Proc. 5th Int. Conf. 3-D Digital Imaging and Modeling (3DIM2005), June 2005, Ottawa, Canada, pp. 2–13.

Further Improving Geometric Fitting

Kenichi Kanatani
Department of Computer Science, Okayama University, Okayama, Japan 700-8530
kanatani@suri.it.okayama-u.ac.jp

Abstract

We give a formal definition of geometric fitting *in a way that suits computer vision applications. We point out that the performance of geometric fitting should be evaluated in the limit of* small noise *rather than in the limit of a large number of data as recommended in the statistical literature. Taking the* KCR lower bound *as an optimality requirement and focusing on the linearized constraint case, we compare the accuracy of Kanatani's renormalization with maximum likelihood (ML) approaches including the FNS of Chojnacki et al. and the HEIV of Leedan and Meer. Our analysis reveals the existence of a method superior to all these.*

1. Introduction

By *geometric fitting*, we mean fitting geometric constraints to observed data and discerning the underlying geometric structure from the coefficients of the fitted equations [10]. A large class of computer vision problems fall into this framework. The simplest one 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 resulting 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 exactly satisfy $F(\boldsymbol{x};\boldsymbol{u}) = 0$ and to simultaneously estimate the true positions $\{\bar{\boldsymbol{x}}_\alpha\}$ and the parameter $\boldsymbol{u}$ that maximize the statistical likelihood. If 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 using Lagrange multipliers. Introducing linear approximation by assuming that 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, $\nabla_{\mathbf{x}} F_\alpha$ denotes the gradient of the function $F(\boldsymbol{x};\boldsymbol{u})$ in eq. (1) with respect to $\boldsymbol{x}$ 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 [10].

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 camera 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 equations [8].

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

2. How Can We Compare Methods?

The reason this question is 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 to this. 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 (the *Bayesian approach*). Others may argue that we should rather focus on the error in the application domain, e.g., if

the value $\hat{\boldsymbol{u}}$ is to be used 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 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. Many prefer the mean square because this generally makes the subsequent analysis easy, but other choices are conceivable: some prefer $\max\|\hat{\boldsymbol{u}} - \boldsymbol{u}\|$; others endorse $E[\|\hat{\boldsymbol{u}} - \boldsymbol{u}\|]$. However, the analysis is still intractably 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. Following them, many computer vision researchers analyze asymptotic behavior 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 a 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 noise is random, the effect of noise is expected to be canceled if observations are repeated; the hidden mechanism will reveal 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 [12, 14].

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 by *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, and their estimation accuracy cannot be improved however many data we observe. Such increasing parameters are called *nuisance parameters* to distinguish them from the remaining *structural parameters*. 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. Recently, various theories have been proposed for introducing the *distribution* of the true positions along the curve and marginalizing them over the distribution. Such formulations are called *semiparametric models* [2, 20, 21].

If we have a lot of data, ML is known to be *not* optimal. In fact, Endoh et al. [7] pointed out that 3-D interpretation from a dense optical flow field by ML is not optimal, and Ohta [20] showed that the semiparametric model yields a better result. Okatani and Deguchi [21] 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 extremely large and the problem has a 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?

An answer to this question was given by Kanatani [10, 11]. In the following, we summarize his formulation.

4. KCR Lower Bound

The fundamental difference of Kanatani's approach from the standard statistical estimation is that it focuses on *small noise* rather than asymptotic analysis for a 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 $\{\boldsymbol{x}_\alpha\}$:

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

Such a function $\hat{\boldsymbol{u}}$ is called an *estimator* of $\boldsymbol{u}$. Let us measure the accuracy of estimator $\hat{\boldsymbol{u}}$ by its *covariance matrix*

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

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

Suppose 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*. 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), doing Taylor expansion in $\Delta\boldsymbol{x}_\alpha$ and $\Delta\boldsymbol{u}$ by assuming that 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 [10] (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 $\boldsymbol{u}$ 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 [10] (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 of the left-hand side from the right is positive semidefinite.

Thus, the covariance matrix of the ML estimator $\hat{\boldsymbol{u}}_{\rm ML}$ attains the lower bound except for $O(\varepsilon^4)$. In this sense, ML is optimal. Chernov and Lesort [3] called eq. (11) the *KCR* (*Kanatani-Cramer-Rao*) *lower bound* and derived it under a weaker condition.

The above result can be extended further. First, we need not assume isotropic and identical Gaussian noise. The same argument applies to a wide class of probability distributions called the *exponential family*. If the noise distribution is different from datum to datum, all we need is to introduce covariance matrices $V[\boldsymbol{x}_\alpha]$ in eq. (5). 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. However, the analysis goes similarly if we introduce pseudoinverse and projection operators [10].

5. 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

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

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* holds for an arbitrary unbiased estimator $\hat{\boldsymbol{\theta}}$ (see, e.g., [10] for the proof):

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

It follows that the covariance matrix of the ML estimator $\hat{\boldsymbol{\theta}}_{\rm ML}$ attains the CR lower bound except for $O(1/n^2)$. In this sense, ML is optimal.

6. 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*.

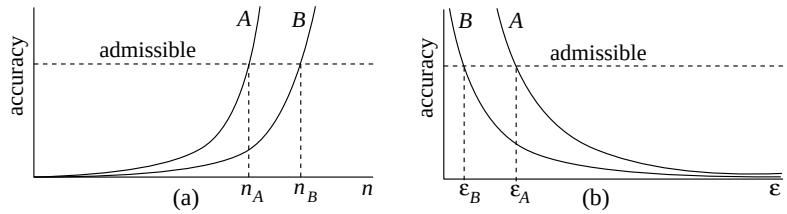

Figure 1. (a) For the standard statistical estimation, it is desired that the accuracy increases rapidly as $n \to \infty$ for the number n of observations, because admissible accuracy can be reached with a smaller number of observations. (b) For geometric fitting, it is desired that the accuracy increases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

Suppose, hypothetically, they change their values each time we observe them (as if in quantum mechanics). 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 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 [13].

This type of duality of interpretation also arises for *model selection*: we obtain the *geometric AIC* and the *geometric MDL* for geometric fitting as counterparts of Akaike's *AIC* (*Akaike information criterion*) [1] and Rissanen's *MDL* (*minimum description length*) [22] for statistical estimation, respectively [13].

7. Linearized Constraints

In many computer vision applications, the constraint (1) can be linearized in the form

$$(\boldsymbol{\xi}(\boldsymbol{x}_\alpha), \boldsymbol{u}) = 0, \tag{16}$$

where $\boldsymbol{\xi}(\,\cdot\,)$ is a (generally nonlinear) mapping from an m-dimensional vector to a p-dimensional vector. In the following, we write $(\boldsymbol{a}, \boldsymbol{b})$ for the inner product of vectors $\boldsymbol{a}$ and $\boldsymbol{b}$. In order to remove scale indeterminacy, we normalize $\boldsymbol{u}$ to $\|\boldsymbol{u}\| = 1$.

Example 1 Suppose we want to fit a quadratic curve (circle, ellipse, parabola, hyperbola, or their degeneracy) to N points $\{(x_\alpha, y_\alpha)\}$, $\alpha = 1, ..., N$, in the plane. The constraint has the form

$$Ax_\alpha^2 + 2Bx_\alpha y_\alpha + Cy_\alpha^2 + 2(Dx_\alpha + Ey_\alpha) + F = 0. \tag{17}$$

If we define

$$\begin{aligned} \boldsymbol{\xi}(x, y) &= (x^2 \;\; 2xy \;\; y^2 \;\; 2x \;\; 2y \;\; 1)^\top, \\ \boldsymbol{u} &= (A \;\; B \;\; C \;\; D \;\; E \;\; F)^\top, \end{aligned} \tag{18}$$

eq. (17) is linearized in the form of eq. (16). □

Example 2 Suppose we have N corresponding points in two images of the same scene viewed from different positions. If point (x_α, y_α) in the first image correspond to (x_α, y_α) in the second, there exists a singular matrix $\boldsymbol{F}$, called the *fundamental matrix*, such that in the absence of noise

$$(\begin{pmatrix} x_\alpha \\ y_\alpha \\ 1 \end{pmatrix}, \boldsymbol{F} \begin{pmatrix} x'_\alpha \\ y'_\alpha \\ 1 \end{pmatrix}) = 0. \tag{19}$$

