat{A}, \Delta\hat{B}) V[\hat{A}, \hat{B}]^{-1} (\Delta\hat{A}, \Delta\hat{B})^\top$$

is subject to a χ^2 distribution with two degrees of freedom, which we abbreviate as

$$\begin{pmatrix} \Delta\hat{A} \\ \Delta\hat{B} \end{pmatrix}^\top V[\hat{A}, \hat{B}]^{-1} \begin{pmatrix} \Delta\hat{A} \\ \Delta\hat{B} \end{pmatrix} \sim \chi_2^2. \tag{4.14}$$

Hence, it has expectation 2. It follows that

$$E[\hat{J}] = (N-2)\sigma^2. \tag{4.15}$$

Moreover, it can be shown that $\hat{J}/\sigma^2 \sim \chi^2_{N-2}$ [112]. Hence, we can reject the hypothesis with *significance level* $a\%$ (or *confidence level* $(100-a)\%$) if

$$\frac{\hat{J}}{\sigma^2} > \chi^2_{a;N-2}, \tag{4.16}$$

where $\chi^2_{a;r}$ is the upper $a\%$ percentile of a χ^2 distribution with r degrees of freedom. The interpretation is as follows. Equation (4.16) holds with probability $a/100$ if the hypothesis is correct. If a is very small, we are observing a very rare event, so it makes sense to question the hypothesis.

4.2.2 Comparison of Models

The above χ^2 test has a serious drawback, namely, determining how to set the significance level a? For a very large value of a the hypothesis is always accepted, while for a very small value of a the hypothesis is always rejected. It appears that this ambiguity can be removed by introducing an *alternative hypothesis*. We decide which hypothesis is more likely than the other. But this entails a more serious problem. Suppose we question whether the observed points $\{(x_\alpha, y_\alpha)\}$ are perturbed positions from a line or from a quadratic curve. In other words, we have the following two models:

$$\text{Model } \mathcal{M}_1: \quad y_\alpha = A + Bx_\alpha + \epsilon_\alpha. \tag{4.17}$$

$$\text{Model } \mathcal{M}_2: \quad y_\alpha = A + Bx_\alpha + Cx_\alpha^2 + \epsilon_\alpha. \tag{4.18}$$

The conclusion is evident: Model $\mathcal{M}_2$ is always favored since a quadratic curve always fits better than a line with a smaller residual.

Thus, the residual is not a good measure for validating a model. In fact, if we are given N points $\{(x_\alpha, y_\alpha)\}$, $\alpha = 1, \ldots, N$, an $(N-1)$th degree polynomial fits them with zero residual. Such an overfit is not desirable because a high degree polynomial fits to the *particular noise* rather than the *true positions* of the observed points; if noise occurred differently, we would obtain a completely different fit.

One way to resolve this problem is *cross-validation*:

1. Divide the data into two sets.
2. Fit a line or a curve to one data set (the *learning set*).
3. Evaluate the sum of squared deviation of the resulting fit from the *other* data set (the *validation set*).

A high degree polynomial may fit very well to the learning set, but should result in a large deviation for the validation set. We need many data to obtain a stable fit, while we also need many data to validate the model accurately. Various techniques have been proposed to deal with a limited number of data. For example:

Jackknife: Remove *one* datum and validate the resulting fit by that removed datum. Repeat by using different validation data.

Bootstrap: Generate the validation set via computer simulation, estimating the mechanism according to which the data have been produced.

AIC can be viewed as *hypothetical cross-validation*; it has the advantage that an *analytical* expression can be obtained for comparing models.

4.2.3 Expected Residual

Consider equation (4.4) again. The reason the residual $\hat{J}$ is not a good measure for model validation is that the ML-estimator $(\hat{A}, \hat{B})$ is determined so as to minimize $\hat{J}$, and hence it is unduly small. In order to validate the model fairly, we need a validation set that does *not* depend on $(\hat{A}, \hat{B})$. The basic underlying idea for the AIC is to introduce a *hypothetical validation set* $\{y^*_\alpha\}$ defined by

$$y^*_\alpha = A + Bx_\alpha + \epsilon^*_\alpha, \qquad \alpha = 1, ..., N \tag{4.19}$$

where $\{\epsilon^*_\alpha\}$ are Gaussian random variables of mean 0 and variance σ^2 *independent of* $\{\epsilon_\alpha\}$. The values $\{y^*_\alpha\}$ are interpreted to be the positions *we would observe if noise occurred differently*. We may also think of them as the positions we may observe if the experiment is repeated again in a different environment. In this sense, we call $\{y^*_\alpha\}$ the *future data*.

Let $(\hat{A}, \hat{B})$ be the ML-estimator obtained from the current data $\{y_\alpha\}$; the fitted line is $y = \hat{A}x + \hat{B}$. It makes sense to validate this fit against the residual *with respect to the future data* $\{y^*_\alpha\}$

$$\hat{J}^* = \sum_{\alpha=1}^{N} (y^*_\alpha - \hat{A} - \hat{B}x_\alpha)^2. \tag{4.20}$$

Since the future data $\{y^*_\alpha\}$ are hypothetical, we cannot evaluate this. However, we can evaluate its *expectation*. Substituting eqs. (4.5) into equation (4.20) and doing the same manipulations used for deriving equation (4.6), we obtain

$$\hat{J}^* = \sum_{\alpha=1}^{N} (\epsilon^*_\alpha)^2 - 2\sum_{\alpha=1}^{N} \epsilon^*_\alpha (\Delta\hat{A} + x_\alpha \Delta\hat{B}) + \sigma^2 \begin{pmatrix} \Delta\hat{A} \\ \Delta\hat{B} \end{pmatrix}^\top \boldsymbol{J} \begin{pmatrix} \Delta\hat{A} \\ \Delta\hat{B} \end{pmatrix}. \tag{4.21}$$

As we have shown earlier, we have

$$\begin{pmatrix} \Delta\hat{A} \\ \Delta\hat{B} \end{pmatrix}^\top \boldsymbol{J} \begin{pmatrix} \Delta\hat{A} \\ \Delta\hat{B} \end{pmatrix} = \begin{pmatrix} \Delta\hat{A} \\ \Delta\hat{B} \end{pmatrix}^\top V[\hat{A},\hat{B}]^{-1} \begin{pmatrix} \Delta\hat{A} \\ \Delta\hat{B} \end{pmatrix} \sim \chi^2_2, \quad (4.22)$$

and hence it has expectation 2. It follows that

$$E[\hat{J}^*] = (N+2)\sigma^2. \quad (4.23)$$

Comparing this with equation (4.15), we observe that *the residual $\hat{J}$ is smaller than its expected value J^* by $4\sigma^2$ in expectation.* This suggests that we can compensate for this decrease by defining

$$AIC = \hat{J} + 4\sigma^2. \quad (4.24)$$

4.2.4 Model Selection

Let us return to the comparison of the models $\mathcal{M}_1$ and $\mathcal{M}_2$ given by eqs. (4.17) and (4.18). Let $\hat{J}_1$ and $\hat{J}_2$ be the residuals for the models $\mathcal{M}_1$ and $\mathcal{M}_2$, respectively. We can analyze the statistical behavior of $\hat{J}_2$ and its expected value $\hat{J}_2^*$ for the future data just as we did for $\hat{J}_1$. The only difference is that the Fisher information matrix for model $\mathcal{M}_2$ is given by

$$\boldsymbol{J}_2 = \frac{1}{\sigma^2} \begin{pmatrix} N & \sum_{\alpha=1}^N x_\alpha & \sum_{\alpha=1}^N x_\alpha^2 \\ \sum_{\alpha=1}^N x_\alpha & \sum_{\alpha=1}^N x_\alpha^2 & \sum_{\alpha=1}^N x_\alpha^3 \\ \sum_{\alpha=1}^N x_\alpha^2 & \sum_{\alpha=1}^N x_\alpha^3 & \sum_{\alpha=1}^N x_\alpha^4 \end{pmatrix}. \quad (4.25)$$

Hence, we have

$$\begin{pmatrix} \Delta\hat{A} \\ \Delta\hat{B} \\ \Delta\hat{C} \end{pmatrix}^\top \boldsymbol{J}_2 \begin{pmatrix} \Delta\hat{A} \\ \Delta\hat{B} \\ \Delta\hat{C} \end{pmatrix} = \begin{pmatrix} \Delta\hat{A} \\ \Delta\hat{B} \\ \Delta\hat{C} \end{pmatrix}^\top V[\hat{A},\hat{B},\hat{C}]^{-1} \begin{pmatrix} \Delta\hat{A} \\ \Delta\hat{B} \\ \Delta\hat{C} \end{pmatrix} \sim \chi^2_3. \quad (4.26)$$

It follows that

$$E[\hat{J}_2] = (N-3)\sigma^2, \qquad E[\hat{J}_2^*] = (N+3)\sigma^2. \quad (4.27)$$

So we define AIC for these two models by

$$AIC_1 = \hat{J}_1 + 4\sigma^2, \qquad AIC_2 = \hat{J}_2 + 6\sigma^2, \quad (4.28)$$

and we favor the model $\mathcal{M}_2$ if $AIC_2 < AIC_1$. In terms of the residuals, this condition takes the form

$$\hat{J}_2 < \hat{J}_1 - 2\sigma^2. \quad (4.29)$$

In other words, a quadratic curve is not chosen unless the residual is smaller by more than $2\sigma^2$.

