Memoirs of the Faculty of Engineering, Okayama University, Vol.39, pp.63-78, January, 2005

Optimality of Maximum Likelihood Estimation for Geometric Fitting and the KCR Lower Bound

Kenichi KANATANI*

Department of Information Technology, Okayama University

Okayama 700-8530 Japan

(Received December 21, 2004)

Geometric fitting is one of the most fundamental problems of computer vision. In [8], the author derived a theoretical accuracy bound (KCR lower bound) for geometric fitting in general and proved that maximum likelihood (ML) estimation is statistically optimal. Recently, Chernov and Lesort [3] proved a similar result, using a weaker assumption. In this paper, we compare their formulation with the author's and describe the background of the problem. We also review recent topics including semiparametric models and discuss remaining issues.

1. What Is the Problem?

By *geometric fitting*, we mean fitting a geometric constraint to observed data and discerning the underlying geometric structure from the coefficients of the fitted equation [8]. A large class of computer vision problems fall into this framework. The simplest example is to fit a parametric curve (e.g., a line, a circle, an ellipse, or a polynomial curve) in the form

$$F(\boldsymbol{x};\boldsymbol{u}) = 0 \tag{1}$$

to N points $\{(x_\alpha, y_\alpha)\}$ in the image, where $\boldsymbol{x} = (x, y)^\top$ is the position vector, and $\boldsymbol{u} = (u_1, ..., u_p)^\top$ is the parameter vector.

For noisy data $\{(x_\alpha, y_\alpha)\}$, no parameter $\boldsymbol{u}$ satisfies $F(\boldsymbol{x}_\alpha;\boldsymbol{u}) = 0$ for all $\alpha = 1$, ..., N, so one often computes a $\boldsymbol{u}$ such that

$$J_{\rm LS} = \sum_{\alpha=1}^{N} F(\boldsymbol{x}_\alpha;\boldsymbol{u})^2 \to \min. \tag{2}$$

This is called the *least-squares* (*LS*) *method* or *algebraic distance minimization*. However, it is widely known that the solution has strong statistical bias.

A better method known to yield higher accuracy is to regard the data $\{\boldsymbol{x}_\alpha\}$ as perturbed from their *true* positions $\{\bar{\boldsymbol{x}}_\alpha\}$ which are exactly on the curve $F(\boldsymbol{x};\boldsymbol{u}) = 0$ and to simultaneously estimate the true positions $\{\bar{\boldsymbol{x}}_\alpha\}$ and the parameter $\boldsymbol{u}$ that maximize the statistical likelihood. If the noise is subject to isotropic, independent, and identical Gaussian distribution, this reduces to the minimization

$$J_{\rm ML} = \sum_{\alpha=1}^{N} \|\boldsymbol{x}_\alpha - \bar{\boldsymbol{x}}_\alpha\|^2 \to \min, \tag{3}$$

subject to the constraint

$$F(\bar{\boldsymbol{x}}_\alpha;\boldsymbol{u}) = 0, \qquad \alpha = 1, ..., N. \tag{4}$$

This is called *maximum likelihood* (*ML*) *estimation* or *geometric distance minimization*.

Eqs. (3) and (4) can be converted to unconstrained minimization by introducing Lagrange multipliers. Introducing linear approximation by assuming that the noise is small, we can rewrite eq. (3) as follows (see Appendix A for the derivation):

$$J_{\rm ML} = \sum_{\alpha=1}^{N} \frac{F(\boldsymbol{x}_\alpha;\boldsymbol{u})^2}{\|\nabla_{\mathbf{x}} F_\alpha\|^2} \to \min. \tag{5}$$

Here, $\nabla_{\mathbf{x}} F_\alpha$ is the gradient of the function $F(\boldsymbol{x};\boldsymbol{u})$ in eq. (1) with respect to $\boldsymbol{x}$, and the subscript α means that the derivative is evaluated at $\boldsymbol{x} = \boldsymbol{x}_\alpha$. This minimization is known to be effective in many problems and is one of the most widely used methods in computer vision applications [8].

This approach is not limited to curve fitting but can be extended to many other problems. For example, given correspondences of feature points over multiple images, the *trajectory* of a particular point can be identified with a single point in the product space of the images, known as the *joint image*. Fitting a geometric constraint derived from the imaging geometry, such as the *epipolar constraint*, the *trifocal constraint*, the *quadrifocal constraint*, or the *affine constraint*, we can compute the camera motion and the 3-D shape of the scene from the coefficients of the fitted equation [6].

We need not assume isotropic and identical Gaussian noise. If the noise distribution is different from datum to datum, all we need is to introduce covari-

*E-mail `kanatani@suri.it.okayama-u.ac.jp` ance matrices[1] $V[\boldsymbol{x}_\alpha]$ in eq. (5). The author showed that the solution of eq. (5) can be systematically computed by a method called *renormalization* [7] when the function $F(\boldsymbol{x}; \boldsymbol{u})$ can be transformed into a linear form in $\boldsymbol{u}$ by change of variables[2]. This method motivated many similar approaches[3]: Leedan and Meer [14] proposed a method called *HEIV*, and Chojnacki et al. [4] generalized it into what they call *FNS*.

However, a still unanswered question is if eq. (5) is really optimal and if better methods exist at all.

2. How Do We Compare Methods?

The reason this question is so difficult to answer is that it is not clear how to measure the "goodness" of a method. For example, we may measure the accuracy of an estimate $\hat{\boldsymbol{u}}$ by the norm $\|\hat{\boldsymbol{u}} - \boldsymbol{u}\|$ of the difference from its true value $\boldsymbol{u}$. However, there are many objections. Some may say that we should take expectation with respect to our belief or experience as to what value the parameter $\boldsymbol{u}$ is likely to take (known as the *Bayesian approach*). Others may argue that we should rather focus on the error in the application domain, e.g., if we use the value $\hat{\boldsymbol{u}}$ for 3-D reconstruction, we should evaluate the reconstruction error that $\hat{\boldsymbol{u}}$ incurs.

Even if we adopt the simplest measure $\|\hat{\boldsymbol{u}} - \boldsymbol{u}\|$, the problem is not solved, because the noise is random and hence an estimate $\hat{\boldsymbol{u}}$ can happen to coincide with the true value $\boldsymbol{u}$, whatever method we use. So, we need to compute the mean square $E[\|\hat{\boldsymbol{u}} - \boldsymbol{u}\|^2]$, where $E[\,\cdot\,]$ is the expectation with respect to the noise distribution. We prefer the mean square simply because this generally makes the subsequent analysis easy, but there are many objections: some say $\max \|\hat{\boldsymbol{u}} - \boldsymbol{u}\|$ should be used; others say $E[\|\hat{\boldsymbol{u}} - \boldsymbol{u}\|]$ is better. However, the analysis is still complicated even if the simplest mean square is used.

For comparing the performance of statistical estimation methods, statisticians usually simplify the analysis by introducing asymptotic approximations as the number n of observations increases. So, many computer vision researchers analyze asymptotic approximations as the number N of data increases for evaluating the performance of geometric fitting. However, is the number N of data really the number of "observations"?

3. How Can We Increase Data?

The tenet of statistics is to observe random phenomenon and discern the underlying mechanism, assuming that the observed data are deterministically generated but corrupted by random noise. We cannot infer the mechanism from only one observation, but because the noise is random, we can expect that the effect of noise is canceled if observations are repeated; the hidden mechanism will reveals itself as the number of observations increases. Hence, statisticians measure the performance of statistical estimation by the rate of the increase of accuracy as the number n of observation increases. However, if we identify the *number N of data* with the "number of observations", many inconsistencies arise [11, 12].