This is called the *epipolar equation* [8]. If we define

$$\begin{aligned} \boldsymbol{\xi}(x, y, x', y') &= (xx' \;\; xy' \;\; x \;\; yx' \;\; yy' \;\; y \;\; x' \;\; y' \;\; 1)^\top, \\ \boldsymbol{u} &= (F_{11} \;\; F_{12} \;\; F_{13} \;\; F_{21} \;\; F_{22} \;\; F_{23} \;\; F_{31} \;\; F_{32} \;\; F_{33})^\top, \end{aligned} \tag{20}$$

eq. (19) is linearized in the form of eq. (16). □

The KCR lower bound for the linearized constraint (16) has the form

$$V_{\mathrm{KCR}}[\hat{\boldsymbol{u}}] = \varepsilon^2 \Big(\sum_{\alpha=1}^{N} \frac{\bar{\boldsymbol{\xi}}_\alpha \bar{\boldsymbol{\xi}}_\alpha^\top}{(\boldsymbol{u}, V_0[\boldsymbol{\xi}_\alpha]\boldsymbol{u})} \Big)^-, \tag{21}$$

where $(\cdots)^-$ denotes pseudoinverse. The symbol $\bar{\boldsymbol{\xi}}_\alpha$ is an abbreviation for $\boldsymbol{\xi}_\alpha(\bar{\boldsymbol{x}}_\alpha)$, and $V_0[\boldsymbol{\xi}_\alpha]$ is the normalized covariance matrix (scaled so that $\varepsilon = 1$) of $\boldsymbol{\xi}(\boldsymbol{x}_\alpha)$: it can be expressed as

$$V_0[\boldsymbol{\xi}_\alpha] = \nabla_{\mathbf{x}} \boldsymbol{\xi}_\alpha^\top \nabla_{\mathbf{x}} \boldsymbol{\xi}_\alpha \tag{22}$$

except for $O(\varepsilon^4)$, where $\nabla_{\mathbf{x}} \boldsymbol{\xi}_\alpha$ denotes the $m \times p$ Jacobian matrix

$$\nabla_{\mathbf{x}} \boldsymbol{\xi} = \begin{pmatrix} \partial \xi_1 / \partial x_1 & \cdots & \partial \xi_p / \partial x_1 \\ \vdots & \ddots & \vdots \\ \partial \xi_1 / \partial x_m & \cdots & \partial \xi_p / \partial x_m \end{pmatrix}. \tag{23}$$

evaluated at $\boldsymbol{x} = \boldsymbol{x}_\alpha$.

8. In Search of an Optimal Estimator

Now, we try to find an optimal estimator $\hat{\boldsymbol{u}}$ that satisfies the KCR lower bound (21). This is diametrically opposite to the conventional approach of finding some method heuristically and doing analysis or simulation *a posteriori* to see if the bound is indeed attained.

The starting point is the observation that the pseudoinverse on the right-hand of eq. (21) reflects the fact that the vector $\boldsymbol{u}$ in eq. (16) is normalized to $\|\boldsymbol{u}\| = 1$. Hence, its domain is the unit sphere S^{p-1} in $\mathcal{R}^p$, its uncertainty being restricted only in the direction orthogonal to $\hat{\boldsymbol{u}}$. The pseudoinverse $(\cdots)^-$ on the right-hand side of eq. (21) projects $\cdots$ onto the tangent space to S^{p-1} at $\hat{\boldsymbol{u}}$.

It follows that the null space of $V_{\rm KCR}[\hat{\boldsymbol{u}}]$ is in the direction of $\hat{\boldsymbol{u}}$, meaning that $\hat{\boldsymbol{u}}$ is the *unit eigenvector of* $V_{\rm KCR}[\hat{\boldsymbol{u}}]$ *for eigenvalue 0*. Thus, if we know the KCR lower bound $V_{\rm KCR}[\hat{\boldsymbol{u}}]$, we can obtain an optimal estimator $\hat{\boldsymbol{u}}$ as its unit eigenvector for eigenvalue 0.

This appears impossible because $V_{\rm KCR}[\hat{\boldsymbol{u}}]$ involves the true values $\{\bar{\boldsymbol{x}}_\alpha\}$ and $\boldsymbol{u}$, which we do not know. However, this can be overcome by approximating $\{\bar{\boldsymbol{x}}_\alpha\}$ by the data $\{\bar{\boldsymbol{x}}_\alpha\}$ and iteratively estimating $\boldsymbol{u}$. All we need to do is analytically evaluate the error incurred by such approximations. If the error in the resulting covariance matrix is $O(\varepsilon^4)$, we are done.

9. Perturbation Theorem

If we define

$$\boldsymbol{M} = \sum_{\alpha=1}^{N} \frac{\boldsymbol{\xi}_\alpha \boldsymbol{\xi}_\alpha^\top}{(\boldsymbol{u}, V_0[\boldsymbol{\xi}_\alpha]\boldsymbol{u})}. \tag{24}$$

the KCR lower bound (21) is written as

$$V_{\rm KCR}[\hat{\boldsymbol{u}}] = \varepsilon^2 \bar{\boldsymbol{M}}^-, \tag{25}$$

where $\bar{\boldsymbol{M}}$ is the value of $\boldsymbol{M}$ obtained by replacing $\{\boldsymbol{\xi}_\alpha\}$ by their true values $\{\bar{\boldsymbol{\xi}}_\alpha\}$.

Since pseudoinverse preserves the null space, the null space of $\bar{\boldsymbol{M}}$ is the same as that of $\bar{\boldsymbol{M}}^-$, hence of $V_{\rm KCR}[\hat{\boldsymbol{u}}]$. It follows that the unit eigenvector of $V_{\rm KCR}[\hat{\boldsymbol{u}}]$ for eigenvalue 0 is *the unit eigenvector of* $\bar{\boldsymbol{M}}$ *for eigenvalue 0*. In fact, we can directly confirm this: the constraint $(\bar{\boldsymbol{\xi}}_\alpha, \boldsymbol{u}) = 0$ implies

$$\bar{\boldsymbol{M}}\boldsymbol{u} = \sum_{\alpha=1}^{N} \frac{\bar{\boldsymbol{\xi}}_\alpha \bar{\boldsymbol{\xi}}_\alpha^\top \boldsymbol{u}}{(\boldsymbol{u}, V_0[\boldsymbol{\xi}_\alpha]\boldsymbol{u})} = \sum_{\alpha=1}^{N} \frac{(\bar{\boldsymbol{\xi}}_\alpha, \boldsymbol{u})\bar{\boldsymbol{\xi}}_\alpha}{(\boldsymbol{u}, V_0[\boldsymbol{\xi}_\alpha]\boldsymbol{u})} = \boldsymbol{0}. \tag{26}$$

However, we do not know the true matrix $\bar{\boldsymbol{M}}$. So, we approximate it by the matrix $\boldsymbol{M}$ in eq. (24) and evaluate the incurred error. Since $\boldsymbol{M}$ is generally nonsingular, it does not have eigenvalue 0. So, we compute the unit eigenvector of $\boldsymbol{M}$ for the smallest[1] eigenvalue λ and let the solution of

$$\boldsymbol{M}\boldsymbol{u} = \lambda \boldsymbol{u}, \tag{27}$$

be $\hat{\boldsymbol{u}}$. However, the matrix $\boldsymbol{M}$ also involves the unknown $\boldsymbol{u}$, so we do iterations: we compute $\boldsymbol{M}$ using the ith estimate $\boldsymbol{u}_i$ and let the solution of eq. (27) be $\boldsymbol{u}_{i+1}$, i = 0, 1, 2, ..., starting from an initial guess. If the iterations converge, the resulting $\hat{\boldsymbol{u}}$ satisfies eq. (27) (up to the convergence threshold).