4.2.5 Noise Estimation

So far we have assumed that the noise variance σ^2 is known. But it is very difficult to predict it a priori in real circumstances. In such a case, we need to estimate it a posteriori. For this, we must distinguish the following two situations:

Model dependent noise The source of noise is not known. *Deviations from an assumed model are defined to be noise.*

Model independent noise The source of noise is known (for example, digitization of the image, edge detection operation, etc.). In other words, *noise is independent of our interpretation of the data.*

Traditional statistical estimation (in economics, politics, medicine, biology, pharmacology, agriculture, etc.) deals with the first situation, but the latter is the case in image analysis, computer vision, and robotics applications.

Here, we adopt the latter view and observe the crucial fact — model $\mathcal{M}_1$ is *included* in model $\mathcal{M}_2$. In other words, model $\mathcal{M}_1$ is a *degenerate* case of model $\mathcal{M}_2$ with $C = 0$. Hence, we can assume that model $\mathcal{M}_2$ is true and question if we can safely assume $C = 0$. Since the general model $\mathcal{M}_2$ holds irrespective of whether $\mathcal{M}_1$ is correct or not, we can estimate σ^2 from $\mathcal{M}_2$. From the first of eqs. (4.27), we have the following unbiased estimator of σ^2:

$$\hat{\sigma}^2 = \frac{\hat{J}_2}{N-3}. \tag{4.30}$$

Using this, we observe that the condition $AIC_1 < AIC_2$ is equivalent to

$$\frac{\hat{J}_1}{\hat{J}_2} < 1 + \frac{2}{N-3}. \tag{4.31}$$

In other words, a quadratic curve can be replaced by a line if $\hat{J}_1$ is smaller than $1 + 2/(N-3)$ times $\hat{J}_2$, which is a milder condition than $\hat{J}_1 < \hat{J}_2$.

4.2.6 Generalization

The above result can easily be extended to a general nonlinear model

$$y_\alpha = f(x_\alpha; \boldsymbol{u}) + \epsilon_\alpha, \tag{4.32}$$

where $\boldsymbol{u}$ is an n-dimensional parameter vector. Let $\hat{\boldsymbol{u}}$ be its ML-estimator. The residual is

$$\hat{J} = \sum_{\alpha=1}^{N} \big(y_\alpha - f(x_\alpha; \hat{\boldsymbol{u}})\big)^2. \tag{4.33}$$

Let us write

$$\hat{\boldsymbol{u}} = \bar{\boldsymbol{u}} + \Delta\hat{\boldsymbol{u}} \tag{4.34}$$

where $\bar{\boldsymbol{u}}$ is the true value of $\boldsymbol{u}$. If we assume that the noise variance σ^2 is small, the deviation $\Delta\hat{\boldsymbol{u}}$ is also small. Substituting equation (4.34) into equation (4.33), expanding it with respect to $\Delta\hat{\boldsymbol{u}}$, and ignoring higher order terms, we can express the residual in the form

$$\hat{J} = \sum_{\alpha=1}^{N} \epsilon_\alpha^2 - \sigma^2 \Delta\hat{\boldsymbol{u}}^\top \boldsymbol{J} \Delta\hat{\boldsymbol{u}} + O(\sigma^3) \tag{4.35}$$

where $\boldsymbol{J}$ is the following Fisher information matrix:

$$\boldsymbol{J} = \frac{1}{\sigma^2} \sum_{\alpha=1}^{N} (\nabla_{\mathbf{u}} \bar{f}_\alpha)(\nabla_{\mathbf{u}} \bar{f}_\alpha)^\top. \tag{4.36}$$

Here, $\nabla_{\mathbf{u}} \bar{f}_\alpha$ is the vector whose ith component is $\partial f(x_\alpha; \boldsymbol{u})/\partial u_i|_{\mathbf{u}=\bar{\mathbf{u}}}$. It can be shown that the covariance matrix of the ML-estimator $\hat{\boldsymbol{u}}$ has the following form [112]:

$$V[\hat{\boldsymbol{u}}] = \boldsymbol{J}^{-1} + O(\sigma^4). \tag{4.37}$$

This states that the ML-estimator attains the Cramer–Rao lower bound given by $\boldsymbol{J}^{-1}$ in the first order [112]. Since

$$\hat{\boldsymbol{u}}^\top V[\hat{\boldsymbol{u}}]^{-1} \hat{\boldsymbol{u}} \sim \chi_n^2, \tag{4.38}$$

we conclude that

$$E[\hat{J}] = (N - n)\sigma^2 + O(\sigma^4), \tag{4.39}$$

$$E[\hat{J}^*] = (N + n)\sigma^2 + O(\sigma^4). \tag{4.40}$$

Ignoring higher order terms, we define the AIC by

$$AIC = \hat{J} + 2n\sigma^2. \tag{4.41}$$

Thus, the AIC equals the residual plus $2\sigma^2$ times the *number of parameters*.

In the above argument, we have ignored higher order terms. They can be ignored if the noise magnitude is small, but it can be shown that they can be ignored even if the noise has finite magnitude, provided that the number of data is large, that is, in the *asymptotic limit* $N \to \infty$ [112].

Let $\mathcal{M}_2$ with n_2 parameters be a general model which should always hold, and let $\mathcal{M}_1$ with n_1 ($< n_2$) parameters be a degenerate model obtained by imposing additional constraint to $\mathcal{M}_2$. We write

$$\mathcal{M}_1 \succ \mathcal{M}_2, \tag{4.42}$$

and say that model $\mathcal{M}_1$ is *stronger* than model $\mathcal{M}_2$ or model $\mathcal{M}_2$ is *weaker* than model $\mathcal{M}_1$. The noise variance σ^2 can be estimated from the weaker model $\mathcal{M}_2$. From equation (4.39), we have the following unbiased estimator of σ^2:

$$\hat{\sigma}^2 = \frac{\hat{J}_2}{N - n_2}. \tag{4.43}$$

Using this, we observe that the stronger model $\mathcal{M}_1$ is favored if its residual $\hat{J}_1$ satisfies

$$\frac{\hat{J}_1}{\hat{J}_2} < 1 + \frac{2(n_2 - n_1)}{N - n_2}. \tag{4.44}$$

4.3 Geometric Line Fitting

4.3.1 Residual Analysis

Now consider the problem of fitting a line or a curve to N given points $(x_1, y_1), \ldots, (x_N, y_N)$. This time, we regard each point (x_α, y_α) as a *geometric entity*, not a *response* (or *observation*) y_α for *input* (or *control*) x_α. Suppose (x_α, y_α) is perturbed from its true position $(\bar{x}_\alpha, \bar{y}_\alpha)$, which we do not know, and write

$$x_\alpha = \bar{x}_\alpha + \Delta x_\alpha, \qquad y_\alpha = \bar{y}_\alpha + \Delta y_\alpha. \tag{4.45}$$

We regard Δx_α and Δy_α as independent Gaussian random variables of mean zero and unknown variance σ^2.