Firstly, it is assumed in statistics that observations can be repeated as many times as desired *in principle*, i.e., except for the fact that observations entail costs and are subject to many constraints in the real world. In contrast, the input for computer vision is images. We may observe *many different images*, but except in simulations we cannot repeatedly observe the *same* image corrupted with different noise. Hence, the number of observation is always $n = 1$.

Secondly, the unknowns for the standard statistical estimation are the parameters of the underlying mechanism, while for geometric fitting the true values of the data are also unknowns. Hence, if we increase the number of data, the number of unknowns also increases accordingly[4], and their estimation accuracy cannot be improved however many data we observe. For curve fitting, for example, we may correctly estimate the true curve by increasing the number of points, but we cannot estimate their true positions on that curve.

Thirdly, we cannot simply *increase* the data but also need to consider *how* we increase them. For line fitting, for example, the fitting accuracy does not improve if we repeatedly add new points in the neighborhood of a particular point. In contrast, the accuracy will dramatically improve if we distribute new points uniformly along the line to be fitted. So, various theories have recently been proposed for assuming or estimating the *distribution* of the true positions along the curve and marginalizing them over the distribution. Such formulations are called *semiparametric models* [2, 13, 16].

4. Is ML Not Optimal?

If we have a lot of data, ML is known to be *not* optimal. In fact, Endoh et al. [5] pointed out that 3-D interpretation from a dense optical flow field by ML is not optimal, and Ohta [13] showed that the semiparametric model yields a better result. Okatani and Deguchi [16] demonstrated that for estimating 3-D shape and motion from multiple images, the semiparametric model can result in higher accuracy. In all cases, however, the procedure is very complicated, and the performance can surpass ML only when the number of data is very large and the problem has a

[1] The datum $\boldsymbol{x}$ and the parameter $\boldsymbol{u}$ can be subject to some constraints, such as being unit vectors. Multiple constraints, each in the form of eq. (1), can exist, and some of them can be overlapping or redundant. The analysis goes similarly if we introduce pseudoinverse and projection operators [8].

[2] This is the case for many problems in computer vision, including line and conic fitting, homography computation, and estimation of the fundamental matrix [8].

[3] A comprehensive review is in [12].

[4] Such increasing parameters are often called *nuisance parameters* to distinguish them from the remaining *structural parameters*.

special form.

On the other hand, ML in the form of eq. (5) is always effective in all practical applications. At present, no method that surpasses ML in usual situations is known. This implies that ML may be optimal in some sense in "usual" situations. If so, in what sense? What are the "usual situations?

The author gave an answer to these questions [8, 9]. The fact that these issues have not been widely discussed within the statistical community seems largely because of the paradigm that statistics is to overcome randomness by repeated observations. Also, statisticians are mostly unfamiliar with geometric fitting problems in the form as appears in computer vision applications.

In the following, we describe the author's formulation and compared it with the recent results of Chernov and Lesort [3].

5. KCR Lower Bound

The fundamental difference of the author's approach from the standard statistical estimation is that the analysis is focused on *small noise* rather than asymptotic analysis for large number n of observations. This is motivated by the fact that computer vision deals with pixel-level small errors, while the traditional statistical estimation is mainly concerned with large errors, e.g., in fieldwork in real environments.

Estimating the parameter $\boldsymbol{u}$ from the data $\{\boldsymbol{x}_\alpha\}$ means finding an estimate $\hat{\boldsymbol{u}}$ expressed as a function of the data $\{\boldsymbol{x}_\alpha\}$:

$$\hat{\boldsymbol{u}} = \hat{\boldsymbol{u}}(\boldsymbol{x}_1, ..., \boldsymbol{x}_N). \tag{6}$$

The function $\hat{\boldsymbol{u}}$ is called an *estimator* of $\boldsymbol{u}$. Consider the *covariance matrix*[5] of estimator $\hat{\boldsymbol{u}}$:

$$V[\hat{\boldsymbol{u}}] = E[(\hat{\boldsymbol{u}} - \boldsymbol{u})(\hat{\boldsymbol{u}} - \boldsymbol{u})^\top]. \tag{7}$$

We assume that each datum $\boldsymbol{x}_\alpha$ is displaced from its true value $\bar{\boldsymbol{x}}_\alpha$ by component-wise independent Gaussian noise of mean 0 and standard deviation ε:

$$\boldsymbol{x}_\alpha = \bar{\boldsymbol{x}}_\alpha + \Delta\boldsymbol{x}_\alpha, \qquad \Delta\boldsymbol{x}_\alpha \sim N(\boldsymbol{0}, \varepsilon^2\boldsymbol{I}). \tag{8}$$

We call ε the *noise level.* The following argument holds for a more general noise distribution[6] [8], but here we concentrate only on the isotropic Gaussian distribution for simplicity.

Let $\Delta\boldsymbol{u}$ be the error in the estimator $\hat{\boldsymbol{u}}$:

$$\hat{\boldsymbol{u}} = \boldsymbol{u} + \Delta\boldsymbol{u}. \tag{9}$$

Substituting eqs. (8) and (9) into eq. (5), using Taylor expansion in $\Delta\boldsymbol{x}_\alpha$ and $\Delta\boldsymbol{u}$ by assuming that the noise is small, and computing the value $\Delta\boldsymbol{u}$ that minimizes eq. (5), we find that the covariance matrix $V[\hat{\boldsymbol{u}}_{\rm ML}]$ of the ML estimator $\hat{\boldsymbol{u}}_{\rm ML}$ can be expanded in ε as follows [8] (see Appendix B for the derivation):

$$V[\hat{\boldsymbol{u}}_{\rm ML}] = \varepsilon^2\left(\sum_{\alpha=1}^N \frac{(\nabla_{\mathbf{u}}\bar{F}_\alpha)(\nabla_{\mathbf{u}}\bar{F}_\alpha)^\top}{\|\nabla_{\mathbf{x}}\bar{F}_\alpha\|^2}\right)^{-1} + O(\varepsilon^4). \tag{10}$$

Here, $\nabla_{\mathbf{u}}\bar{F}_\alpha$ denotes the gradient of the function $F(\boldsymbol{x};\boldsymbol{u})$ in eq. (1) with respect to $\boldsymbol{u}$, and $\bar{F}_\alpha$ means that the derivation is evaluated at $\boldsymbol{x} = \bar{\boldsymbol{x}}_\alpha$.

We can also show that the first term on the right-hand side of eq. (10) is a lower bound on an arbitrary unbiased estimator $\hat{\boldsymbol{u}}$ in the following sense [8] (see Appendix C for the derivation):

$$V[\hat{\boldsymbol{u}}] \succ \varepsilon^2\left(\sum_{\alpha=1}^N \frac{(\nabla_{\mathbf{u}}\bar{F}_\alpha)(\nabla_{\mathbf{u}}\bar{F}_\alpha)^\top}{\|\nabla_{\mathbf{x}}\bar{F}_\alpha\|^2}\right)^{-1}. \tag{11}$$

Here, $\succ$ denotes that the difference between the left-hand side and the right-hand side is a positive semidefinite symmetric matrix.