Now, we evaluate to what extent the resulting $\hat{\boldsymbol{u}}$ approximates the true $\boldsymbol{u}$. Let

$$\boldsymbol{\xi}_\alpha = \bar{\boldsymbol{\xi}}_\alpha + \Delta\boldsymbol{\xi}_\alpha. \tag{28}$$

The error in $\boldsymbol{M}$ is

$$\begin{aligned}
\Delta\boldsymbol{M} &= \boldsymbol{M} - \bar{\boldsymbol{M}} \\
&= \sum_{\alpha=1}^{N} \frac{(\bar{\boldsymbol{\xi}}_\alpha+\Delta\boldsymbol{\xi}_\alpha)(\bar{\boldsymbol{\xi}}_\alpha+\Delta\boldsymbol{\xi}_\alpha)^\top}{(\boldsymbol{u}, V_0[\boldsymbol{\xi}_\alpha]\boldsymbol{u})} - \sum_{\alpha=1}^{N} \frac{\bar{\boldsymbol{\xi}}_\alpha \bar{\boldsymbol{\xi}}_\alpha^\top}{(\boldsymbol{u}, V_0[\boldsymbol{\xi}_\alpha]\boldsymbol{u})} \\
&= \sum_{\alpha=1}^{N} \frac{\Delta\boldsymbol{\xi}_\alpha \bar{\boldsymbol{\xi}}_\alpha^\top + \bar{\boldsymbol{\xi}}_\alpha \Delta\boldsymbol{\xi}_\alpha^\top}{(\boldsymbol{u}, V_0[\boldsymbol{\xi}_\alpha]\boldsymbol{u})} + \sum_{\alpha=1}^{N} \frac{\Delta\boldsymbol{\xi}_\alpha \Delta\boldsymbol{\xi}_\alpha^\top}{(\boldsymbol{u}, V_0[\boldsymbol{\xi}_\alpha]\boldsymbol{u})} \\
&= \sum_{\alpha=1}^{N} \frac{\Delta\boldsymbol{\xi}_\alpha \bar{\boldsymbol{\xi}}_\alpha^\top + \bar{\boldsymbol{\xi}}_\alpha \Delta\boldsymbol{\xi}_\alpha^\top}{(\boldsymbol{u}, V_0[\boldsymbol{\xi}_\alpha]\boldsymbol{u})} + O(\varepsilon^2).
\end{aligned} \tag{29}$$

According to the *perturbation theorem*, the perturbation of $\bar{\boldsymbol{M}}$ into $\bar{\boldsymbol{M}} - \Delta\boldsymbol{M}$ induces the perturbation of $\boldsymbol{u}$ as follows [10]:

$$\hat{\boldsymbol{u}} = \boldsymbol{u} + \bar{\boldsymbol{M}}^- \Delta\boldsymbol{M}\boldsymbol{u} + O(\varepsilon^2). \tag{30}$$

Its covariance matrix is evaluated as follows:

$$\begin{aligned}
V[\hat{\boldsymbol{u}}] &= E[(\hat{\boldsymbol{u}} - \boldsymbol{u})(\hat{\boldsymbol{u}} - \boldsymbol{u})^\top] \\
&= E[\bar{\boldsymbol{M}}^- \Delta\boldsymbol{M}\boldsymbol{u}\boldsymbol{u}^\top \Delta\boldsymbol{M}\bar{\boldsymbol{M}}^-] + O(\varepsilon^4) \\
&= E[\bar{\boldsymbol{M}}^- \sum_{\alpha=1}^{N} \frac{\Delta\boldsymbol{\xi}_\alpha \bar{\boldsymbol{\xi}}_\alpha^\top + \bar{\boldsymbol{\xi}}_\alpha \Delta\boldsymbol{\xi}_\alpha^\top}{(\boldsymbol{u}, V_0[\boldsymbol{\xi}_\alpha]\boldsymbol{u})} \boldsymbol{u}\boldsymbol{u}^\top \\
&\qquad \sum_{\beta=1}^{N} \frac{\Delta\boldsymbol{\xi}_\beta \bar{\boldsymbol{\xi}}_\beta^\top + \bar{\boldsymbol{\xi}}_\beta \Delta\boldsymbol{\xi}_\beta^\top}{(\boldsymbol{u}, V_0[\boldsymbol{\xi}_\beta]\boldsymbol{u})} \bar{\boldsymbol{M}}^-] + O(\varepsilon^4) \\
&= E[\bar{\boldsymbol{M}}^- \sum_{\alpha=1}^{N} \frac{(\Delta\boldsymbol{\xi}_\alpha, \boldsymbol{u})(\Delta\boldsymbol{\xi}_\beta, \boldsymbol{u})\bar{\boldsymbol{\xi}}_\alpha \bar{\boldsymbol{\xi}}_\beta^\top}{(\boldsymbol{u}, V_0[\boldsymbol{\xi}_\alpha]\boldsymbol{u})(\boldsymbol{u}, V_0[\boldsymbol{\xi}_\beta]\boldsymbol{u})} \bar{\boldsymbol{M}}^-] + O(\varepsilon^4) \\
&= \bar{\boldsymbol{M}}^- \sum_{\alpha,\beta=1}^{N} \frac{(\boldsymbol{u}, E[\Delta\boldsymbol{\xi}_\alpha \Delta\boldsymbol{\xi}_\beta^\top]\boldsymbol{u})\bar{\boldsymbol{\xi}}_\alpha \bar{\boldsymbol{\xi}}_\beta^\top}{(\boldsymbol{u}, V_0[\boldsymbol{\xi}_\alpha]\boldsymbol{u})(\boldsymbol{u}, V_0[\boldsymbol{\xi}_\beta]\boldsymbol{u})} \bar{\boldsymbol{M}}^- + O(\varepsilon^4) \\
&= \bar{\boldsymbol{M}}^- \sum_{\alpha,\beta=1}^{N} \frac{(\boldsymbol{u}, \varepsilon^2 \delta_{\alpha\beta} V_0[\boldsymbol{\xi}_\alpha]\boldsymbol{u})\bar{\boldsymbol{\xi}}_\alpha \bar{\boldsymbol{\xi}}_\beta^\top}{(\boldsymbol{u}, V_0[\boldsymbol{\xi}_\alpha]\boldsymbol{u})(\boldsymbol{u}, V_0[\boldsymbol{\xi}_\beta]\boldsymbol{u})} \bar{\boldsymbol{M}}^- + O(\varepsilon^4) \\
&= \bar{\boldsymbol{M}}^- \sum_{\alpha=1}^{N} \frac{\varepsilon^2 \bar{\boldsymbol{\xi}}_\alpha \bar{\boldsymbol{\xi}}_\alpha^\top}{(\boldsymbol{u}, V_0[\boldsymbol{\xi}_\alpha]\boldsymbol{u})} \bar{\boldsymbol{M}}^- + O(\varepsilon^4)
\end{aligned}$$

[1]The matrix $\boldsymbol{M}$ is positive semidefinite by construction, so its eigenvalues are all nonnegative.

$$= \varepsilon^2 \bar{\boldsymbol{M}}^- \bar{\boldsymbol{M}} \bar{\boldsymbol{M}}^- + O(\varepsilon^4) = \varepsilon^2 \bar{\boldsymbol{M}}^- + O(\varepsilon^4)$$
$$= V_{\mathrm{KCR}}[\hat{\boldsymbol{u}}] + O(\varepsilon^4). \tag{31}$$

Here, $\delta_{\alpha\beta}$ is the Kronecker delta, taking 1 for $\alpha = \beta$ and 0 otherwise. In the above derivation, we use the equality $E[\Delta\boldsymbol{\xi}_\alpha \Delta\boldsymbol{\xi}_\beta^\top] = \delta_{\alpha\beta} V_0[\boldsymbol{\xi}_\alpha]$, which follows from the assumption that noise in each $\boldsymbol{x}_\alpha$ is independent. The remainder term is $O(\varepsilon^4)$. This is a consequence of the fact that the noise distribution is symmetric with respect to the origin, hence terms of all odd degrees in ε vanish in expectation.

Thus, we find that the unit eigenvector $\hat{\boldsymbol{u}}$ of $\boldsymbol{M}$ in eq. (24) for the smallest eigenvalue is *optimal* in the sense that its covariance matrix attains the KCR lower bound $V_{\mathrm{KCR}}[\hat{\boldsymbol{u}}]$ except for $O(\varepsilon^4)$.

10. Bias Removal

Not being satisfied with this, let us go further. Can this be what we could do? Can't we further improve the accuracy?