Our task is to detect the geometric structure or *geometric model* of the unknown true positions $\{(\bar{x}_\alpha, \bar{y}_\alpha)\}$. For example, we are interested to see if

$$A\bar{x}_\alpha + B\bar{y}_\alpha + C = 0, \qquad \alpha = 1, ..., N, \tag{4.46}$$

for *some* (A, B, C). In order to remove the scale indeterminacy, we normalize $\boldsymbol{n} = (A, B, C)^\top$ to a unit vector. Suppose we fit a line to the observed positions $\{(x_\alpha, y_\alpha)\}$. An optimal solution is obtained by MLE: We minimize

$$J = \sum_{\alpha=1}^{N} \Big((x_\alpha - \bar{x}_\alpha)^2 + (y_\alpha - \bar{y}_\alpha)^2 \Big) \tag{4.47}$$

with respect to $\{(\bar{x}_\alpha, \bar{y}_\alpha)\}$ *subject to the constraint* (4.46). The parameter vector $\boldsymbol{n} = (A, B, C)^\top$ is determined so that the resulting residual is minimum.

Let $\hat{A}x + \hat{B}y + \hat{C} = 0$ be the resulting fit. The positions $(\hat{x}_\alpha, \hat{y}_\alpha)$ that minimize equation (4.47) such that $\hat{A}\hat{x}_\alpha + \hat{B}\hat{y}_\alpha + \hat{C} = 0$ are given by

$$\hat{x}_\alpha = x_\alpha - \frac{\hat{A}(\hat{A}x_\alpha + \hat{B}y_\alpha + \hat{C})}{\hat{A}^2 + \hat{B}^2}, \qquad \hat{y}_\alpha = y_\alpha - \frac{\hat{B}(\hat{A}x_\alpha + \hat{B}y_\alpha + \hat{C})}{\hat{A}^2 + \hat{B}^2}. \tag{4.48}$$

Geometrically, $(\hat{x}_\alpha, \hat{y}_\alpha)$ is the foot of the perpendicular line drawn from (x_α, y_α) onto the fitted line $\hat{A}x + \hat{B}y + \hat{C} = 0$. The corresponding residual is

$$\hat{J} = \sum_{\alpha=1}^{N} \Big((x_\alpha - \hat{x}_\alpha)^2 + (y_\alpha - \hat{y}_\alpha)^2 \Big). \tag{4.49}$$

This residual is not a good measure for model selection because the fitted line is determined so that the residual is minimized. A fair measure is the residual with respect to the *future data* $\{(x^*_\alpha, y^*_\alpha)\}$. Future data are independent of $\{(x_\alpha, y_\alpha)\}$, but have the same probability distribution

$$\hat{J}^* = \sum_{\alpha=1}^{N} \Big((x^*_\alpha - \hat{x}_\alpha)^2 + (y^*_\alpha - \hat{y}_\alpha)^2 \Big). \tag{4.50}$$

Substituting eqs. (4.48) into equation (4.49), we obtain

$$\hat{J} = \frac{\sum_{\alpha=1}^{N} (\hat{A}x_\alpha + \hat{B}y_\alpha + \hat{C})^2}{\hat{A}^2 + \hat{B}^2}. \tag{4.51}$$

Using bars to denote the true values, we write

$$\hat{A} = \bar{A} + \Delta\hat{A}, \qquad \hat{B} = \bar{B} + \Delta\hat{B}, \qquad \hat{C} = \bar{C} + \Delta\hat{C}, \tag{4.52}$$

$$x_\alpha = \bar{x}_\alpha + \Delta x_\alpha, \qquad y_\alpha = \bar{y}_\alpha + \Delta y_\alpha. \tag{4.53}$$

We assume that the perturbations $(\Delta\hat{A}, \Delta\hat{B}, \Delta\hat{C})$ are linearly related to $(\Delta x_\alpha, \Delta y_\alpha)$ to a first approximation, and hence are $O(\sigma)$. Then, we can express the residual in the form

$$\begin{aligned} \hat{J} = \frac{1}{\bar{A}^2 + \bar{B}^2} \Big(& \sum_{\alpha=1}^{N} (\bar{A}^2 \Delta x_\alpha^2 + 2\bar{A}\bar{B}\Delta x_\alpha \Delta y_\alpha + \bar{B}^2 \Delta y_\alpha^2) \\ & + 2 \begin{pmatrix} \sum_{\alpha=1}^{N} (\bar{A}\Delta x_\alpha + \bar{B}\Delta y_\alpha)\bar{x}_\alpha \\ \sum_{\alpha=1}^{N} (\bar{A}\Delta x_\alpha + \bar{B}\Delta y_\alpha)\bar{y}_\alpha \\ \sum_{\alpha=1}^{N} (\bar{A}\Delta x_\alpha + \bar{B}\Delta y_\alpha) \end{pmatrix}^\top \Delta\hat{\boldsymbol{n}} \Big) \\ & - \sigma^2 \Delta\hat{\boldsymbol{n}}^\top \boldsymbol{M} \Delta\hat{\boldsymbol{n}} + O(\sigma^3) \end{aligned} \tag{4.54}$$

where $\Delta\hat{\boldsymbol{n}} = (\Delta\hat{A}, \Delta\hat{B}, \Delta\hat{C})^\top$ and

$$\boldsymbol{M} = \frac{1}{\sigma^2(\bar{A}^2 + \bar{B}^2)} \begin{pmatrix} \sum_{\alpha=1}^N \bar{x}_\alpha^2 & \sum_{\alpha=1}^N \bar{x}_\alpha \bar{y}_\alpha & \sum_{\alpha=1}^N \bar{x}_\alpha \\ \sum_{\alpha=1}^N \bar{y}_\alpha \bar{x}_\alpha & \sum_{\alpha=1}^N \bar{y}_\alpha^2 & \sum_{\alpha=1}^N \bar{y}_\alpha \\ \sum_{\alpha=1}^N \bar{x}_\alpha & \sum_{\alpha=1}^N \bar{y}_\alpha & N \end{pmatrix}. \quad (4.55)$$

Differentiating equation (4.54) with respect to $\Delta\hat{\boldsymbol{n}}$ and setting the result equal to zero, we obtain

$$\sigma^2 \boldsymbol{M} \Delta\hat{\boldsymbol{n}} = -\frac{1}{(\bar{A}^2 + \bar{B}^2)} \begin{pmatrix} \sum_{\alpha=1}^N (\bar{A}\Delta x_\alpha + \bar{B}\Delta y_\alpha)\bar{x}_\alpha \\ \sum_{\alpha=1}^N (\bar{A}\Delta x_\alpha + \bar{B}\Delta y_\alpha)\bar{y}_\alpha \\ \sum_{\alpha=1}^N (\bar{A}\Delta x_\alpha + \bar{B}\Delta y_\alpha) \end{pmatrix}. \quad (4.56)$$

Using this, we can rewrite equation (4.54) in the form

$$\hat{J} = \frac{\sum_{\alpha=1}^N (\bar{A}^2\Delta x_\alpha^2 + 2\bar{A}\bar{B}\Delta x_\alpha \Delta y_\alpha + \bar{B}^2 \Delta y_\alpha^2)}{\bar{A}^2 + \bar{B}^2} - \sigma^2 \Delta\hat{\boldsymbol{n}}^\top \boldsymbol{M} \Delta\hat{\boldsymbol{n}} + O(\sigma^3). \quad (4.57)$$

Expectation of the first term on the righthand side is

$$\frac{\sum_{\alpha=1}^N (\bar{A}^2\sigma^2 + \bar{B}^2\sigma^2)}{\bar{A}^2 + \bar{B}^2} = N\sigma^2. \quad (4.58)$$

From equation (4.56), we can evaluate the covariance matrix $V[\hat{\boldsymbol{n}}] = E[\Delta\hat{\boldsymbol{n}}\Delta\hat{\boldsymbol{n}}^\top]$ in the following form [112]:

$$V[\hat{\boldsymbol{n}}] = \boldsymbol{M}^- + O(\sigma^4). \quad (4.59)$$

Here, the superscript "$-$" denotes the (*Moore–Penrose*) *generalized inverse* [112]. Since $\hat{\boldsymbol{n}}$ is normalized to a unit vector, the three components are not independent. The generalized inverse is applied so as to keep the perturbations of $\boldsymbol{n}$ in the orientation orthogonal to $\boldsymbol{n}$. As a result, the covariance matrix $V[\hat{\boldsymbol{n}}]$ is a singular matrix of rank 2; its null space is spanned by $\boldsymbol{n}$. Equation (4.59) states the fact that the ML-estimator for this type of problem attains the *accuracy lower bound* given by $\boldsymbol{M}^{-1}$ [112].