Thus, the covariance matrix of the ML estimator $\hat{\boldsymbol{u}}_{\rm ML}$ attains the lower bound in the first order in ε (i.e., if terms $O(\varepsilon^4)$ are ignored). In this sense, ML is optimal. Chernov and Lesort [3] called the right-hand side of eq. (11) the *KCR* (*Kanatani-Cramer-Rao*) *lower bound* .

6. CR Lower Bound

The KCR lower bound is different from the well known CR (Cramer-Rao) lower bound: the difference is less in the bound than in the *problem.* As mentioned earlier, statistical estimation is to discern the hidden mechanism by repeating observations. This is formalized as estimation of the parameter $\boldsymbol{\theta}$ by observing n independent instances $\boldsymbol{x}_1$, ..., $\boldsymbol{x}_n$ of a random variable $\boldsymbol{X}$ occurring according to an assumed probability density $p(\boldsymbol{x};\boldsymbol{\theta})$. *Maximum likelihood* (*ML*) *estimation* is to compute the value $\hat{\boldsymbol{\theta}}_{\rm ML}$ of $\boldsymbol{\theta}$ that maximizes the *likelihood*

$$L = \prod_{i=1}^n p(\boldsymbol{x}_i;\boldsymbol{\theta}). \tag{12}$$

Considering the asymptotic limit $n \to \infty$ and invoking the *law of large numbers*, which states that the sample mean of independent instances of a random variable converges to its expectation as $n \to \infty$, together with the *central limit theorem*, which states that the distribution of the sample mean can be asymptotically approximated by a Gaussian distribution, we can show under a fairly general condition that the covariance matrix $V[\hat{\boldsymbol{\theta}}_{\rm ML}]$ of the ML estimator $\hat{\boldsymbol{\theta}}_{\rm ML}$ is expanded in $1/n$ in the form

$$V[\hat{\boldsymbol{\theta}}_{\rm ML}] = \frac{1}{n}\boldsymbol{J}^{-1} + O(\frac{1}{n^2}), \tag{13}$$

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

$$\boldsymbol{J} = E[\Big(\nabla_\theta \log p(\boldsymbol{x};\boldsymbol{\theta})\Big)\Big(\nabla_\theta \log p(\boldsymbol{x};\boldsymbol{\theta})\Big)^\top]. \tag{14}$$

[5] Its trace ${\rm tr}V[\hat{\boldsymbol{u}}] = E[\|\hat{\boldsymbol{u}} - \boldsymbol{u}\|^2]$ is the *mean-square* error.

[6] The same argument applies to a wide class of probability distributions called the *exponential family* [8].

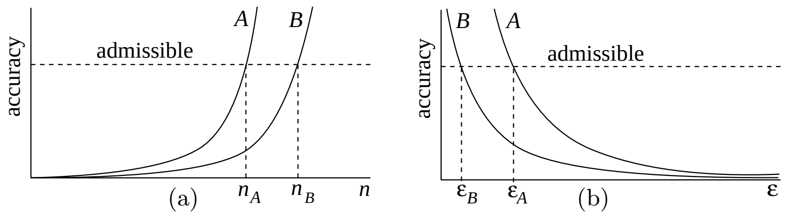

Figure 1: (a) For the standard statistical estimation, it is desired that the accuracy increases rapidly as $n \to \infty$ for the number n of observations, because admissible accuracy can be reached with a smaller number of observations. (b) For geometric fitting, it is desired that the accuracy increases rapidly as $\varepsilon \to 0$ for the noise level ε, because larger data uncertainty can be tolerated for admissible accuracy.

The expectation $E[\,\cdot\,]$ is taken with respect to the probability density $p(\boldsymbol{x};\boldsymbol{\theta})$. The first term on the right-hand side of eq. (13) is called the *CR* (*Cramer-Rao*) *lower bound*, and the following *Cramer-Rao inequality* can be proved for an arbitrary unbiased estimator $\hat{\boldsymbol{\theta}}$:

$$V[\hat{\boldsymbol{\theta}}] \succ \frac{1}{n}\boldsymbol{J}^{-1}. \tag{15}$$

It follows that the covariance matrix of the ML estimator $\hat{\boldsymbol{\theta}}_{\rm ML}$ attains the CR lower bound in the first order (i.e., if terms $O(1/n^2)$ are ignored) in the asymptotic limit $n \to \infty$ of the number n of observations. This fact is known as the *asymptotic efficiency* of ML, and in this sense ML is optimal for the standard statistical estimation.

7. Duality of Interpretation

Thus, the KCR lower bound and the CR lower bound are different concepts. Yet, there is something common in their formalisms.

The reason why the performance of the standard statistical estimation is evaluated in the asymptotic limit $n \to \infty$ of the number n of observations is that a method whose accuracy increases rapidly as $n \to \infty$ can attain admissible accuracy with a fewer number of observations (Fig. 1(a)). Such a method is desirable if we consider the cost of observations in real situations.

In contrast, the performance of geometric fitting should be evaluated in the limit $\varepsilon \to 0$ of the noise level ε, because a method whose accuracy increases rapidly as $\varepsilon \to 0$ can tolerate larger uncertainty for admissible accuracy (Fig. 1(b)). Such a method is preferable if we consider the uncertainty inherent of image processing operations.

Now, consider the following thought experiment. For geometric fitting, the image data may not be exact due to the uncertainty of image processing operations, but they always have the same value however many times we observe them. Suppose, hypothetically, they change their values each time we observe them. Then, we would obtain n different values for n observations. Under independent Gaussian noise, an optimal estimate of the true value is their sample mean. As is well known, the standard deviation of a sample mean of n observations is $1/\sqrt{n}$ times that of the individual observations.

Thus, repeating such hypothetical observations is equivalent to reducing the noise level ε to $\varepsilon/\sqrt{n}$. It follows that the perturbation analysis for $\varepsilon \to 0$ is mathematically equivalent to the asymptotic analysis for $n \to \infty$ of the number n of hypothetical observations. This is the reason why the asymptotic approximation $\cdots + O(1/\sqrt{n^k})$ for the standard statistical estimation corresponds to $\cdots + O(\varepsilon^k)$ for the geometric fitting counterpart [10].

This type of duality of interpretation also arises for *model selection*: we obtain the *geometric AIC* and the *geometric MDL* for geometric fitting as the counterparts of Akaike's *AIC* (*Akaike information criterion*) [1] and Rissanen's *MDL* (*minimum description length*) [17], respectively [10].

8. Condition for Optimality

Since its first introduction in [8], the KCR lower bound in eq. (11) has scarcely been recognized by the computer vision community. Even today, there are some who doubt its validity[7]. Recently, however, Chernov and Lesort [3] proved that the KCR lower bound holds under a weaker assumption.

Eq. (11) is derived by assuming *unbiasedness* [8] (see Appendix C):

$$E[\hat{\boldsymbol{u}}] = \boldsymbol{u}. \tag{16}$$

Chernov and Lesort [3] replaced this by the following *consistency*:

$$\lim_{\varepsilon \to 0} \hat{\boldsymbol{u}} = \boldsymbol{u}. \tag{17}$$

This states that *the estimate* $\hat{\boldsymbol{u}}$ *gives the true value* $\boldsymbol{u}$ *in the absence of noise*. This is trivially confirmed for all practical estimation methods[8], since methods that do not satisfy this are not worth considering.