The annoying fact is that the second term $\bar{\boldsymbol{M}}^- \Delta\boldsymbol{M}\boldsymbol{u}$ on the right-hand side of eq. (30) is not zero in expectation, i.e., it has statistical bias. Eq. (29) implies that $\Delta\boldsymbol{M}$ is unbiased except for $O(\varepsilon^2)$, but if we do not ignore $O(\varepsilon^2)$, we see that

$$E[\Delta\boldsymbol{M}] = \sum_{\alpha=1}^N \frac{E[\Delta\boldsymbol{\xi}_\alpha]\bar{\boldsymbol{\xi}}_\alpha^\top + \bar{\boldsymbol{\xi}}_\alpha E[\Delta\boldsymbol{\xi}_\alpha^\top]}{(\boldsymbol{u}, V_0[\boldsymbol{\xi}_\alpha]\boldsymbol{u})} + \sum_{\alpha=1}^N \frac{E[\Delta\boldsymbol{\xi}_\alpha \Delta\boldsymbol{\xi}_\alpha^\top]}{(\boldsymbol{u}, V_0[\boldsymbol{\xi}_\alpha]\boldsymbol{u})}$$
$$= \sum_{\alpha=1}^N \frac{\varepsilon^2 V_0[\boldsymbol{\xi}_\alpha]}{(\boldsymbol{u}, V_0[\boldsymbol{\xi}_\alpha]\boldsymbol{u})} = \varepsilon^2 \boldsymbol{N}, \tag{32}$$

where we define

$$\boldsymbol{N} = \sum_{\alpha=1}^N \frac{V_0[\boldsymbol{\xi}_\alpha]}{(\boldsymbol{u}, V_0[\boldsymbol{\xi}_\alpha]\boldsymbol{u})}. \tag{33}$$

Hence, the expectation of eq. (30) is

$$E[\hat{\boldsymbol{u}}] = \boldsymbol{u} + \varepsilon^2 \bar{\boldsymbol{M}}^- \boldsymbol{N}\boldsymbol{u} + O(\varepsilon^2). \tag{34}$$

Can we remove the term $\varepsilon^2 \bar{\boldsymbol{M}}^- \boldsymbol{N}\boldsymbol{u}$?

After careful examinations, we find that this can be done if eq. (24) is replaced by

$$\hat{\boldsymbol{M}} = \boldsymbol{M} - \varepsilon^2 \boldsymbol{N}. \tag{35}$$

If we let $\hat{\boldsymbol{u}}$ be the unit eigenvector of $\hat{\boldsymbol{M}}$ for the smallest eigenvalue, eq. (30) is replaced by

$$\hat{\boldsymbol{u}} = \boldsymbol{u} + \bar{\boldsymbol{M}}^- \Delta\hat{\boldsymbol{M}}\boldsymbol{u} + O(\varepsilon^2), \tag{36}$$

doing the same perturbation analysis, and

$$E[\Delta\hat{\boldsymbol{M}}] = E[\hat{\boldsymbol{M}} - \bar{\boldsymbol{M}}] = E[\boldsymbol{M} - \bar{\boldsymbol{M}} - \varepsilon^2 \boldsymbol{N}] = \boldsymbol{O}. \tag{37}$$

This does not affect the fact that the covariance matrix attains the KCR lower bound except for $O(\varepsilon^4)$, because in eq. (31) the difference between $\Delta\boldsymbol{M}$ and $\Delta\hat{\boldsymbol{M}}$ is absorbed in the remainder term $O(\varepsilon^4)$.

11. Renormalization

Since the noise level ε on the right-hand side of eq. (35) is unknown, we need to estimate it. This is easily done by choosing the value of ε^2 in eq. (35) so that $\hat{\boldsymbol{M}}$ has eigenvalue 0. Suppose we use a tentative value ε^2, and let λ be the smallest (in absolute value) eigenvalue of $\hat{\boldsymbol{M}}$ with the unit eigenvector $\hat{\boldsymbol{u}}$. If $\lambda \neq 0$, we increment the current ε^2 by c so that $(\hat{\boldsymbol{M}} - c\boldsymbol{N})\hat{\boldsymbol{u}} = \boldsymbol{0}$, or

$$(\hat{\boldsymbol{u}}, (\hat{\boldsymbol{M}} - c\boldsymbol{N})\hat{\boldsymbol{u}}) = (\hat{\boldsymbol{u}}, \hat{\boldsymbol{M}}\hat{\boldsymbol{u}}) - c(\hat{\boldsymbol{u}}, \boldsymbol{N}\hat{\boldsymbol{u}})$$
$$= \lambda - c(\hat{\boldsymbol{u}}, \boldsymbol{N}\hat{\boldsymbol{u}}) = 0. \tag{38}$$

Hence, $c = \lambda/(\hat{\boldsymbol{u}}, \boldsymbol{N}\hat{\boldsymbol{u}})$. We iterate this process until $\lambda \approx 0$. If we incorporate this iteration into the computation of the eigenvector $\boldsymbol{u}$ of $\hat{\boldsymbol{M}}$, we obtain the following scheme:

1. Guess an initial value $\boldsymbol{u}_0$, and let $c_0 = 0$.
2. Letting $\boldsymbol{M}_{i-1}$ and $\boldsymbol{N}_{i-1}$ be, respectively, the matrices $\boldsymbol{M}$ and $\boldsymbol{N}$ in eqs. (24) and (33) computed using the $(i-1)$th estimate $\boldsymbol{u}_{i-1}$, solve the eigenvalue problem
$$(\boldsymbol{M}_{i-1} - c_{i-1}\boldsymbol{N}_{i-1})\boldsymbol{u} = \lambda\boldsymbol{u}, \tag{39}$$
and let $\boldsymbol{u}_i$ be the unit eigenvector for the smallest (in absolute value) eigenvalue.
3. If λ is sufficiently close to 0, stop and return $\boldsymbol{u}_i$ as $\hat{\boldsymbol{u}}$. Else, let
$$c_i \leftarrow c_{i-1} + \frac{\lambda}{(\boldsymbol{u}_i, \boldsymbol{N}(\boldsymbol{u}_{i-1})\boldsymbol{u}_i)}, \tag{40}$$
and go pack to Step 2 after letting $\boldsymbol{u}_{i-1} \leftarrow \boldsymbol{u}_i$.

This is nothing but the *renormalization* of Kanatani [9, 10, 14], though he introduced this by an intuition different from the above reasoning.

If the renormalization iterations converge, we have $(\boldsymbol{M} - c\boldsymbol{N})\hat{\boldsymbol{u}} = \boldsymbol{0}$. Computing the inner product with $\hat{\boldsymbol{u}}$ on both sides, we have

$$(\hat{\boldsymbol{u}}, (\boldsymbol{M} - c\boldsymbol{N})\hat{\boldsymbol{u}}) = (\hat{\boldsymbol{u}}, \boldsymbol{M}\hat{\boldsymbol{u}}) - c(\hat{\boldsymbol{u}}, \boldsymbol{N}\hat{\boldsymbol{u}}) = 0. \tag{41}$$

Hence,

$$c = \frac{(\hat{\boldsymbol{u}}, \boldsymbol{M}\hat{\boldsymbol{u}})}{(\hat{\boldsymbol{u}}, \boldsymbol{N}\hat{\boldsymbol{u}})}. \tag{42}$$

It is difficult to evaluate the expectation of c exactly, because $\hat{\boldsymbol{u}}$ depends on not only $\boldsymbol{M}$ but also c itself. However, if we let $\hat{\boldsymbol{u}} \approx \boldsymbol{u}$ to a first approximation, we obtain

$$E[c] = \frac{(\boldsymbol{u}, E[\boldsymbol{M}]\boldsymbol{u})}{(\boldsymbol{u}, \boldsymbol{N}\boldsymbol{u})} = \frac{(\boldsymbol{u}, \varepsilon^2 \boldsymbol{N}\boldsymbol{u})}{(\boldsymbol{u}, \boldsymbol{N}\boldsymbol{u})} = \varepsilon^2. \tag{43}$$

If $\hat{\boldsymbol{u}}$ is a good approximation of $\boldsymbol{u}$, which is usually the case, the error in the above approximation is expected to be a higher order term $O(\varepsilon^4)$. Thus,

$$E[\hat{\boldsymbol{M}} - \bar{\boldsymbol{M}}] = \varepsilon^2 \boldsymbol{N} - E[c]\boldsymbol{N} = O(\varepsilon^4). \tag{44}$$

The initial guess $\boldsymbol{u}_0$ is given, for example, by the unit eigenvector of

$$\boldsymbol{M}_{\rm LS} = \sum_{\alpha=1}^{N} \boldsymbol{\xi}_\alpha \boldsymbol{\xi}_\alpha^\top \tag{45}$$

for the smallest eigenvalue. This is simply the least-square method (2), for it minimizes the sum of squares

$$J_{\rm LS} = \sum_{\alpha=1}^{N} (\boldsymbol{\xi}_\alpha, \boldsymbol{u})^2. \tag{46}$$

12. Controversies about Renormalization

Kanatani's renormalization turned out to produce highly accurate values in many computer vision applications, and it is now an indispensable tools for computing the fundamental matrix and homographies for 3-D reconstruction and image mosaicing applications [16, 17]. It is used across the world and incorporated in some commercial products, too.