Since $\Delta\hat{\boldsymbol{n}}^\top V[\hat{\boldsymbol{n}}]^- \Delta\hat{\boldsymbol{n}} \sim \chi_2^2$, it has expectation 2. It follows that

$$E[\hat{J}] = (N-2)\sigma^2 + O(\sigma^4). \quad (4.60)$$

4.3.2 Geometric AIC

Now we evaluate the expectation of $\hat{J}^*$. Equation (4.50) can be rewritten as

$$\begin{aligned}\hat{J}^* = &\sum_{\alpha=1}^{N}\Big((x_\alpha^* - \bar{x}_\alpha)^2 + (y_\alpha^* - \bar{y}_\alpha)^2\Big) \\ &+ 2\sum_{\alpha=1}^{N}\Big((x_\alpha^* - \bar{x}_\alpha)(\bar{x}_\alpha - \hat{x}_\alpha) + (y_\alpha^* - \bar{y}_\alpha)(\bar{y}_\alpha - \hat{y}_\alpha)\Big) \\ &+ \sum_{\alpha=1}^{N}\Big((\bar{x}_\alpha - \hat{x}_\alpha)^2 + (\bar{y}_\alpha - \hat{y}_\alpha)^2\Big). \end{aligned} \tag{4.61}$$

The expectation of the first term is $2N\sigma^2$. Since $\{(x_\alpha^*, y_\alpha^*)\}$ is independent of $\{(x_\alpha, y_\alpha)\}$, and hence of $\{(\hat{x}_\alpha, \hat{y}_\alpha)\}$, the expectation of the second term is zero. Hence,

$$E[J^*] = 2N\sigma^2 + e \tag{4.62}$$

where

$$e = E[\sum_{\alpha=1}^{N}\Big((\bar{x}_\alpha - \hat{x}_\alpha)^2 + (\bar{y}_\alpha - \hat{y}_\alpha)^2\Big). \tag{4.63}$$

Let $(\tilde{x}_\alpha, \tilde{y}_\alpha)$ be the foot of the perpendicular line drawn from $(\hat{x}_\alpha, \hat{y}_\alpha)$ onto the "true" line $\bar{A}x + \bar{B}y + \bar{C} = 0$. Just as we have eqs. (4.48) for the fitted line $\hat{A}x + \hat{B}y + \hat{C} = 0$, we have

$$\tilde{x}_\alpha = \hat{x}_\alpha - \frac{\bar{A}(\bar{A}\hat{x}_\alpha + \bar{B}\hat{y}_\alpha + \bar{C})}{\bar{A}^2 + \bar{B}^2}, \qquad \tilde{y}_\alpha = \hat{y}_\alpha - \frac{\bar{B}(\bar{A}\hat{x}_\alpha + \bar{B}\hat{y}_\alpha + \bar{C})}{\bar{A}^2 + \bar{B}^2}. \tag{4.64}$$

Since the distributions of $(\hat{x}_\alpha - \tilde{x}_\alpha, \hat{y}_\alpha - \tilde{y}_\alpha)$ and $(\tilde{x}_\alpha - \bar{x}_\alpha, \tilde{y}_\alpha - \bar{y}_\alpha)$ are uncorrelated, we have

$$e = e_1 + e_2, \tag{4.65}$$

where

$$\begin{aligned} e_1 &= E[\sum_{\alpha=1}^{N}\Big((\hat{x}_\alpha - \tilde{x}_\alpha)^2 + (\hat{y}_\alpha - \tilde{y}_\alpha)^2\Big)], \\ e_2 &= E[\sum_{\alpha=1}^{N}\Big((\tilde{x}_\alpha - \bar{x}_\alpha)^2 + (\tilde{y}_\alpha - \bar{y}_\alpha)^2\Big)]. \end{aligned} \tag{4.66}$$

The deviation of $(\tilde{x}_\alpha, \tilde{y}_\alpha)$ along the line $\bar{A}x + \bar{B}y + \bar{C} = 0$ is purely due to noise and not affected by the fitting process. Hence,

$$e_2 = N\sigma^2. \tag{4.67}$$

From eqs. (4.64), we have

$$\sum_{\alpha=1}^{N}\Big((\hat{x}_\alpha-\tilde{x}_\alpha)^2+(\hat{y}_\alpha-\tilde{y}_\alpha)^2\Big)=\sum_{\alpha=1}^{N}\frac{(\bar{A}\hat{x}_\alpha+\bar{B}\hat{y}_\alpha+\bar{C})^2}{\bar{A}^2+\bar{B}^2}. \tag{4.68}$$

Since $(\hat{x}_\alpha,\hat{x}_\alpha)$ is on the fitted line $\hat{A}x+\hat{B}y+\hat{C}=0$, we have

$$0=\hat{A}\hat{x}_\alpha+\hat{B}\hat{y}_\alpha+\hat{C}=(\bar{A}+\Delta\hat{A})\hat{x}_\alpha+(\bar{B}+\Delta\hat{B})\hat{y}_\alpha+(\bar{C}+\Delta\hat{C}). \tag{4.69}$$

Hence,

$$\bar{A}\hat{x}_\alpha+\bar{B}\hat{y}_\alpha+\bar{C}=-\Delta\hat{\boldsymbol{n}}^\top\begin{pmatrix}\hat{x}_\alpha\\ \hat{y}_\alpha\\ 1\end{pmatrix}. \tag{4.70}$$

It follows that

$$\begin{aligned}&\sum_{\alpha=1}^{N}\frac{(\bar{A}\hat{x}_\alpha+\bar{B}\hat{y}_\alpha+\bar{C})^2}{\bar{A}^2+\bar{B}^2}\\ &=\frac{1}{\bar{A}^2+\bar{B}^2}\Delta\hat{\boldsymbol{n}}^\top\begin{pmatrix}\sum_{\alpha=1}^{N}\hat{x}_\alpha^2 & \sum_{\alpha=1}^{N}\hat{x}_\alpha\hat{y}_\alpha & \sum_{\alpha=1}^{N}\hat{x}_\alpha\\ \sum_{\alpha=1}^{N}\hat{y}_\alpha\hat{x}_\alpha & \sum_{\alpha=1}^{N}\hat{y}_\alpha^2 & \sum_{\alpha=1}^{N}\hat{y}_\alpha\\ \sum_{\alpha=1}^{N}\hat{x}_\alpha & \sum_{\alpha=1}^{N}\hat{y}_\alpha & N\end{pmatrix}\Delta\hat{\boldsymbol{n}}\\ &=\sigma^2\Delta\hat{\boldsymbol{n}}^\top\boldsymbol{M}\Delta\hat{\boldsymbol{n}}+O(\sigma^3)=\sigma^2\Delta\hat{\boldsymbol{n}}^\top V[\hat{\boldsymbol{n}}]^-\Delta\hat{\boldsymbol{n}}+O(\sigma^3),\end{aligned} \tag{4.71}$$

where we have noted that $\hat{x}_\alpha=\bar{x}_\alpha+O(\sigma)$ and $\hat{y}_\alpha=\bar{y}_\alpha+O(\sigma)$ and used eqs. (4.59). Since $\Delta\hat{\boldsymbol{n}}^\top V[\hat{\boldsymbol{n}}]^-\Delta\hat{\boldsymbol{n}}\sim\chi_2^2$, it has expectation 2. If we note that $E[O(\sigma^3)]=O(\sigma^4)$, we conclude that

$$e_1=2\sigma^2+O(\sigma^4). \tag{4.72}$$

From eqs. (4.65) and (4.67), we have

$$e=(N+2)\sigma^2+O(\sigma^4). \tag{4.73}$$

From equation (4.62), we thus obtain

$$E[J^*]=(3N+2)\sigma^2+O(\sigma^4). \tag{4.74}$$

Comparing this with equation (4.60), we observe that *the residual $\hat{J}$ is smaller than its expected value $\hat{J}^*$ by $2(N+2)\sigma^2$ in the first order in expectation*. So we define the *geometric AIC* by

$$AIC=\hat{J}+2(N+2)\sigma^2. \tag{4.75}$$

4.4 General Geometric Model Selection

The above result can be generalized as follows. Let $\{\boldsymbol{a}_\alpha\}$ for $\alpha = 1, ..., N$ be N instances of an m-dimensional variable $\boldsymbol{a}$ constrained to be in an m'-dimensional manifold $\mathcal{A} \subset \mathcal{R}^m$, which we call the *data space*. We assume that each $\boldsymbol{a}_\alpha$ is a perturbed datum

$$\boldsymbol{a}_\alpha = \bar{\boldsymbol{a}}_\alpha + \Delta\boldsymbol{a}_\alpha \tag{4.76}$$

where $\Delta\boldsymbol{a}_\alpha$ is an independent Gaussian random variable of mean $\mathbf{0}$ and covariance matrix $V[\boldsymbol{a}_\alpha]$, which may be different from datum to datum and dependent on orientation. Assuming that the covariance matrix $V[\boldsymbol{a}_\alpha]$ is known only up to scale, we decompose it into an unknown *noise level* ϵ and a known *normalized covariance matrix* $V_0[\boldsymbol{a}_\alpha]$ in the form

$$V[\boldsymbol{a}_\alpha] = \epsilon^2 V_0[\boldsymbol{a}_\alpha]. \tag{4.77}$$

We are interested in validating a model that may govern the true values $\{\bar{\boldsymbol{a}}_\alpha\}$. Consider the following model:

$$F^{(k)}(\bar{\boldsymbol{a}}_\alpha, \boldsymbol{u}) = 0, \qquad k = 1, ..., L. \tag{4.78}$$