Suppose the data $\boldsymbol{x}_\alpha$ are m-dimensional vectors and the parameter $\boldsymbol{u}$ is a p-dimensional vector. Substituting eqs. (8) into the right-hand side of eq. (6) and using Taylor expansion, we see that the consistency condition (17) implies

$$\hat{\boldsymbol{u}} = \boldsymbol{u} + \sum_{\alpha=1}^{N} \Big(\nabla_{\mathbf{x}_\alpha} \hat{\boldsymbol{u}}\Big) \Delta \boldsymbol{x}_\alpha + O(\varepsilon^2), \tag{18}$$

[7] Some dismissed the result, saying that it appeared only in the author's monograph, not in peer-reviewed journals. The truth is that the result was submitted to journals but rejected as not being useful in practice.

[8] This is not so for standard statistical estimation problems, i.e., it is not easy to prove the consistency in the sense that the estimate converges to the true value in some probabilistic sense as the number n of observations goes to infinity.

where $\nabla_{\mathbf{x}_\alpha}\hat{\boldsymbol{u}}$ denotes the $p \times m$ matrix whose (ij) element is $\partial \hat{u}_i / \partial x_{j\alpha}$; derivatives are evaluated at $\boldsymbol{x}_\alpha = \bar{\boldsymbol{x}}_\alpha$, $\alpha = 1, ..., N$. From eq. (18), we have

$$(\hat{\boldsymbol{u}}-\boldsymbol{u})(\hat{\boldsymbol{u}}-\boldsymbol{u})^\top = \sum_{\alpha,\beta=1}^{N} \Big(\nabla_{\mathbf{x}_\alpha}\hat{\boldsymbol{u}}\Big)\Delta\boldsymbol{x}_\alpha \Delta\boldsymbol{x}_\beta^\top \Big(\nabla_{\mathbf{x}_\beta}\hat{\boldsymbol{u}}\Big)^\top + (\text{terms of order 3 or higher in } \{\Delta\boldsymbol{x}_\alpha\}). \qquad (19)$$

Taking expectation on both sides, we obtain the covariance matrix $V[\hat{\boldsymbol{u}}]$ in eq. (7) in the form

$$V[\hat{\boldsymbol{u}}] = \varepsilon^2 \sum_{\alpha=1}^{N} \Big(\nabla_{\mathbf{x}_\alpha}\hat{\boldsymbol{u}}\Big)\Big(\nabla_{\mathbf{x}_\alpha}\hat{\boldsymbol{u}}\Big)^\top + O(\varepsilon^4), \qquad (20)$$

where we use the fact that noise is independent for each α and hence from eq. (8) we have[9]

$$E[\Delta\boldsymbol{x}_\alpha \Delta\boldsymbol{x}_\beta^\top] = \delta_{\alpha\beta}\varepsilon^2 \boldsymbol{I}. \qquad (21)$$

The remainder term on the right-hand side of eq. (20) is $O(\varepsilon^4)$ because of the symmetry of the noise distribution: the third order terms in $\Delta\boldsymbol{x}_\alpha$ are 0 in expectation

9. Derivation of the KCR Lower Bound

Chernov and Lesort [3] derived the KCR lower bound in much the same way as in the original derivation in [8], using the *variational principle* with respect to the true values $\{\bar{\boldsymbol{x}}_\alpha\}$ and the parameter $\boldsymbol{u}$. If we perturb $\bar{\boldsymbol{x}}_\alpha$ and $\boldsymbol{u}$ into $\bar{\boldsymbol{x}}_\alpha + \delta\bar{\boldsymbol{x}}_\alpha$ and $\boldsymbol{u} + \delta\boldsymbol{u}$ in such a way that eq. (4) is not violated, i.e., $F(\bar{\boldsymbol{x}}_\alpha + \delta\bar{\boldsymbol{x}}_\alpha; \boldsymbol{u} + \delta\boldsymbol{u}) = 0$, we have for arbitrary perturbations[10] $\{\delta\bar{\boldsymbol{x}}_\alpha\}$ and $\delta\boldsymbol{u}$

$$(\nabla_{\mathbf{x}}\bar{F}_\alpha, \delta\bar{\boldsymbol{x}}_\alpha) + (\nabla_{\mathbf{u}}\bar{F}_\alpha, \delta\boldsymbol{u}) = 0, \qquad (22)$$

where in the following we write $(\boldsymbol{a}, \boldsymbol{b})$ for the the inner product of vectors $\boldsymbol{a}$ and $\boldsymbol{b}$. The notations $\nabla_{\mathbf{x}}\bar{F}_\alpha$ and $\nabla_{\mathbf{u}}\bar{F}_\alpha$ have the same meaning as in eq. (10).

From the definition (6) of the estimator $\hat{\boldsymbol{u}}$ and the consistency condition (17), we have the identity $\boldsymbol{u} = \hat{\boldsymbol{u}}(\bar{\boldsymbol{x}}_1, ..., \bar{\boldsymbol{x}}_N)$. Hence,

$$\sum_{\alpha=1}^{N} \Big(\nabla_{\mathbf{x}_\alpha}\hat{\boldsymbol{u}}\Big)\delta\bar{\boldsymbol{x}}_\alpha = \delta\boldsymbol{u}, \qquad (23)$$

for arbitrary variations $\{\delta\bar{\boldsymbol{x}}_\alpha\}$ and $\delta\boldsymbol{u}$ that satisfy eq. (22). From this, we conclude

$$\sum_{\alpha=1}^{N} \Big(\nabla_{\mathbf{x}_\alpha}\hat{\boldsymbol{u}}\Big)\Big(\nabla_{\mathbf{x}_\alpha}\hat{\boldsymbol{u}}\Big)^\top \succ \left(\sum_{\alpha=1}^{N} \frac{(\nabla_{\mathbf{u}}\bar{F}_\alpha)(\nabla_{\mathbf{u}}\bar{F}_\alpha)^\top}{\|\nabla_{\mathbf{x}}\bar{F}_\alpha\|^2}\right)^{-1}, \qquad (24)$$

by invoking the following lemma:

Lemma 1 *Let $\boldsymbol{a}_1$, ..., $\boldsymbol{a}_N$ be nonzero m-dimensional vectors, and $\boldsymbol{b}_1$, ..., $\boldsymbol{b}_N$ p-dimensional vectors $\boldsymbol{b}_1$, ..., $\boldsymbol{b}_N$ spanning $\mathcal{R}^p$. If there exist $p \times m$ matrices $\boldsymbol{A}_1$, ..., $\boldsymbol{A}_N$ such that equality*

$$\sum_{\alpha=1}^{N} \boldsymbol{A}_\alpha \boldsymbol{x}_\alpha = \boldsymbol{y} \qquad (25)$$

holds for any m-dimensional vectors $\boldsymbol{x}_1$, ..., $\boldsymbol{x}_N$ and an arbitrary p-dimensional vector $\boldsymbol{u}$ that satisfy

$$(\boldsymbol{a}_\alpha, \boldsymbol{x}_\alpha) + (\boldsymbol{b}_\alpha, \boldsymbol{y}) = 0, \qquad \alpha = 1, ..., N, \qquad (26)$$

then the following inequality holds:

$$\sum_{\alpha=1}^{N} \boldsymbol{A}_\alpha \boldsymbol{A}_\alpha^\top \succ \left(\sum_{\alpha=1}^{N} \frac{\boldsymbol{b}_\alpha \boldsymbol{b}_\alpha^\top}{\|\boldsymbol{a}_\alpha\|^2}\right)^{-1}. \qquad (27)$$

Chernov and Lesort [3] proved this lemma by arguments similar to the author's [8]. The right-hand side of eq. (24) is nothing but the KCR lower bound except for the multiplier ε^2. From (10), we see that the covariance matrix $V[\hat{\boldsymbol{u}}_{\mathrm{ML}}]$ of the ML estimator $\hat{\boldsymbol{u}}_{\mathrm{ML}}$ satisfies the lower bound except for terms $O(\varepsilon^4)$.