However, questions and doubts have constantly been raised about its interpretation. This is because Kanatani introduced renormalization as a *bias removal procedure* [9, 10, 14]. But, if bias removal is the sole purpose, why don't we start with the matrix $\boldsymbol{M}_{\rm LS}$?

Kanatani endorsed the use of the matrix $\boldsymbol{M}$ in eq. (24) in an analogy with ML. Extending this view, Chojnacki et al. [4] asserted that renormalization is an approximate method for ML and proposed a new method called *FNS* (*fundamental numerical scheme*) for directly computing ML [5]. They also pointed out that in this respect the *HEIV* (*heteroscedastic errors-in-variables*) of Leedan and Meer [18] falls in the same category [6], too. From these observations, Chojnacki et al. [4] asserted superiority of the FNS and the HEIV over renormalization.

From the description in the preceding section, however, it is now evident that *renormalization has nothing to do with ML*. The use of the matrix $\boldsymbol{M}$ is justified only by realizing that what we really want is *the eigenvector of the KCR lower bound*. The only link of renormalization with ML is the fact that the ML estimator also satisfies the KCR bound in the leading term [15].

Thus, Kanatani's renormalization is justified in this new light. However, a new question arises. Why is removing the second term on the right-hand side of eq. (34) effective? The right-hand side has the remainder term $O(\epsilon^2)$ after all. Since the removed term is also $O(\epsilon^2)$, the order of approximation does not change.

Yet, it has been proven by simulations and real data experiments that the removal of that term results in significant improvement of accuracy (see, e.g., [16, 17]). It has also been confirmed that the accuracy of renormalization is practically comparable to the FNS the HEIV. Is this a miraculous coincidence[2]?

Evidently, we couldn't answer this question as long as we are restricted to first order analysis: we are forced to do second order analysis.

13. Second Order Perturbation

We now write

$$\boldsymbol{M} = \bar{\boldsymbol{M}} + \Delta_1 \boldsymbol{M} + \Delta_2 \boldsymbol{M}, \tag{47}$$

where Δ_1 and Δ_2 designate perturbations of orders $O(\varepsilon)$ and $O(\varepsilon^2)$, respectively. From the derivation of eq. (29), we find that

$$\Delta_1 \boldsymbol{M} = \sum_{\alpha=1}^{N} \frac{\Delta\boldsymbol{\xi}_\alpha \bar{\boldsymbol{\xi}}_\alpha^\top + \bar{\boldsymbol{\xi}}_\alpha \Delta\boldsymbol{\xi}_\alpha^\top}{(\boldsymbol{u}, V_0[\boldsymbol{\xi}_\alpha]\boldsymbol{u})}, \tag{48}$$

$$\Delta_2 \boldsymbol{M} = \sum_{\alpha=1}^{N} \frac{\Delta\boldsymbol{\xi}_\alpha \Delta\boldsymbol{\xi}_\alpha^\top}{(\boldsymbol{u}, V_0[\boldsymbol{\xi}_\alpha]\boldsymbol{u})}. \tag{49}$$

Eq. (27) can be written in the form

$$\begin{aligned}&(\bar{\boldsymbol{M}} + \Delta_1 \boldsymbol{M} + \Delta_2 \boldsymbol{M})(\boldsymbol{u} + \Delta_1 \boldsymbol{u} + \Delta_2 \boldsymbol{u} + \cdots)\\ &= (\Delta_1 \lambda + \Delta_2 \lambda + \cdots)(\boldsymbol{u} + \Delta_1 \boldsymbol{u} + \Delta_2 \boldsymbol{u} + \cdots).\end{aligned} \tag{50}$$

Comparing terms of $O(1)$, $O(\varepsilon)$, and $O(\varepsilon^2)$ on both sides, we obtain the following expressions (see Appendix D for the derivation):

$$\Delta_1 \boldsymbol{u} = -\bar{\boldsymbol{M}}^- \Delta_1 \boldsymbol{M} \boldsymbol{u}, \tag{51}$$

$$\begin{aligned}\Delta_2 \boldsymbol{u} = &-\bar{\boldsymbol{M}}^- \Delta_2 \boldsymbol{M} \boldsymbol{u} + \bar{\boldsymbol{M}}^- \Delta_1 \boldsymbol{M} \bar{\boldsymbol{M}}^- \Delta_1 \boldsymbol{M} \boldsymbol{u}\\ &-\|\bar{\boldsymbol{M}}^- \Delta_1 \boldsymbol{M} \boldsymbol{u}\|^2 \boldsymbol{u}.\end{aligned} \tag{52}$$

Since $E[\Delta_1 \boldsymbol{M}] = \boldsymbol{O}$, we have $E[\Delta_1 \boldsymbol{u}] = \boldsymbol{0}$. Since $E[\Delta_2 \boldsymbol{M}] = \varepsilon^2 \boldsymbol{N}$, the expectation of $\Delta_2 \boldsymbol{u}$ is

$$\begin{aligned}E[\Delta_2 \boldsymbol{u}] = &-\varepsilon^2 \bar{\boldsymbol{M}}^- \boldsymbol{N} \boldsymbol{u} + \bar{\boldsymbol{M}}^- E[\Delta_1 \boldsymbol{M} \bar{\boldsymbol{M}}^- \Delta_1 \boldsymbol{M}] \boldsymbol{u}\\ &-E[\|\bar{\boldsymbol{M}}^- \Delta_1 \boldsymbol{M} \boldsymbol{u}\|^2] \boldsymbol{u}.\end{aligned} \tag{53}$$

We can show $E[\|\bar{\boldsymbol{M}}^- \Delta_1 \boldsymbol{M} \boldsymbol{u}\|^2] = \varepsilon^2 {\rm tr}(\bar{\boldsymbol{M}}^-)$ (see Appendix E for the proof). Since renormalization removes the bias term $-\varepsilon^2 \bar{\boldsymbol{M}}^- \boldsymbol{N} \boldsymbol{u}$, the renormalization solution $\hat{\boldsymbol{u}}_{\rm RN}$ has the following expectation:

$$\begin{aligned}E[\hat{\boldsymbol{u}}_{\rm RN}] = &\boldsymbol{u} + \bar{\boldsymbol{M}}^- E[\Delta_1 \boldsymbol{M} \bar{\boldsymbol{M}}^- \Delta_1 \boldsymbol{M}] \boldsymbol{u}\\ &-\varepsilon^2 {\rm tr}(\bar{\boldsymbol{M}}^-) \boldsymbol{u} + O(\varepsilon^4).\end{aligned} \tag{54}$$

14. Errors in Maximum Likelihood

We now compare eq. (54) with ML. For the linearized constraint (16), the ML minimization (5) reduces to

$$J_{\rm ML} = \sum_{\alpha=1}^{N} \frac{(\boldsymbol{\xi}_\alpha, \boldsymbol{u})^2}{(\boldsymbol{u}, V_0[\boldsymbol{\xi}_\alpha]\boldsymbol{u})} \to \min. \tag{55}$$

[2] Nikolai Chernov, the author of [3], described so in a personal communication with the author.

This is an approximation to the true ML of eq. (3) for small noise, but since we are concerned with perturbation for small ε (recall the arguments in Sec. 6), we call this simply ML. The FNS of Chojnacki et al. [5] the HEIV of Leedan and Meer [18], and the recent method of Mühlich and Mester [19], which is a variant of the method known as *equilibration* or *whitening*, all aim to minimize eq. (55).