Here $\boldsymbol{u}$ is an n-dimensional vector that parameterizes the constraint, and $F^{(k)}(\boldsymbol{a}, \boldsymbol{u})$ is a smooth function of variables $\boldsymbol{a}$ and $\boldsymbol{u}$. We assume that the domain of $\boldsymbol{u}$ is not the entire n-dimensional space $\mathcal{R}^n$ but an n'-dimensional manifold $\mathcal{U} \subset \mathcal{R}^n$, which we call the *parameter space*. We also assume that the above L equations are not necessarily independent and that the essential number of constraining equations is r ($\leq L$), which we call the *rank* of the constraint. This means that the data $\{\boldsymbol{a}_\alpha\}$ are on a *manifold* $\mathcal{S}$ of *codimension* r in the m'-dimensional data space $\mathcal{A}$ (hence, the dimension of $\mathcal{S}$ is $d = m' - r$). We call $\mathcal{S}$ the *geometric model* corresponding to equation (4.78).

For convenience, we hereafter use the notation

$$\|\boldsymbol{x}\|_\alpha^2 = \boldsymbol{x}^\top V_0[\boldsymbol{a}_\alpha]^- \boldsymbol{x}. \tag{4.79}$$

MLE minimizes the sum of the squared *Mahalanobis distances*

$$J = \sum_{\alpha=1}^{N} \|\boldsymbol{a}_\alpha - \bar{\boldsymbol{a}}_\alpha\|_\alpha^2 \tag{4.80}$$

subject to the constraint (4.78) for some $\boldsymbol{u}$.

Let $\hat{\boldsymbol{u}}$ be the resulting ML-estimator of $\boldsymbol{u}$. The ML-estimator of $\bar{\boldsymbol{a}}_\alpha$ is given by

$$\hat{\boldsymbol{a}}_\alpha = \boldsymbol{a}_\alpha - \sum_{k,l=1}^{L} \hat{W}_\alpha^{(kl)} \hat{F}_\alpha^k V_0[\boldsymbol{a}_\alpha](\nabla_{\mathbf{a}} \hat{F}_\alpha^{(l)}) \tag{4.81}$$

$$(\hat{W}_\alpha^{(kl)}) = \Big((\nabla_{\mathbf{a}} \hat{F}_\alpha^{(k)})^\top V_0[\boldsymbol{a}_\alpha] (\nabla_{\mathbf{a}} \hat{F}_\alpha^{(l)}) \Big)_r^- \tag{4.82}$$

where $(\cdot)_r^-$ denotes the (Moore–Penrose) generalized inverse of rank r; it is computed by transforming the matrix into its canonical form, replacing the smallest $L-r$ eigenvalues by zeros, and transforming it back into its original frame [112]. Equation (4.82) states that $\hat{W}_\alpha^{(kl)}$ is the (kl) element of the generalized inverse of rank r of the matrix whose (kl) element is $(\nabla_{\mathbf{a}} \hat{F}_\alpha^{(k)})^\top V_0[\boldsymbol{a}_\alpha] (\nabla_{\mathbf{a}} \hat{F}_\alpha^{(l)})$, where $\hat{F}_\alpha^{(k)}$ is an abbreviation of $F^{(k)}(\boldsymbol{a}_\alpha, \hat{\boldsymbol{u}})$, and $\nabla_{\mathbf{a}}(\cdots)$ is a vector whose ith component is $\partial(\cdots)/\partial a_i$. Geometrically, equation (4.81) defines the *Mahalanobis projection* of $\boldsymbol{a}_\alpha$ onto the manifold $\mathcal{S}$ [112]. The residual $\hat{J}$ and the expected residual $\hat{J}^*$ have the form

$$\hat{J} = \sum_{\alpha=1}^{N} \|\boldsymbol{a}_\alpha - \hat{\boldsymbol{a}}_\alpha\|_\alpha^2, \tag{4.83}$$

$$\hat{J}^* = \sum_{\alpha=1}^{N} \|\boldsymbol{a}_\alpha^* - \hat{\boldsymbol{a}}_\alpha\|_\alpha^2, \tag{4.84}$$

where $\{\boldsymbol{a}_\alpha^*\}$ are the future data that have the same distribution as $\{\boldsymbol{a}_\alpha\}$ but are independent of them. It can be shown [112] that

$$E[\hat{J}] = (rN - n')\epsilon^2 + O(\epsilon^4), \tag{4.85}$$

$$E[\hat{J}^*] = ((r+2d)N + n')\epsilon^2 + O(\epsilon^4) \tag{4.86}$$

where d is the dimension of the manifold $\mathcal{S}$. Thus, $\hat{J}$ is smaller than its expected value $\hat{J}^*$ by $2(dN+n')\epsilon^2$ in the first order in expectation. So we define the geometric AIC by

$$AIC(\mathcal{S}) = \hat{J} + 2(dN + n')\epsilon^2. \tag{4.87}$$

The number $dN + n'$ can be interpreted as the *effective degree of freedom of the model*: the true values $\{\bar{\boldsymbol{a}}_\alpha\}$ can be anywhere in the d-dimensional manifold $\mathcal{S}$, so they have dN degrees of freedom; the vector $\boldsymbol{u}$ that parameterizes the manifold $\mathcal{S}$ has n' degrees of freedom.

Let $\mathcal{S}_1$ be a model obtained by imposing an additional constraint to model $\mathcal{S}_2$, that is, $\mathcal{S}_1 \succ \mathcal{S}_2$. Suppose model $\mathcal{S}_1$ has dimension d_1, codimension r_1, and n_1' degrees of freedom, and model $\mathcal{S}_2$ has dimension d_2, codimension r_2, and n_2' degrees of freedom. Let $\hat{J}_1$ and $\hat{J}_2$ be their respective residuals. The squared noise level ϵ^2 can be estimated from the weaker model $\mathcal{M}$. From equation (4.85), we obtain the following estimator of ϵ^2, which is unbiased in the first order:

$$\hat{\epsilon}^2 = \frac{\hat{J}_2}{r_2 N - n_2'}. \tag{4.88}$$

Using this, we observe that a stronger model $\mathcal{S}_1$ is favored if its residual $\hat{J}_1$ satisfies

$$\frac{\hat{J}_1}{\hat{J}_2} < 1 + \frac{2(d_2 - d_1)N + 2(n'_2 - n'_1)}{r_2 N - n'_2}. \tag{4.89}$$

4.5 Geometric C_P

We now compare the geometric AIC with Mallows' C_P [144]. Like other criteria, C_P was also defined for traditional statistical estimation (typically for linear regression), so we need some modification to ensure that it fits in the framework of geometric estimation.

For the geometric AIC, the true positions $\{\bar{\boldsymbol{a}}_\alpha\}$ of the observed data $\{\boldsymbol{a}_\alpha\}$ are assumed to satisfy the constraint given by equation (4.78). For Mallows' C_P, however, we assume that *they do not*. In other words, there exists no value of $\boldsymbol{u}$ that satisfies $F^{(k)}(\bar{\boldsymbol{a}}_\alpha, \boldsymbol{u}) = 0$, $k = 1, ..., r$. Differently put, we assume a priori that *the model in question is not correct*. Our interest is *how closely can the model approximate the reality?*

This idea is more in line with the traditional domains of statistics, where any model is an artificial and imaginary entity, which only approximates the reality, than with the domains of engineering, where usually all models are completely and exactly known—we just do not know which is out there.