10. Observations

The theory of Chernov and Lesort [3] is an expansion of the author's theory [8] in that they do not use the unbiasedness assumption (16). Also, their argument is clearer[11] and easier to understand than the author's [8] (see Appendix C). On the other hand, the author's derivation imposes the lower bound on the covariance matrix $V[\hat{\boldsymbol{u}}]$ itself (irrespective of the noise level ε), while Chernov and Lesort [3] only derived the lower bound on the leading term in the expansion (20) in ε. Their reasoning sounds natural if we note that the consistency (17) implies the unbiasedness (16) in the limit $\varepsilon \to 0$. Reflecting the weaker assumption, their conclusion is somewhat weaker but still sufficient for characterizing properties for $\varepsilon \to 0$. In particular, the optimality of ML follows.

Chernov and Lesort [3] also pointed out a rather surprising fact. Analyzing the KCR lower bound, they showed that seemingly suboptimal methods can be optimal. One example is the problem of fitting a circle

$$(x-a)^2 + (y-b)^2 = R^2 \qquad (28)$$

to given points $\{(x_\alpha, y_\alpha)\}$, $\alpha = 1$, ..., N. The LS of eq. (2) becomes

$$J_{\mathrm{LS}} = \sum_{\alpha=1}^{N} ((x_\alpha - a)^2 + (y_\alpha - b)^2 - R^2)^2 \to \min, \qquad (29)$$

while the ML of eq. (3) has the form

$$J_{\mathrm{ML}} = \frac{1}{4}\sum_{\alpha=1}^{N} \frac{((x_\alpha - a)^2 + (y_\alpha - b)^2 - R^2)^2}{(x_\alpha - a)^2 + (y_\alpha - b)^2} \to \min. \qquad (30)$$

[9] The symbol $\delta_{\alpha\beta}$ is the Kronecker delta, taking 1 for $\alpha = \beta$ and 0 otherwise.

[10] This is not the usual Taylor expansion but an *identity* for *infinitesimal variations* $\delta\bar{\boldsymbol{x}}_\alpha$ and $\delta\boldsymbol{u}$, so no higher order terms appear. This corresponds to what is known as *principle of virtual work* in mechanics. Note that $\delta\bar{\boldsymbol{x}}_\alpha$ is a hypothetical variation of the *true* value $\bar{\boldsymbol{x}}_\alpha$, not the observation noise $\Delta\boldsymbol{x}_\alpha$.

[11] However, the proof of Lemma 1 is not easy, requiring as sophisticated mathematical techniques as in the proof of the author's.

It can be shown that both attain the KCR lower bound in the first order, and in this sense both are optimal [3]. In general, modification of eq. (3) in the form

$$J = \sum_{\alpha=1}^{N} \frac{c(\boldsymbol{u})F(\boldsymbol{x}_\alpha;\boldsymbol{u})^2}{\|\nabla_{\mathbf{x}} F_\alpha\|^2} \to \min \tag{31}$$

does not affect the covariance matrix $V[\hat{\boldsymbol{u}}]$ of the resulting estimator $\hat{\boldsymbol{u}}$ in the first order, where $c(\boldsymbol{u})$ is an arbitrary positive function of $\boldsymbol{u}$ (see Appendix D). Eq. (28) is obtained from eq. (30) by inserting $c(a,b,R) = R$ to the numerator; replacing the denominator in eq. (30) by R does not affect the solution as far as the leading term in ε is concerned.

Chernov and Lesort [3] conducted simulations and confirmed that the solution of the LS of eq. (29) and the solution of the ML of eq. (30) behave quite similarly when the noise is extremely small. As the noise increases, however, ML generally performs better than LS, but surprisingly LS is better than ML above a certain noise level in some situations. Chernov and Lesort [3] pointed out that the cause of this anomaly can be traced back to a hidden singularity[12] in eq. (28).

11. Conclusions

As we have seen, the KCR lower bound is the most important characterization of geometric fitting. As Chernov and Lesort [3] showed, however, methods that are optimal in the sense of the KCR lower bound may perform differently in the presence of large noise. In this sense, finding additional characterization that complements the KCR lower bound remains a crucial problem.

Recently, Mühlich and Mester [15] proposed a new fitting technique for problems for which the constraint (1) can be linearized in $\boldsymbol{u}$ (including line and conic fitting, homography computation, and estimation of the fundamental matrix). For such problems, the author's renormalization, the HEIV of Leedan and Meer [14], and the FNS of Chojnacki et al. [4] all yield a solution that attains the KCR lower bound in the first order. Mühlich and Mester [15] extended a technique called *whitening* or *equilibration* and showed that their method, though not optimal in the sense of the KCR lower bound, can produce a solution with comparable or higher accuracy with less computational failures when the noise is large.

One of the major reasons why such attempts have not been made until recently seem to lie in the fact that computer vision researchers are likely to take textbooks of statistics and discourses of distinguished statisticians for granted and blindly follow the asymptotic analysis as $N \to \infty$ for the number N of data. Rather, computer vision researchers should bring forth theories and analyses specific to computer vision applications. The studies of Chernov and Lesort [3] and Mühlich and Mester [15] are good examples.

Acknowledgments: This work was supported in part by the Ministry of Education, Culture, Sports, Science and Technology, Japan, under the Grant in Aid for Scientific Research C(2) (No. 15500113).

References

[1] H. Akaike, A new look at the statistical model identification, IEEE Trans. Autom. Control, **16**-6 (1977), 716–723.

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

[3] N. Chernov and C. Lesort, Statistical efficiency of curve fitting algorithms, *Comput. Stat. Data Anal.*, **47**-4 (2004-11), 713–728.

[4] W. Chojnacki, M. J. Brooks, A. van den Hengel and D. Gawley, On the fitting of surfaces to data with covariances, *IEEE Trans. Patt. Anal. Mach. Intell.*, **22**-11 (2000), 1294–1303.

[5] T. Endoh, T. Toriu, and N. Tagawa, A superior estimator to the maximum likelihood estimator on 3-D motion estimation from noisy optical flow, *IEICE Trans. Inf. & Sys.*, **E77-D**-11 (1994), 1240–1246.

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

[7] K. Kanatani Renormalization for unbiased estimation, *Proc. 4th Int. Conf. Comput. Vision*, May, 1993, Berlin, Germnay, pp. 599–606.

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

[9] K. Kanatani, Cramer-Rao lower bounds for curve fitting, *Graphical Models Image Processing*, **60**-2 (1998), 93–99.

[10] K. Kanatani, Uncertainty modeling and model selection for geometric inference, *IEEE Trans. Patt. Anal. Machine Intell.*, **26**-10 (2004), 1307–1319.

[11] K. Kanatani, For geometric inference from images, what kind of statistical model is necessary? *Sys. Comp. Japan*, **35**-6 (2004), 1–9.

[12] K. Kanatani, Uncertainty modeling and geometric inference, *Memoirs of the Faculty of Engineering*, **38**-1/2 (2004), 39–60.