Differentiating $J_{\rm ML}$ with respect to $\boldsymbol{u}$, we obtain

$$\nabla_{\mathbf{u}} J_{\rm ML} = \sum_{\alpha=1}^{N} \frac{2(\boldsymbol{\xi}_\alpha, \boldsymbol{u})\boldsymbol{\xi}_\alpha}{(\boldsymbol{u}, V_0[\boldsymbol{\xi}_\alpha]\boldsymbol{u})} - \sum_{\alpha=1}^{N} \frac{2(\boldsymbol{\xi}_\alpha, \boldsymbol{u})^2 V_0[\boldsymbol{\xi}_\alpha]\boldsymbol{u}}{(\boldsymbol{u}, V_0[\boldsymbol{\xi}_\alpha]\boldsymbol{u})^2}. \quad (56)$$

Hence, the ML estimator $\hat{\boldsymbol{u}}_{\rm ML}$ is the solution of

$$\boldsymbol{M}\hat{\boldsymbol{u}} = \boldsymbol{L}\hat{\boldsymbol{u}}, \quad (57)$$

where we define

$$\boldsymbol{M} = \sum_{\alpha=1}^{N} \frac{\boldsymbol{\xi}_\alpha \boldsymbol{\xi}_\alpha^\top}{(\boldsymbol{u}, V_0[\boldsymbol{\xi}_\alpha]\boldsymbol{u})}, \quad \boldsymbol{L} = \sum_{\alpha=1}^{N} \frac{(\boldsymbol{\xi}_\alpha, \boldsymbol{u})^2 V_0[\boldsymbol{\xi}_\alpha]}{(\boldsymbol{u}, V_0[\boldsymbol{\xi}_\alpha]\boldsymbol{u})^2}. \quad (58)$$

The FNS and the HEIV both solve eq. (57) by iterations.

One may wonder if eq. (56) vanishes only in the direction orthogonal to $\boldsymbol{u}$ because the minimization (55) should be subject to the normalization $\|\boldsymbol{u}\| = 1$. However, eq. (55) is a *homogeneous form of degree 0* in $\boldsymbol{u}$ and hence is invariant to scale change of $\boldsymbol{u}$. It follows that $\nabla_{\mathbf{u}} J_{\rm ML}$ is identically 0 in the direction of $\boldsymbol{u}$, hence 0 in all directions [5].

The perturbation of $\bar{\boldsymbol{M}}$ is written in the form of eqs. (47)$\sim$(49). For $\boldsymbol{L}$, we observe

$$\begin{aligned}\boldsymbol{L} &= \sum_{\alpha=1}^{N} \frac{(\bar{\boldsymbol{\xi}}_\alpha + \Delta\boldsymbol{\xi}_\alpha, \boldsymbol{u})^2 V_0[\boldsymbol{\xi}_\alpha]}{(\boldsymbol{u}, V_0[\boldsymbol{\xi}_\alpha]\boldsymbol{u})^2} = \frac{(\Delta\boldsymbol{\xi}_\alpha, \boldsymbol{u})^2 V_0[\boldsymbol{\xi}_\alpha]}{(\boldsymbol{u}, V_0[\boldsymbol{\xi}_\alpha]\boldsymbol{u})^2} \\ &= \Delta_2 \boldsymbol{L}. \end{aligned} \quad (59)$$

In other words, $\boldsymbol{L}$ is $O(\varepsilon^2)$ from the beginning, so eq. (57) is written in the form

$$\begin{aligned}&(\bar{\boldsymbol{M}} + \Delta_1\boldsymbol{M} + \Delta_2\boldsymbol{M})(\boldsymbol{u} + \Delta_1\boldsymbol{u} + \Delta_2\boldsymbol{u} + \cdots) \\ &= \Delta_2\boldsymbol{L}(\boldsymbol{u} + \Delta_1\boldsymbol{u} + \Delta_2\boldsymbol{u} + \cdots). \end{aligned} \quad (60)$$

Comparing terms of $O(1)$, $O(\varepsilon)$, and $O(\varepsilon^2)$ on both sides, we obtain the following expressions (see Appendix F for the derivation):

$$\Delta_1\boldsymbol{u} = -\bar{\boldsymbol{M}}^- \Delta_1\boldsymbol{M}\boldsymbol{u}, \quad (61)$$

$$\begin{aligned}\Delta_2\boldsymbol{u} = &-\bar{\boldsymbol{M}}^- \Delta_2\boldsymbol{M}\boldsymbol{u} + \bar{\boldsymbol{M}}^- \Delta_1\boldsymbol{M}\bar{\boldsymbol{M}}^- \Delta_1\boldsymbol{M}\boldsymbol{u} \\ &+\bar{\boldsymbol{M}}^- \Delta_2\boldsymbol{L}\boldsymbol{u} - \|\bar{\boldsymbol{M}}^- \Delta_1\boldsymbol{M}\boldsymbol{u}\|^2 \boldsymbol{u}. \end{aligned} \quad (62)$$

We have already seen that $E[\Delta_1\boldsymbol{u}] = \boldsymbol{0}$ and $E[\Delta_2\boldsymbol{M}] = \varepsilon\boldsymbol{N}$. From eq. (59), the expectation of $\Delta_2\boldsymbol{L}$ is

$$\begin{aligned}E[\Delta_2\boldsymbol{L}] &= E[\sum_{\alpha=1}^{N} \frac{(\Delta\boldsymbol{\xi}_\alpha, \boldsymbol{u})^2 V_0[\boldsymbol{\xi}_\alpha]}{(\boldsymbol{u}, V_0[\boldsymbol{\xi}_\alpha]\boldsymbol{u})^2}] \\ &= \sum_{\alpha=1}^{N} \frac{(\boldsymbol{u}, E[\Delta\boldsymbol{\xi}_\alpha \Delta\boldsymbol{\xi}_\alpha^\top]\boldsymbol{u}) V_0[\boldsymbol{\xi}_\alpha]}{(\boldsymbol{u}, V_0[\boldsymbol{\xi}_\alpha]\boldsymbol{u})^2} \\ &= \sum_{\alpha=1}^{N} \frac{(\boldsymbol{u}, \varepsilon^2 V_0[\boldsymbol{\xi}_\alpha]\boldsymbol{u}) V_0[\boldsymbol{\xi}_\alpha]}{(\boldsymbol{u}, V_0[\boldsymbol{\xi}_\alpha]\boldsymbol{u})^2} = \varepsilon^2 \sum_{\alpha=1}^{N} \frac{V_0[\boldsymbol{\xi}_\alpha]}{(\boldsymbol{u}, V_0[\boldsymbol{\xi}_\alpha]\boldsymbol{u})} \\ &= \varepsilon^2 \boldsymbol{N}. \end{aligned} \quad (63)$$

Thus, the expectation of the ML estimator $\hat{\boldsymbol{u}}_{\rm ML}$ is

$$\begin{aligned}E[\hat{\boldsymbol{u}}_{\rm ML}] = \boldsymbol{u} &+ \bar{\boldsymbol{M}}^- E[\Delta_1\boldsymbol{M}\bar{\boldsymbol{M}}^- \Delta_1\boldsymbol{M}]\boldsymbol{u} \\ &-\varepsilon^2 {\rm tr}(\bar{\boldsymbol{M}}^-)\boldsymbol{u} + O(\varepsilon^4). \end{aligned} \quad (64)$$

This coincides with eq. (54).

15. Toward Further Improvement

Our conclusions are summarized as follows:

1. Renormalization is not an approximate solution technique for ML. It is to compute the solution that satisfies the KCR lower bound followed by removal of one of the $O(\varepsilon^2)$ bias terms.
2. The difference of the renormalization solution $\hat{\boldsymbol{u}}_{\rm RN}$ from the ML estimator $\hat{\boldsymbol{u}}_{\rm ML}$ is in expectation
$$E[\hat{\boldsymbol{u}}_{\rm RN} - \hat{\boldsymbol{u}}_{\rm ML}] = O(\varepsilon^4). \quad (65)$$
3. The covariance matrices $V[\hat{\boldsymbol{u}}_{\rm RN}]$ and $V[\hat{\boldsymbol{u}}_{\rm LM}]$ of the renormalization solution $\boldsymbol{u}_{\rm RN}$ and the ML estimator $\boldsymbol{u}_{\rm ML}$ both attain the KCR lower bound $V_{\rm KCR}[\hat{\boldsymbol{u}}]$ except for $O(\varepsilon^4)$.
$$\begin{aligned}V[\hat{\boldsymbol{u}}_{\rm RN}] &= V_{\rm KCR}[\hat{\boldsymbol{u}}] + O(\varepsilon^4), \\ V[\hat{\boldsymbol{u}}_{\rm ML}] &= V_{\rm KCR}[\hat{\boldsymbol{u}}] + O(\varepsilon^4). \end{aligned} \quad (66)$$
4. The covariance matrices $V[\hat{\boldsymbol{u}}_{\rm RN}]$ and $V[\hat{\boldsymbol{u}}_{\rm ML}]$ coincide except for $O(\varepsilon^6)$.
$$V[\hat{\boldsymbol{u}}_{\rm RN}] = V[\hat{\boldsymbol{u}}_{\rm ML}] + O(\varepsilon^6). \quad (67)$$
5. The renormalization solution $\hat{\boldsymbol{u}}_{\rm RN}$ and the ML estimator $\hat{\boldsymbol{u}}_{\rm ML}$ share a common error term $\bar{\boldsymbol{M}}^- E[\Delta_1\boldsymbol{M}\bar{\boldsymbol{M}}^- \Delta_1\boldsymbol{M}]\boldsymbol{u}$.