Given observations $\{\boldsymbol{a}_\alpha\}$, we fit a model surface (i.e., a manifold) $\hat{\mathcal{S}}$ to them. Let $\{\hat{\boldsymbol{a}}_\alpha\}$ be their (Mahalanobis) projections onto $\hat{\mathcal{S}}$ (for simplicity, we hereafter omit "Mahalanobis"). The residual is

$$J = \sum_{\alpha=1}^{N} \|\boldsymbol{a}_\alpha - \hat{\boldsymbol{a}}_\alpha\|_\alpha^2. \tag{4.90}$$

If we are to claim that the model is good, $\hat{\boldsymbol{a}}_\alpha$ should not be far apart from its true position $\bar{\boldsymbol{a}}_\alpha$. So we measure the goodness of the model by

$$K = \sum_{\alpha=1}^{N} \|\hat{\boldsymbol{a}}_\alpha - \bar{\boldsymbol{a}}_\alpha\|_\alpha^2. \tag{4.91}$$

Although we cannot compute K because it involves unknown values $\{\bar{\boldsymbol{a}}_\alpha\}$, we can compute its expectation.

First, suppose our model $\mathcal{S}$ has no degrees of freedom; it is a fixed manifold, and hence $\hat{\mathcal{S}} = \mathcal{S}$. Let $\hat{\bar{\boldsymbol{a}}}_\alpha$ be the projection of $\bar{\boldsymbol{a}}_\alpha$ onto $\hat{\mathcal{S}}$, noting that, unlike for the geometric AIC, the true positions $\{\bar{\boldsymbol{a}}_\alpha\}$ are not on the model surface $\mathcal{S}$. The deviation $\hat{\boldsymbol{a}}_\alpha - \hat{\bar{\boldsymbol{a}}}_\alpha$ is the projection of the noise $\Delta\boldsymbol{a}_\alpha = \boldsymbol{a}_\alpha - \bar{\boldsymbol{a}}_\alpha$ onto $\hat{\mathcal{S}}$ if $\Delta\boldsymbol{a}_\alpha$ is sufficiently small, which we hereafter assume. Hence, it has zero mean. We can also see that $\|\hat{\boldsymbol{a}}_\alpha - \hat{\bar{\boldsymbol{a}}}_\alpha\|_\alpha^2/\epsilon^2 \sim \chi_d^2$. Hence,

it has expectation d. It follows that

$$E[K] = E[\sum_{\alpha=1}^{N} \|\hat{\boldsymbol{a}}_\alpha - \hat{\bar{\boldsymbol{a}}}_\alpha\|_\alpha^2] + \sum_{\alpha=1}^{N} \|\hat{\bar{\boldsymbol{a}}}_\alpha - \bar{\boldsymbol{a}}_\alpha\|_\alpha^2 = dN\epsilon^2 + \sum_{\alpha=1}^{N} \|\hat{\bar{\boldsymbol{a}}}_\alpha - \bar{\boldsymbol{a}}_\alpha\|_\alpha^2. \tag{4.92}$$

Note that $\sum_{\alpha=1}^{N} \|\hat{\bar{\boldsymbol{a}}}_\alpha - \bar{\boldsymbol{a}}_\alpha\|_\alpha^2$ is not a random variable.

Since $\hat{\boldsymbol{a}}_\alpha - \hat{\bar{\boldsymbol{a}}}_\alpha$ is the projection of $\boldsymbol{a}_\alpha - \bar{\boldsymbol{a}}_\alpha$ onto $\hat{\mathcal{S}}$, the "normal component" $\boldsymbol{a}_\alpha - \hat{\boldsymbol{a}}_\alpha$ is independent of the "tangential component" $\hat{\boldsymbol{a}}_\alpha - \hat{\bar{\boldsymbol{a}}}_\alpha$. Hence,

$$\begin{aligned} E[\sum_{\alpha=1}^{N} \|\boldsymbol{a}_\alpha - \hat{\bar{\boldsymbol{a}}}_\alpha\|_\alpha^2] &= E[\sum_{\alpha=1}^{N} \|\boldsymbol{a}_\alpha - \hat{\boldsymbol{a}}_\alpha\|_\alpha^2] + E[\sum_{\alpha=1}^{N} \|\hat{\boldsymbol{a}}_\alpha - \hat{\bar{\boldsymbol{a}}}_\alpha\|_\alpha^2] \\ &= E[J] + dN\epsilon^2 \end{aligned} \tag{4.93}$$

On the other hand, we have

$$\begin{aligned} E[\sum_{\alpha=1}^{N} \|\boldsymbol{a}_\alpha - \hat{\bar{\boldsymbol{a}}}_\alpha\|_\alpha^2] &= E[\sum_{\alpha=1}^{N} \|\boldsymbol{a}_\alpha - \bar{\boldsymbol{a}}_\alpha\|_\alpha^2] + \sum_{\alpha=1}^{N} \|\bar{\boldsymbol{a}}_\alpha - \hat{\bar{\boldsymbol{a}}}_\alpha\|_\alpha^2 \\ &= m'N\epsilon^2 + \sum_{\alpha=1}^{N} \|\bar{\boldsymbol{a}}_\alpha - \hat{\bar{\boldsymbol{a}}}_\alpha\|_\alpha^2. \end{aligned} \tag{4.94}$$

Noting that $r = m' - d$, we conclude that

$$E[J] = \sum_{\alpha=1}^{N} \|\bar{\boldsymbol{a}}_\alpha - \hat{\bar{\boldsymbol{a}}}_\alpha\|_\alpha^2 + rN\epsilon^2. \tag{4.95}$$

So far, we have ignored the degree of freedom of the model $\mathcal{S}$. Suppose it has n' degrees of freedom. Let $\hat{\mathcal{S}}$ be the surface fitted to the data $\{\boldsymbol{a}_\alpha\}$, and $\hat{\bar{\boldsymbol{a}}}_\alpha$ the projection of $\bar{\boldsymbol{a}}_\alpha$ onto it. The deviation of $\hat{\mathcal{S}}$ from $\{\bar{\boldsymbol{a}}_\alpha\}$ is measured by $\sum_{\alpha=1}^{N} \|\hat{\bar{\boldsymbol{a}}}_\alpha - \bar{\boldsymbol{a}}_\alpha\|_\alpha^2$. Now, let $\bar{\mathcal{S}}$ be the fit to the true positions $\{\bar{\boldsymbol{a}}_\alpha\}$ (recall that the model does *not* fit to the true positions exactly), and let $\bar{\bar{\boldsymbol{a}}}_\alpha$ be the projection of $\bar{\boldsymbol{a}}_\alpha$ onto $\bar{\mathcal{S}}$. The deviation of $\bar{\mathcal{S}}$ from $\{\bar{\boldsymbol{a}}_\alpha\}$ is measured by

$$\bar{J} = \sum_{\alpha=1}^{N} \|\bar{\bar{\boldsymbol{a}}}_\alpha - \bar{\boldsymbol{a}}_\alpha\|_\alpha^2, \tag{4.96}$$

which is not a random variable. It can be shown that because of the n' degrees of freedom of the model, $\bar{J}$ is smaller than $\sum_{\alpha=1}^{N} \|\hat{\bar{\boldsymbol{a}}}_\alpha - \bar{\boldsymbol{a}}_\alpha\|_\alpha^2$ by $n'\epsilon^2$ in expectation:

$$\bar{J} = E[\sum_{\alpha=1}^{N} \|\hat{\bar{\boldsymbol{a}}}_\alpha - \bar{\boldsymbol{a}}_\alpha\|_\alpha^2] - n'\epsilon^2. \tag{4.97}$$

From equation (4.92), we have

$$E[K] = \bar{J} + (dN + n')\epsilon^2. \tag{4.98}$$

On the other hand, the residual J also decreases because we adjust $\hat{\mathcal{S}}$ so that J is minimized. The amount of this decrease can be shown to be $n'\epsilon^2$ in expectation:

$$E[J] = \bar{J} + (rN - n')\epsilon^2 \tag{4.99}$$

Noting that $r = m' - d$, we conclude that

$$E[K] = E[J] + (2(dN + n') - m'N)\epsilon^2. \tag{4.100}$$

Thus, K is larger than the residual J by $(2(dN + n') - m'N)\epsilon^2$ in expectation. This observation leads to the following definition of the *geometric* C_P:

$$C_P(\mathcal{S}) = \sum_{\alpha=1}^{N} \|\boldsymbol{a}_\alpha - \hat{\boldsymbol{a}}_\alpha\|_\alpha^2 + (2(dN + n') - m'N)\epsilon^2. \tag{4.101}$$

But we see that

$$C_P(\mathcal{S}) = AIC(\mathcal{S}) - m'N\epsilon^2. \tag{4.102}$$

Since the dimension m' of the data space $\mathcal{A}$ and the number N of the data $\{\boldsymbol{a}_\alpha\}$ have nothing to do with the choice of the model $\mathcal{S}$, we observe that *the geometric C_P is equivalent to the geometric AIC*.