[13] N. Ohta, Motion parameter estimation from optical flow without nuisance parameters, *3rd Int. Workshop on Statistical and Computational Theory of Vision* October 2003, Nice, France: `http://www.stat.ucla.edu/~sczhu/Workshops/SCTV2003.html`

[14] Y. Leedan and P. Meer, Heteroscedastic regression in computer vision: Problems with bilinear constraint, *Int. J. Comput. Vision.*, **37**-2 (2000), 127–150.

[15] M. Mühlich and R. Mester, Unbiased errors-in-variables estimation using generalized eigensystem analysis, *Proc. 2nd Workshop on Statistical Methods in Video Processing*, May 2004, Prague, Czech, pp. 38–49.

[16] T. Okatani and K. Deguchi, Toward a statistically optimal method for estimating geometric relations from noisy data: Cases of linear relations, *Proc. IEEE Conf. Comput. Vision Pattern Recog.*, June 2003, Madison, WI, U.S.A., Vol. 1, pp. 432–439.

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

Appendix

A: Linear Approximation of ML

Substituting $\bar{\boldsymbol{x}}_\alpha = \boldsymbol{x}_\alpha - \Delta\boldsymbol{x}_\alpha$ into eq. (4) and assuming that the noise term $\Delta\boldsymbol{x}_\alpha$ is small, we obtain

[12] Hence, this observation does not apply to most problems of computer vision including conic fitting. In general, ML is far superior to LS.

the linear approximation

$$F_\alpha - (\nabla_{\mathbf{x}} F_\alpha, \Delta\boldsymbol{x}_\alpha) = \mathbf{0}. \tag{32}$$

Introduce Lagrange multipliers λ_α for this constraint, and let

$$L = \frac{1}{2}\sum_{\alpha=1}^{N} \|\Delta\boldsymbol{x}_\alpha\|^2 + \sum_{\alpha=1}^{N} \lambda_\alpha (F_\alpha - (\nabla_{\mathbf{x}} F_\alpha, \Delta\boldsymbol{x}_\alpha)). \tag{33}$$

The solution $\Delta\boldsymbol{x}_\alpha$ that minimizes L subject to the constraint (32) satisfies $\nabla_{\Delta\mathbf{x}_\alpha} L = \mathbf{0}$, $\alpha = 1, ..., N$, or

$$\Delta\boldsymbol{x}_\alpha - \lambda_\alpha \nabla_{\mathbf{x}} F_\alpha = \mathbf{0}. \tag{34}$$

Hence, $\Delta\boldsymbol{x}_\alpha = \lambda_\alpha \nabla_{\mathbf{x}} F_\alpha$. Substitution of this into eq. (32) yields

$$F_\alpha - (\nabla_{\mathbf{x}} F_\alpha, \lambda_\alpha \nabla_{\mathbf{x}} F_\alpha) = 0, \tag{35}$$

from which we obtain λ_α in the form

$$\lambda_\alpha = \frac{F_\alpha}{\|\nabla_{\mathbf{x}} F_\alpha\|^2}. \tag{36}$$

Thus, eq. (3) is rewritten in the form

$$\begin{aligned} J_{\mathrm{ML}} &= \sum_{\alpha=1}^{N} \|\lambda_\alpha \nabla_{\mathbf{x}} F_\alpha\|^2 = \sum_{\alpha=1}^{N} \frac{F_\alpha^2}{\|\nabla_{\mathbf{x}} F_\alpha\|^4} \|\nabla_{\mathbf{x}} F_\alpha\|^2 \\ &= \sum_{\alpha=1}^{N} \frac{F_\alpha^2}{\|\nabla_{\mathbf{x}} F_\alpha\|^2}, \end{aligned} \tag{37}$$

resulting in eq. (5). □

B: Covariance Matrix of ML

After substitution of eqs. (8) and (9) into eq. (5) and Taylor expansion, the function J_{ML} is written in the following form:

$$J_{\mathrm{ML}} = \sum_{\alpha=1}^{N} \frac{((\nabla_{\mathbf{x}} \bar{F}_\alpha, \Delta\boldsymbol{x}_\alpha) + (\nabla_{\mathbf{u}} \bar{F}_\alpha, \Delta\boldsymbol{u}))^2}{\|\nabla_{\mathbf{x}} \bar{F}_\alpha\|^2} + O(\varepsilon^3). \tag{38}$$

Replacing $\|\nabla_{\mathbf{x}} F_\alpha\|^2$ by $\|\nabla_{\mathbf{x}} \bar{F}_\alpha\|^2$ in the denominator on the right hand side does not affect the leading term because the numerator is $O(\varepsilon^2)$; the difference is absorbed into the remainder term $O(\varepsilon^3)$.

If we find $\Delta\boldsymbol{u}$ that minimizes eq. (38), the ML estimator $\hat{\boldsymbol{u}}_{\mathrm{ML}}$ is given by $\boldsymbol{u} + \Delta\boldsymbol{u}$. The solution $\Delta\boldsymbol{u}$ is obtained by solving $\nabla_{\Delta\mathbf{u}} J_{\mathrm{ML}} = \mathbf{0}$. Since the first term on the right-hand side of eq. (38) is a quadratic form in $\Delta\boldsymbol{u}_\alpha$, we obtain

$$\begin{aligned} &2\sum_{\alpha=1}^{N} \frac{((\nabla_{\mathbf{x}} \bar{F}_\alpha, \Delta\boldsymbol{x}_\alpha) + (\nabla_{\mathbf{u}} \bar{F}_\alpha, \Delta\boldsymbol{u}))\nabla_{\mathbf{u}} \bar{F}_\alpha}{\|\nabla_{\mathbf{x}} \bar{F}_\alpha\|^2} \\ &= O(\varepsilon^2), \end{aligned} \tag{39}$$

which is rewritten in the form

$$\begin{aligned} &\sum_{\alpha=1}^{N} \frac{(\nabla_{\mathbf{u}} \bar{F}_\alpha)(\nabla_{\mathbf{u}} \bar{F}_\alpha)^\top}{\|\nabla_{\mathbf{x}} \bar{F}_\alpha\|^2} \Delta\boldsymbol{u} \\ &= -\sum_{\alpha=1}^{N} \frac{(\nabla_{\mathbf{u}} \bar{F}_\alpha)(\nabla_{\mathbf{x}} \bar{F}_\alpha)^\top}{\|\nabla_{\mathbf{x}} \bar{F}_\alpha\|^2} \Delta\boldsymbol{x}_\alpha + O(\varepsilon^2). \end{aligned} \tag{40}$$

From this, we obtain

$$\begin{aligned} &\sum_{\alpha=1}^{N} \frac{(\nabla_{\mathbf{u}} \bar{F}_\alpha)(\nabla_{\mathbf{u}} \bar{F}_\alpha)^\top}{\|\nabla_{\mathbf{x}} \bar{F}_\alpha\|^2} \Delta\boldsymbol{u}\Delta\boldsymbol{u}^\top \sum_{\beta=1}^{N} \frac{(\nabla_{\mathbf{u}} \bar{F}_\beta)(\nabla_{\mathbf{u}} \bar{F}_\beta)^\top}{\|\nabla_{\mathbf{x}} \bar{F}_\beta\|^2} \\ &= \sum_{\alpha,\beta=1}^{N} \frac{(\nabla_{\mathbf{u}} \bar{F}_\alpha)(\nabla_{\mathbf{x}} \bar{F}_\alpha)^\top}{\|\nabla_{\mathbf{x}} \bar{F}_\alpha\|^2} \Delta\boldsymbol{x}_\alpha \Delta\boldsymbol{x}_\beta^\top \frac{(\nabla_{\mathbf{x}} \bar{F}_\beta)(\nabla_{\mathbf{u}} \bar{F}_\beta)^\top}{\|\nabla_{\mathbf{x}} \bar{F}_\alpha\|^2} \\ &\quad + O(\varepsilon^3). \end{aligned} \tag{41}$$