The last fact implies that we can obtain a method superior to both renormalization and ML by estimating and subtracting that error term. This is currently under investigation.

It seems that one of the reasons this type of analysis has not been attempted in the past is that computer vision researchers are likely to take textbooks of statistics 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 their applications. This paper demonstrates how promising such an attempt can be.

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 and A. van den Hengel, Rationalising the renormalisation method of Kanatani, *J. Math. Imaging Vision*, **14**-1 (2001), 21–38.

[5] 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.

[6] W. Chojnacki, M. J. Brooks, A. van den Hengel and D. Gawley, From FNS to HEIV: A link between two vision parameter estimation methods, *IEEE Trans. Patt. Anal. Mach. Intell.*, **26**-2 (2004-2), 264–268.

[7] 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 (1994-11), 1240–1246.

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

[9] K. Kanatani, Renormalization for unbiased estimation, *Proc. 4th Int. Conf. Comput. Vision (ICCV'93)*, May 1993, Berlin, Germany, pp. 599–606.

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

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

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

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

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

[15] K. Kanatani, Optimality of maximum likelihood estimation for geometric fitting and the KCR lower bound, *Mem. Fac. Eng. Okayama Univ.*, **39** (2005), 63–70.

[16] K. Kanatani and N. Ohta, Comparing optimal three-dimensional reconstruction for finite motion and optical flow, *J. Electronic Imaging*, **12**-3 (2003), 478–488.

[17] K. Kanatani, N. Ohta and Y. Kanazawa, Optimal homography computation with a reliability measure, *IEICE Trans. Inf. & Sys.*, **E83-D**-7 (2000-7), 1369–1374.

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

[19] 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.

[20] 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`

[21] 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.

[22] 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 the linear approximation

$$F_\alpha - (\nabla_{\mathbf{x}} F_\alpha, \Delta\boldsymbol{x}_\alpha) = 0. \tag{68}$$

Introducing Lagrange multipliers λ_α, let

$$L = \frac{1}{2}\sum_{\alpha=1}^{N} \|\Delta\boldsymbol{x}_\alpha\|^2 + \sum_{\alpha=1}^{N} \lambda_\alpha (F_\alpha - (\nabla_{\mathbf{x}} F_\alpha, \Delta\boldsymbol{x}_\alpha)). \tag{69}$$

The solution $\Delta\boldsymbol{x}_\alpha$ that minimizes L subject to the constraint (68) satisfies $\nabla_{\Delta\mathbf{x}_\alpha} L = \mathbf{0}$, $\alpha = 1, ..., N$, or

$$\Delta\boldsymbol{x}_\alpha - \lambda_\alpha \nabla_{\mathbf{x}} F_\alpha = \mathbf{0}. \tag{70}$$

Hence, $\Delta\boldsymbol{x}_\alpha = \lambda_\alpha \nabla_{\mathbf{x}} F_\alpha$. Substitution of this into eq. (68) yields

$$F_\alpha - (\nabla_{\mathbf{x}} F_\alpha, \lambda_\alpha \nabla_{\mathbf{x}} F_\alpha) = 0, \tag{71}$$

from which we obtain λ_α in the form

$$\lambda_\alpha = \frac{F_\alpha}{\|\nabla_{\mathbf{x}} F_\alpha\|^2}. \tag{72}$$

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

$$\begin{aligned} J_{\mathrm{ML}} &= \sum_{\alpha=1}^{N} \|\lambda_\alpha \nabla_{\mathbf{x}} F_\alpha\|^2 = \sum_{\alpha=1}^{N} \frac{F_\alpha^2}{\|\nabla_{\mathbf{x}} F_\alpha\|^4} \|\nabla_{\mathbf{x}} F_\alpha\|^2 \\ &= \sum_{\alpha=1}^{N} \frac{F_\alpha^2}{\|\nabla_{\mathbf{x}} F_\alpha\|^2}. \end{aligned} \tag{73}$$

B: Covariance Matrix of ML

After substitution of eqs. (8) and (9) into eq. (5) and doing Taylor expansion, J_{ML} is written as

$$J_{\mathrm{ML}} = \sum_{\alpha=1}^{N} \frac{((\nabla_{\mathbf{x}} \bar{F}_\alpha, \Delta\boldsymbol{x}_\alpha) + (\nabla_{\mathbf{u}} \bar{F}_\alpha, \Delta\boldsymbol{u}))^2}{\|\nabla_{\mathbf{x}} \bar{F}_\alpha\|^2} + O(\varepsilon^3), \tag{74}$$

where $\|\nabla_{\mathbf{x}} F_\alpha\|^2$ in the denominator is replaced by $\|\nabla_{\mathbf{x}} \bar{F}_\alpha\|^2$, which 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. (74), the ML estimator $\hat{\boldsymbol{u}}_{\rm ML}$ is given by $\boldsymbol{u}+\Delta\boldsymbol{u}$. The solution $\Delta\boldsymbol{u}$ is obtained by solving $\nabla_{\Delta\mathbf{u}}J_{\rm ML}=\mathbf{0}$. Since the first term on the right-hand side of eq. (74) is a quadratic form in $\Delta\boldsymbol{u}_\alpha$, we have

$$\nabla_{\Delta\mathbf{u}}J_{\rm ML}=2\sum_{\alpha=1}^{N}\frac{((\nabla_{\mathbf{x}}\bar{F}_\alpha,\Delta\boldsymbol{x}_\alpha)+(\nabla_{\mathbf{u}}\bar{F}_\alpha,\Delta\boldsymbol{u}))\nabla_{\mathbf{u}}\bar{F}_\alpha}{\|\nabla_{\mathbf{x}}\bar{F}_\alpha\|^2}+O(\varepsilon^2). \tag{75}$$

Letting this be 0, we have

$$\sum_{\alpha=1}^{N}\frac{(\nabla_{\mathbf{u}}\bar{F}_\alpha)(\nabla_{\mathbf{u}}\bar{F}_\alpha)^\top}{\|\nabla_{\mathbf{x}}\bar{F}_\alpha\|^2}\Delta\boldsymbol{u} =-\sum_{\alpha=1}^{N}\frac{(\nabla_{\mathbf{u}}\bar{F}_\alpha)(\nabla_{\mathbf{x}}\bar{F}_\alpha)^\top}{\|\nabla_{\mathbf{x}}\bar{F}_\alpha\|^2}\Delta\boldsymbol{x}_\alpha+O(\varepsilon^2), \tag{76}$$

from which we obtain

$$\sum_{\alpha=1}^{N}\frac{(\nabla_{\mathbf{u}}\bar{F}_\alpha)(\nabla_{\mathbf{u}}\bar{F}_\alpha)^\top}{\|\nabla_{\mathbf{x}}\bar{F}_\alpha\|^2}\Delta\boldsymbol{u}\Delta\boldsymbol{u}^\top\sum_{\beta=1}^{N}\frac{(\nabla_{\mathbf{u}}\bar{F}_\beta)(\nabla_{\mathbf{u}}\bar{F}_\beta)^\top}{\|\nabla_{\mathbf{x}}\bar{F}_\beta\|^2} =\sum_{\alpha,\beta=1}^{N}\frac{(\nabla_{\mathbf{u}}\bar{F}_\alpha)(\nabla_{\mathbf{x}}\bar{F}_\alpha)^\top}{\|\nabla_{\mathbf{x}}\bar{F}_\alpha\|^2}\Delta\boldsymbol{x}_\alpha\Delta\boldsymbol{x}_\beta^\top\frac{(\nabla_{\mathbf{x}}\bar{F}_\beta)(\nabla_{\mathbf{u}}\bar{F}_\beta)^\top}{\|\nabla_{\mathbf{x}}\bar{F}_\alpha\|^2} +O(\varepsilon^3). \tag{77}$$