This is one of the many evidences of the fact that we always end up with something like the AIC as long as we manipulate residuals and model parameters. If we want a different result, we need to introduce a very different discipline, such as that of the Bayesian school.

4.6 Bayesian Approaches

4.6.1 MDL

One of the widely adopted criteria other than AIC is the *minimum description length* (MDL) of Rissanen [175, 176]. The underlying logic is as follows: Any model can explain the data to some extent if the parameters are appropriately adjusted. The MDL principle demands that the model should explain the data very well and at the same time have a *simple structure*. Rissanen's idea is to measure the simplicity (or, rather, "complexity") of the model *and* the data by *information theoretic code length*. It has turned out that the resulting MDL has a very similar structure to the AIC in one respect and it has a very distinct feature in other respects.

First, we have to code the data. The code length can be minimized by assigning a short codeword to a datum that is likely to occur while assigning a long codeword to a datum that is unlikely to occur. It turns out that the code length of the data is proportional to the negative logarithmic likelihood and hence equivalent to the residual in the case of Gaussian noise. So this part is identical to AIC.

Next, we have to code the model. In order to do this, we need to introduce the "likelihood of models," because information theoretic coding is impossible without a probabilistic structure. We imagine a set of parameterized distributions (i.e., models) and assume some kind of *a priori probability distribution,* or *prior* for short. We first code the choice of the model and then code that chosen model. It turns out that even if we adopt something like a uniform distribution of models and parameters, the code length has a very complicated form that typically involves such a term as $n \log N$, where n is the number of the parameters involved in the probability density and N is the number of data.

Thus, the MDL is essentially Bayesian. Not only does it depend on the statistical evidence given by the data, but it depends also our *belief* about the likelihood of the candidate models. Another aspect of MDL is the quantization process: The information theoretic coding can be defined only for a discrete set of symbols and strings. Since the likelihood of the data and the prior of the models are usually given as continuous probability densities, they are approximated by discrete histograms. This process complicates both the derivation and the resulting form of the MDL. Also, a philosophical question remains unanswered: Why do we need quantization for dealing with problems for which everything is continuous?

If MDL is favored, it is probably because of the following two reasons. First, MDL has been found to possess nice asymptotic properties for linear regression models [15, 249]. Of course, we cannot say anything for a finite number of data. Second, MDL can be defined for models to which it is practically impossible to introduce a statistical structure (hence, AIC cannot be defined); one simply assumes whatever prior one favors. This is regarded as a merit by many practitioners.

In view of these features, MDL suits *guessing* the structure of the problems for which the true mechanism is not known or cannot be defined. For example, it can be used for image segmentation [65, 129], but the resulting mechanism is *hypothetical,* whichever model is chosen. This is a big contrast to such a problem as structure from motion, for which we know that the epipolar constraint should necessarily hold. Thus, MDL has much more flexibility and applicability than AIC because we can freely adjust its value by introducing an appropriate prior. This flexibility often results in too much arbitrariness.

4.6.2 BIC

The *Bayesian information criterion* (BIC) of Schwarz [193] is obtained by a straightforward generalization of the *maximum a posteriori probability* (or *MAP*) estimation, a core technique of the Bayesian approach. We assume a prior for the model and combine it with the likelihood of the data observed under a particular model (the *conditional likelihood*) by means of the *Bayes formula*. BIC adopts as the prior the exponential distribution with respect to the complexity of the model; simple models are assumed to be more likely, while complicated models are assumed to be exponentially less likely.

It turns out that the BIC consists of the residual term and the penalty term, just as the AIC, but the penalty term has the form of $n \log N$ just as in MDL. Theoretically, BIC has asymptotic properties almost identical to MDL unless one assumes very unnatural priors. Compared with AIC, BIC results in slightly different decisions only in the asymptotic limit. BIC can be regarded as a hybrid of AIC and MDL; it is very similar to AIC in some respects and to MDL in others.

Thus, although the Bayesian approach introduces somewhat new criteria, what is predicted in practical situations is not much different from the non-Bayesian approach as long as one does not assume extraordinary priors.

4.7 Noise Estimation

4.7.1 Source of Noise

There are two issues about noise: One is whether or not the noise depends on the model we are assuming; the other is how to estimate its magnitude.

As was discussed earlier, noise in the traditional domains of statistics is *defined* to be the deviations from an assumed model; the source of noise is not known or defined, and noise is *part* of the model. In the domains of engineering, in contrast, noise depends on the *data acquisition process* (e.g., the resolution of the camera and the display and the accuracy of image processing techniques involved), but not on our interpretation of the image. In such domains, the noise level ϵ can be easily estimated from the residual $\hat{J}$. Since $\hat{J}/\epsilon^2 \sim \chi^2_{rN-n'}$ if the model is correct (see eqs. (4.83) and (4.85)), and since we know that the model is correct, we obtain an estimate

$$\hat{\epsilon}^2 = \frac{\hat{J}}{rN - n'}. \tag{4.103}$$

This is an *unbiased estimator* of ϵ^2, that is,

$$E[\hat{\epsilon}^2] = \epsilon^2. \tag{4.104}$$

Using $\hat{\epsilon}^2$ for ϵ^2, we can apply the geometric AIC to test if we can *strengthen* the model.

Mallows' C_P and BIC adopt the strategy of estimating the noise magnitude from a general model and then testing stronger models, because they are based on statistical analysis of the data (BIC is Bayesian *in derivation*, but once defined its use is the same as AIC). Cross-validation by jackknife or bootstrap has the advantage that we need not explicitly estimate noise; their experimental procedure implicitly incorporates noise estimation. MDL is usually coupled with the model dependent approach, because one is interested in how accurately and how succinctly one can *describe* the phenomenon rather than in doing model-based data analysis.

AIC, on the other hand, admits both approaches, because it has two different aspects: data analysis and model description. As described earlier, it can be viewed as data analysis by cross validation using future data and also as model description in terms of *relative entropy* (the Kullback–Leibler information [125]). In order to shift from the data-analysis view to the model-description view, we simply replace the residual J by $-\log$(likelihood). Then equation (4.87) is replaced by

$$AIC_L(\mathcal{S}) = \frac{J}{\epsilon^2} + m'N \log 2\pi\epsilon^2 + 2(dN + n'). \tag{4.105}$$

The righthand side can be written as $AIC(\mathcal{S})/\epsilon^2 + m'N \log 2\pi\epsilon^2$, so $AIC_L(\mathcal{S})$ and $AIC(\mathcal{S})$ are equivalent as long as ϵ^2 is regarded as a constant. But if we regard ϵ^2 as a *parameter*, it should be estimated in such a way that $AIC_L(\mathcal{S})$ is minimized. Since the model parameter term $(2dN + n')$ has nothing to do with ϵ^2, this is equivalent to MLE, and the solution is obtained in the form

$$\hat{\epsilon}^2_{MLE} = \frac{J}{m'N}. \tag{4.106}$$

Substituting this into equation (4.105), we obtain

$$AIC_L(\mathcal{S}) = m'N \left(\log J + \log \frac{2\pi}{m'N} + 1 \right) + 2(dN + n'). \tag{4.107}$$

Since the model independent terms are irrelevant, we conclude that we should compare $m'N \log J + 2(dN + n')$ for different models.

Note that equation (4.106) is different from equation (4.103). Is equation (4.106) justified? The answer is "no." In fact, since $\hat{J}/\epsilon^2 \sim \chi^2_{rN-n'}$, we have

$$E[\hat{\epsilon}^2_{MLE}] = \frac{rN - n'}{m'N}\epsilon^2, \tag{4.108}$$

which is considerably smaller than ϵ^2. In the limit $N \to \infty$, the above expression becomes equal to r/m times ϵ^2. But is MLE not optimal, as stated in the textbooks of statistics?