Taking expectation on both sides and recalling the definition $V[\hat{\boldsymbol{u}}_{\mathrm{ML}}] = E[\Delta\boldsymbol{u}\Delta\boldsymbol{u}^\top]$, we obtain

$$\begin{aligned} &\sum_{\alpha=1}^{N} \frac{(\nabla_{\mathbf{u}} \bar{F}_\alpha)(\nabla_{\mathbf{u}} \bar{F}_\alpha)^\top}{\|\nabla_{\mathbf{x}} \bar{F}_\alpha\|^2} V[\hat{\boldsymbol{u}}_{\mathrm{ML}}] \sum_{\beta=1}^{N} \frac{(\nabla_{\mathbf{u}} \bar{F}_\beta)(\nabla_{\mathbf{u}} \bar{F}_\beta)^\top}{\|\nabla_{\mathbf{x}} \bar{F}_\beta\|^2} \\ &= \sum_{\alpha=1}^{N} \frac{(\nabla_{\mathbf{u}} \bar{F}_\alpha)(\nabla_{\mathbf{x}} \bar{F}_\alpha)^\top}{\|\nabla_{\mathbf{x}} \bar{F}_\alpha\|^2} \frac{(\nabla_{\mathbf{x}} \bar{F}_\alpha)(\nabla_{\mathbf{u}} \bar{F}_\alpha)^\top}{\|\nabla_{\mathbf{x}} \bar{F}_\alpha\|^2} + O(\varepsilon^4) \\ &= \sum_{\alpha=1}^{N} \frac{(\nabla_{\mathbf{u}} \bar{F}_\alpha)(\nabla_{\mathbf{x}} \bar{F}_\alpha)^\top}{\|\nabla_{\mathbf{x}} \bar{F}_\alpha\|^2} + O(\varepsilon^4), \end{aligned} \tag{42}$$

where we have used eq. (21) and the fact that $E[O(\varepsilon^3)] = O(\varepsilon^4)$. From eq. (42) follows eq. (10). □

C: Derivation of the KCR Lower Bound

The original derivation of the KCR lower bound is as follows [8]. The unbiasedness condition (16) is rewritten as

$$E[\hat{\boldsymbol{u}} - \boldsymbol{u}] = \mathbf{0}, \tag{43}$$

which should be an *identity* in $\{\bar{\boldsymbol{x}}_\alpha\}$ and $\boldsymbol{u}$ that satisfies eq. (4).

From the definition of the expectation $E[\,\cdot\,]$, the infinitesimal variation of $E[\hat{\boldsymbol{u}} - \boldsymbol{u}]$ is[13]

$$\begin{aligned} \delta \int (\hat{\boldsymbol{u}} - \boldsymbol{u}) p_1 \cdots p_N d\boldsymbol{x} &= -\int (\delta\boldsymbol{u}) p_1 \cdots p_N d\boldsymbol{x} \\ &+ \sum_{\alpha=1}^{N} \int (\hat{\boldsymbol{u}} - \boldsymbol{u}) p_1 \cdots \delta p_\alpha \cdots p_N d\boldsymbol{x} \\ &= -\delta\boldsymbol{u} + \int (\hat{\boldsymbol{u}} - \boldsymbol{u}) \sum_{\alpha=1}^{N} (p_1 \cdots \delta p_\alpha \cdots p_N) d\boldsymbol{x}, \end{aligned} \tag{44}$$

where $\int d\boldsymbol{x}$ is a shorthand of $\int \cdots \int d\boldsymbol{x}_1 \cdots \boldsymbol{x}_N$. By assumption, the probability density of $\boldsymbol{x}_\alpha$ is

$$p(\boldsymbol{x}_\alpha) = \frac{1}{(\sqrt{2\pi})^n \varepsilon^n} e^{-\|\boldsymbol{x}_\alpha - \bar{\boldsymbol{x}}_\alpha\|^2 / 2\varepsilon^2}, \tag{45}$$

which we abbreviate to p_α. The infinitesimal variation of eq. (45) with respect to $\bar{\boldsymbol{x}}_\alpha$ is

$$\delta p_\alpha = (\boldsymbol{l}_\alpha, \delta\bar{\boldsymbol{x}}_\alpha) p_\alpha, \tag{46}$$

[13] Recall that we consider variations in $\{\bar{\boldsymbol{x}}_\alpha\}$ (not $\{\boldsymbol{x}_\alpha\}$) and $\boldsymbol{u}$. Since the estimator $\hat{\boldsymbol{u}}$ is a function of the data $\{\boldsymbol{x}_\alpha\}$, it does not change for these variations. The variation $\delta\boldsymbol{u}$ is independent of $\{\boldsymbol{x}_\alpha\}$, so it can be moved outside the integral $\int d\boldsymbol{x}$. Also note that $\int p_1 \cdots \delta p_\alpha \cdots p_N d\boldsymbol{x} = 1$.

where we define the *score* $\boldsymbol{l}_\alpha$ by

$$\boldsymbol{l}_\alpha \equiv \nabla_{\bar{\mathbf{x}}_\alpha} \log p_\alpha = \frac{\boldsymbol{x}_\alpha - \bar{\boldsymbol{x}}_\alpha}{\varepsilon^2}. \tag{47}$$

Since eq. (43) is an identity in $\{\bar{\boldsymbol{x}}_\alpha\}$ and $\boldsymbol{u}$ that satisfies eq. (4), the variation (44) should vanish for arbitrary variations $\{\delta\bar{\boldsymbol{x}}_\alpha\}$ and $\delta\boldsymbol{u}$ that satisfy eq. (22). Substituting eq. (46) into eq. (44), we conclude that

$$E[(\hat{\boldsymbol{u}} - \boldsymbol{u})\sum_{\alpha=1}^{N} \boldsymbol{l}_\alpha^\top \delta\bar{\boldsymbol{x}}_\alpha] = \delta\boldsymbol{u}, \tag{48}$$

for arbitrary variations $\{\delta\bar{\boldsymbol{x}}_\alpha\}$ and $\delta\boldsymbol{u}$ that satisfy eq. (22).

Consider the following particular variations $\{\delta\bar{\boldsymbol{x}}_\alpha\}$:

$$\delta\bar{\boldsymbol{x}}_\alpha = -\frac{(\nabla_{\mathbf{x}}\bar{F}_\alpha)(\nabla_{\mathbf{u}}\bar{F}_\alpha)^\top}{\|\nabla_{\mathbf{x}}\bar{F}_\alpha\|^2}\delta\boldsymbol{u}. \tag{49}$$

It is easy to confirm that eq. (22) is identically satisfied. Substituting eq. (49) into eq. (48), we obtain

$$E[(\hat{\boldsymbol{u}} - \boldsymbol{u})\sum_{\alpha=1}^{N} \boldsymbol{m}_\alpha^\top]\delta\boldsymbol{u} = -\delta\boldsymbol{u}, \tag{50}$$

where we define the vectors $\{\boldsymbol{m}_\alpha\}$ by

$$\boldsymbol{m}_\alpha = \frac{(\nabla_{\mathbf{u}}\bar{F}_\alpha)(\nabla_{\mathbf{x}}\bar{F}_\alpha)^\top}{\|\nabla_{\mathbf{x}}\bar{F}_\alpha\|^2}\boldsymbol{l}_\alpha. \tag{51}$$

Since eq. (48) should hold for arbitrary variations $\{\delta\bar{\boldsymbol{x}}_\alpha\}$ and $\delta\boldsymbol{u}$ that satisfy eq. (22), eq. (50) should hold for arbitrary *unconstrained* variations $\delta\boldsymbol{u}$, which means

$$E[(\hat{\boldsymbol{u}} - \boldsymbol{u})\sum_{\alpha=1}^{N} \boldsymbol{m}_\alpha^\top] = -\boldsymbol{I}. \tag{52}$$

Using this and recalling the definition (7) of the covariance matrix $V[\hat{\boldsymbol{u}}]$, we obtain

$$E[\begin{pmatrix} \hat{\boldsymbol{u}} - \boldsymbol{u} \\ \sum_{\alpha=1}^{N} \boldsymbol{m}_\alpha \end{pmatrix}\begin{pmatrix} \hat{\boldsymbol{u}} - \boldsymbol{u} \\ \sum_{\alpha=1}^{N} \boldsymbol{m}_\alpha \end{pmatrix}^\top] = \begin{pmatrix} V[\hat{\boldsymbol{u}}] & -\boldsymbol{I} \\ -\boldsymbol{I} & \boldsymbol{M} \end{pmatrix}, \tag{53}$$

where we define the matrix $\boldsymbol{M}$ by

$$\begin{aligned} \boldsymbol{M} &= E[\Big(\sum_{\alpha=1}^{N}\boldsymbol{m}_\alpha\Big)\Big(\sum_{\beta=1}^{N}\boldsymbol{m}_\beta\Big)^\top] \\ &= \sum_{\alpha,\beta=1}^{N} \frac{(\nabla_{\mathbf{u}}\bar{F}_\alpha)(\nabla_{\mathbf{x}}\bar{F}_\alpha)^\top}{\|\nabla_{\mathbf{x}}\bar{F}_\alpha\|^2} E[\boldsymbol{l}_\alpha \boldsymbol{l}_\beta] \frac{(\nabla_{\mathbf{x}}\bar{F}_\alpha)(\nabla_{\mathbf{u}}\bar{F}_\alpha)^\top}{\|\nabla_{\mathbf{x}}\bar{F}_\alpha\|^2} \\ &= \frac{1}{\varepsilon^2}\frac{(\nabla_{\mathbf{u}}\bar{F}_\alpha)(\nabla_{\mathbf{u}}\bar{F}_\alpha)^\top}{\|\nabla_{\mathbf{x}}\bar{F}_\alpha\|^2}. \end{aligned} \tag{54}$$

In the above equation, we use the identity $E[\boldsymbol{l}_\alpha\boldsymbol{l}_\beta^\top] = \delta_{\alpha\beta}\boldsymbol{I}/\varepsilon^4$, which is easily confirmed from eqs. (21) and (47). The matrix $\boldsymbol{J}_\alpha \equiv E[\boldsymbol{l}_\alpha\boldsymbol{l}_\alpha^\top]$ is the *Fisher information matrix* of the distribution p_α and that $E[\boldsymbol{l}_\alpha\boldsymbol{l}_\beta^\top] = \delta_{\alpha\beta}\boldsymbol{J}_\alpha$ if the distributions $\{p_\alpha\}$ are mutually independent.

Since the inside of the expectation $E[\,\cdot\,]$ on the left-hand side of eq. (53) is evidently a positive semidefinite symmetric matrix, so is the right-hand side. It follows the following is also a positive semidefinite symmetric matrix:

$$\begin{aligned} &\begin{pmatrix} \boldsymbol{I} & \boldsymbol{M}^{-1} \\ & \boldsymbol{M}^{-1} \end{pmatrix}\begin{pmatrix} V[\hat{\boldsymbol{u}}] & -\boldsymbol{I} \\ -\boldsymbol{I} & \boldsymbol{M} \end{pmatrix}\begin{pmatrix} \boldsymbol{I} & \\ \boldsymbol{M}^{-1} & \boldsymbol{M}^{-1} \end{pmatrix} \\ &= \begin{pmatrix} V[\hat{\boldsymbol{u}}] - \boldsymbol{M}^{-1} & \\ & \boldsymbol{M}^{-1} \end{pmatrix}. \end{aligned} \tag{55}$$

From this, we conclude that $V[\hat{\boldsymbol{u}}] - \boldsymbol{M}^{-1}$ should be a positive semidefinite symmetric matrix, implying eq. (11). □

The above proof is for the simplest case, but the same result holds for more general cases[14]. If we have multiple constraints, which may not be independent of each other, or if the domains of the data and the parameters are constrained, we can introduce generalized inverses and projection operators to go along the same argument [8]. If the error distribution is not Gaussian or different from datum to datum, the score $\boldsymbol{l}_\alpha$ and the Fisher information matrix $\boldsymbol{J}_\alpha$ take very complicated forms, yet the basic logic remains the same [8].

D: Weighted Least Squares

Comparing eqs. (3) and (31), we can write

$$\tilde{J}_{\rm ML}(\boldsymbol{u}) = c(\boldsymbol{u})J_{\rm ML}(\boldsymbol{u}). \tag{56}$$

If $c(\boldsymbol{u})$ is perturbed into $c(\boldsymbol{u} + \Delta\boldsymbol{u}) = c(\boldsymbol{u}) + (\nabla_{\mathbf{u}}c, \Delta\boldsymbol{u}) + \cdots$, we have $\tilde{J}_{\rm ML}(\boldsymbol{u}+\Delta\boldsymbol{u}) = c(\boldsymbol{u})J_{\rm ML}(\boldsymbol{u}+\Delta\boldsymbol{u}) + O(\varepsilon^3)$, because $J_{\rm ML}$ is $O(\varepsilon^2)$. Hence, differentiation eq. (56) has the form

$$\nabla\tilde{J}_{\rm ML} = c(\boldsymbol{u})\nabla J_{\rm ML} + O(\varepsilon^2). \tag{57}$$

It follows that the solution of $\nabla\tilde{J}_{\rm ML} = \boldsymbol{0}$ and the solution of $\nabla J_{\rm ML} = \boldsymbol{0}$ coincide except for $O(\varepsilon^2)$. Thus, their covariance matrices coincide except for $O(\varepsilon^4)$.

Chernov and Lesort [3] further proved that the solution of the *weighted least squares method* in the form

$$\tilde{J} = \sum_{\alpha=1}^{N} w_\alpha(\boldsymbol{x}_\alpha;\boldsymbol{u})F(\boldsymbol{x}_\alpha;\boldsymbol{u})^2 \to \min \tag{58}$$

is optimal in the sense of the KCR lower bound if and only if

$$w_\alpha(\boldsymbol{x}_\alpha;\boldsymbol{u}) = \frac{c(\boldsymbol{u})}{\|\nabla_{\mathbf{x}}F_\alpha\|^2}. \tag{59}$$

In other words, no forms other than eq. (31) can attain the KCR lower bound in the first order.

[14] However, the description becomes extremely clumsy and cumbersome with a lot of symbols. One of the reasons why the author's theory [8] was doubted or rejected by journals may be that the proof was done in the most general setting.