4.7.2 Trap of MLE

This issue is very subtle and deep. And this is one of the fundamental difference of geometric fitting as opposed to classical statistical estimation such as linear/nonlinear regression. Consider the line fitting for homogeneous isotropic noise of variance σ^2, for example. Let us estimate the variance σ^2 by MLE. The likelihood function for observation $\{(x_\alpha, y_\alpha)\}$ is given by

$$p = \prod_{\alpha=1}^{N} \frac{1}{2\pi\sigma^2} \exp\left(-\frac{(x_\alpha - \bar{x}_\alpha)^2 + (y_\alpha - \bar{y}_\alpha)^2}{2\sigma^2}\right). \tag{4.109}$$

Taking the logarithm, we obtain

$$-2\log p = 2N\log(2\pi\sigma^2) + \frac{1}{\sigma^2}\sum_{\alpha=1}^{N}\left((x_\alpha - \bar{x}_\alpha)^2 + (y_\alpha - \bar{y}_\alpha)^2\right). \tag{4.110}$$

MLE means minimizing this with respect to $\{(\bar{x}_\alpha, \bar{y}_\alpha)\}$ and σ^2 subject to the constraint $A\bar{x}_\alpha + B\bar{y}_\alpha + C = 0$. Since the first term on the righthand side of equation (4.110) depends only on σ^2, we can obtain the MLE-estimator of $(\hat{A}, \hat{B}, \hat{C})$ of (A, B, C) by minimizing the second term independently of the variance σ^2. Then, the ML-estimator of $(\bar{x}_\alpha, \bar{y}_\alpha)$ is given by eqs. (4.48). Hence, the value of σ^2 that minimizes equation (4.110) is given by

$$\hat{\sigma}^2_{MLE} = \frac{1}{2N}\sum_{\alpha=1}^{N}\left((x_\alpha - \hat{x}_\alpha)^2 + (y_\alpha - \hat{y}_\alpha)^2\right). \tag{4.111}$$

Since $\sum_{\alpha=1}^{N}\left((x_\alpha - \hat{x}_\alpha)^2 + (y_\alpha - \hat{y}_\alpha)^2\right)/\sigma^2 \sim \chi^2_{N-2}$, we have

$$E[\hat{\sigma}^2_{MLE}] = \frac{N-2}{2N}\sigma^2. \tag{4.112}$$

This means that the MLE-estimator $\hat{\sigma}^2_{MLE}$ is a very poor estimator of σ^2; it is approximately half σ^2 for a large N.

This contradicts our belief that MLE produces a good estimator, which should be asymptotically optimal as stated in the textbooks of statistics. If we carefully examine the textbooks, however, we find that this desirable property is based on the fact that *accuracy of estimation improves as the number of data increases*.

But this is *not* guaranteed for geometric fitting. Consider line fitting, for example. As the number of points increases, the fitted line $\hat{A}x + \hat{B}y + \hat{C} = 0$ *usually* approaches the true line $\bar{A}x + \bar{B}y + \bar{C} = 0$, but the projection $(\hat{x}_\alpha, \hat{y}_\alpha)$ of the data point (x_α, y_α) onto the fitted line *never* approaches its true position $(\bar{x}_\alpha, \bar{y}_\alpha)$ since the fitting process has no effect over the "tangential displacement" along the fitted line; this is the cause of the underestimation (4.112). Even the fitted line $\hat{A}x + \hat{B}y + \hat{C} = 0$ does not

approach its true position as the number of data points increases if they are located close to each other in a single cluster. Does this contradict to the statements in the textbooks of statistics?

If we carefully examine the textbooks again, we find that in statistics "number of data" actually means "number of observations": Each observation is assumed to be *an independent sample from a common probability density* that describes the phenomenon. The ML-estimator converges to its true value as the number of observations n goes to infinity. In fact, the basic strategy of statistics is to beat randomness by *repeated observations*.

The situation is completely different in geometric fitting. For line fitting, for example, the unknowns are the coefficients (A, B, C) of the line to be fitted *and* the true positions $(\bar{x}_\alpha, \bar{y}_\alpha)$ of the data (x_α, y_α). Hence, as the number N of "data points" increases, *the number of unknowns increases at the same rate*. The coefficients (A, B, C), which remain the same as the number of data points increases, are called the *structural parameters* or the *parameters of interest*, while the true positions $(\bar{x}_\alpha, \bar{y}_\alpha)$, whose number increases as the number of data points increases, are called the *nuisance parameters*.

Thus, when we observe an increased number of data points, they still represent "one" sample from a *new probability density* with an increased number of unknowns. In other words, the number of observations n is always one however large is N (the number of the data points). Thus, the asymptotic optimality of MLE is never realized for geometric fitting.

Suppose hypothetically we could repeat observations of the *same* points $(\bar{x}_\alpha, \bar{y}_\alpha)$ n times and obtain n sets of data $\{(x_\alpha^{(i)}, y_\alpha^{(i)})\}$, $i = 1, ..., n$, $\alpha = 1, ..., N$. Then the averages

$$x_\alpha = \frac{1}{n}\sum_{i=1}^{n} x_\alpha^{(i)}, \qquad y_\alpha = \frac{1}{n}\sum_{i=1}^{n} y_\alpha^{(i)}, \tag{4.113}$$

of the computed positions $\{(x_\alpha, y_\alpha)\}$ would have errors of $O(1/\sqrt{n})$ times the original errors. In other words, increasing the number of (hypothetical) observations n would *effectively reduce the noise variance* σ^2. Thus, *the optimality of MLE for geometric fitting holds in the limit of small noise*, $\sigma^2 \to 0$.

4.8 Concluding Remarks

The traditional approach in computer vision is to assume whatever constraint that might apply to the observed images and make inference by fitting the assumed constraint. If the constraint contains unknown parameters, they are estimated by least squares. The reliability of the estimation is tested by random-number simulation and subjective evaluation of real image examples for which the ground truth is not known.

However, we cannot arrive at a definitive conclusion regardless of the number of experiments we repeat, because the conclusion is always for that particular environment in which the experiments are conducted. The same conclusion may not hold in another environment. Also, there are many factors that affect the performance of the system, but we do not know which of these affects it and how. This type of knowledge can be obtained only by *analysis*.

Yet, *analysis requires assumptions*. This is the main reason why analysis is not popular, apart from its technical difficulties and mathematical complications. Since noise is a random phenomenon, we need statistical analysis, which in turn requires certain assumptions about noise. The standard technique is to assume that the noise is Gaussian, as we have done so far.

Gaussian noise is an idealization and a mathematical convenience, and real noise is not exactly Gaussian. There is, however, no essential difficulty in extending the theory to deal with non-Gaussian noise. In fact, one only needs to consider the negative logarithmic likelihood instead of the residual. Then, the generalized inverse of the Fisher information matrix, which is defined by applying differential calculus, plays the role of the covariance matrix of Gaussian noise [112]. However, such an extension does not have much practical significance because of the difficulty of estimating the parameters of a non-Gaussian noise distribution. In practice, the Gaussian noise assumption is the most realistic, unless we know something that is definitely non-Gaussian about the noise we encounter.

Of course, the Gaussian noise assumption does not lead to a *robust* system, which ideally should work invariably well (or not poorly) under *any* noise distributions. The basic approach for this is detecting *outliers* that have very different characteristics from the other data. Research in this direction is in progress [153, 209, 216], but the approach has been mostly heuristic. In order to introduce a rigorous analysis to outlier detection, we need to *model* the outliers. Customarily, inlier noise is assumed to be Gaussian but outlier noise is regarded as *anything else*, which is the major difficulty for outlier analysis. For deeper analysis, one needs to model the outlier distribution and test if that model is appropriate. This could be done by applying some kind of model selection criterion. This approach is already in progress [219, 227].

Here we have mainly focused on AIC, but this does not mean that AIC is superior to others such as BIC and MDL. As was pointed out in the Introduction (section 4.1), there exists no definite *criterion for a criterion*. Also, all existing criteria are like AIC in many respects. Hence, all we need to do is apply something like AIC to our benefit in computer vision problems, and AIC should be the first to be tested because of its simplicity.