Taking expectation on both sides, we obtain

$$\sum_{\alpha=1}^{N}\frac{(\nabla_{\mathbf{u}}\bar{F}_\alpha)(\nabla_{\mathbf{u}}\bar{F}_\alpha)^\top}{\|\nabla_{\mathbf{x}}\bar{F}_\alpha\|^2}V[\hat{\boldsymbol{u}}_{\rm ML}]\sum_{\beta=1}^{N}\frac{(\nabla_{\mathbf{u}}\bar{F}_\beta)(\nabla_{\mathbf{u}}\bar{F}_\beta)^\top}{\|\nabla_{\mathbf{x}}\bar{F}_\beta\|^2} =\sum_{\alpha=1}^{N}\frac{(\nabla_{\mathbf{u}}\bar{F}_\alpha)(\nabla_{\mathbf{x}}\bar{F}_\alpha)^\top}{\|\nabla_{\mathbf{x}}\bar{F}_\alpha\|^2}\frac{(\nabla_{\mathbf{x}}\bar{F}_\alpha)(\nabla_{\mathbf{u}}\bar{F}_\alpha)^\top}{\|\nabla_{\mathbf{x}}\bar{F}_\alpha\|^2}+O(\varepsilon^4) =\sum_{\alpha=1}^{N}\frac{(\nabla_{\mathbf{u}}\bar{F}_\alpha)(\nabla_{\mathbf{x}}\bar{F}_\alpha)^\top}{\|\nabla_{\mathbf{x}}\bar{F}_\alpha\|^2}+O(\varepsilon^4), \tag{78}$$

where we have used the relations

$$E[\Delta\boldsymbol{x}_\alpha\Delta\boldsymbol{x}_\beta^\top]=\delta_{\alpha\beta}\varepsilon^2\boldsymbol{I}, \tag{79}$$

and $E[O(\varepsilon^3)]=O(\varepsilon^4)$. From eq. (78) follows eq. (10).

C: Derivation of the KCR Lower Bound

We assume that estimator $\hat{\boldsymbol{u}}$ is unbiased, i.e.,

$$E[\hat{\boldsymbol{u}}-\boldsymbol{u}]=\mathbf{0}, \tag{80}$$

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[3]

$$\delta\int(\hat{\boldsymbol{u}}-\boldsymbol{u})p_1\cdots p_N d\boldsymbol{x}=-\int(\delta\boldsymbol{u})p_1\cdots p_N d\boldsymbol{x} +\sum_{\alpha=1}^{N}\int(\hat{\boldsymbol{u}}-\boldsymbol{u})p_1\cdots\delta p_\alpha\cdots p_N d\boldsymbol{x} =-\delta\boldsymbol{u}+\int(\hat{\boldsymbol{u}}-\boldsymbol{u})\sum_{\alpha=1}^{N}(p_1\cdots\delta p_\alpha\cdots p_N)d\boldsymbol{x}, \tag{81}$$

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

$$p(\boldsymbol{x}_\alpha)=\frac{1}{(\sqrt{2\pi})^n\varepsilon^n}e^{-\|\boldsymbol{x}_\alpha-\bar{\boldsymbol{x}}_\alpha\|^2/2\varepsilon^2}, \tag{82}$$

which we abbreviate to p_α. The infinitesimal variation of eq. (82) with respect to $\bar{\boldsymbol{x}}_\alpha$ is

$$\delta p_\alpha=(\boldsymbol{l}_\alpha,\delta\bar{\boldsymbol{x}}_\alpha)p_\alpha, \tag{83}$$

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

$$\boldsymbol{l}_\alpha\equiv\nabla_{\bar{\mathbf{x}}_\alpha}\log p_\alpha=\frac{\boldsymbol{x}_\alpha-\bar{\boldsymbol{x}}_\alpha}{\varepsilon^2}. \tag{84}$$

Since $\{\bar{\boldsymbol{x}}_\alpha\}$ and $\boldsymbol{u}$ are constrained by eq. (4), their variations are constrained to be

$$(\nabla_{\mathbf{x}}\bar{F}_\alpha,\delta\bar{\boldsymbol{x}}_\alpha)+(\nabla_{\mathbf{u}}\bar{F}_\alpha,\delta\boldsymbol{u})=0. \tag{85}$$

Because eq. (80) is an identity in $\{\bar{\boldsymbol{x}}_\alpha\}$ and $\boldsymbol{u}$ that satisfies eq. (4), the variation (81) should vanish for arbitrary variations $\{\delta\bar{\boldsymbol{x}}_\alpha\}$ and $\delta\boldsymbol{u}$ that satisfy eq. (85). Substituting eq. (83) into eq. (81), we conclude that

$$E[(\hat{\boldsymbol{u}}-\boldsymbol{u})\sum_{\alpha=1}^{N}\boldsymbol{l}_\alpha^\top\delta\bar{\boldsymbol{x}}_\alpha]=\delta\boldsymbol{u}, \tag{86}$$

for arbitrary variations $\{\delta\bar{\boldsymbol{x}}_\alpha\}$ and $\delta\boldsymbol{u}$ that satisfy eq. (85).

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

$$\delta\bar{\boldsymbol{x}}_\alpha=-\frac{(\nabla_{\mathbf{x}}\bar{F}_\alpha)(\nabla_{\mathbf{u}}\bar{F}_\alpha)^\top}{\|\nabla_{\mathbf{x}}\bar{F}_\alpha\|^2}\delta\boldsymbol{u}. \tag{87}$$

It is easy to confirm that eq. (85) is identically satisfied. Substituting eq. (87) into eq. (86), we obtain

$$E[(\hat{\boldsymbol{u}}-\boldsymbol{u})\sum_{\alpha=1}^{N}\boldsymbol{m}_\alpha^\top]\delta\boldsymbol{u}=-\delta\boldsymbol{u}, \tag{88}$$

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

$$\boldsymbol{m}_\alpha=\frac{(\nabla_{\mathbf{u}}\bar{F}_\alpha)(\nabla_{\mathbf{x}}\bar{F}_\alpha)^\top}{\|\nabla_{\mathbf{x}}\bar{F}_\alpha\|^2}\boldsymbol{l}_\alpha. \tag{89}$$

Because eq. (86) should hold for arbitrary variations $\{\delta\bar{\boldsymbol{x}}_\alpha\}$ and $\delta\boldsymbol{u}$ that satisfy eq. (85), eq. (88) should hold for arbitrary *unconstrained* variations $\delta\boldsymbol{u}$, which means

$$E[(\hat{\boldsymbol{u}}-\boldsymbol{u})\sum_{\alpha=1}^{N}\boldsymbol{m}_\alpha^\top]=-\boldsymbol{I}. \tag{90}$$

[3] 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 $\int p_1\cdots p_N d\boldsymbol{x}=1$.

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

$$E[\begin{pmatrix} \hat{\boldsymbol{u}}-\boldsymbol{u} \\ \sum_{\alpha=1}^N \boldsymbol{m}_\alpha \end{pmatrix}\begin{pmatrix} \hat{\boldsymbol{u}}-\boldsymbol{u} \\ \sum_{\alpha=1}^N \boldsymbol{m}_\alpha \end{pmatrix}^\top]=\begin{pmatrix} V[\hat{\boldsymbol{u}}] & -\boldsymbol{I} \\ -\boldsymbol{I} & \boldsymbol{M} \end{pmatrix}, \quad (91)$$

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

$$\begin{aligned} \boldsymbol{M} &= E[\Big(\sum_{\alpha=1}^N \boldsymbol{m}_\alpha\Big)\Big(\sum_{\beta=1}^N \boldsymbol{m}_\beta\Big)^\top] \\ &= \sum_{\alpha,\beta=1}^N \frac{(\nabla_{\mathbf{u}}\bar{F}_\alpha)(\nabla_{\mathbf{x}}\bar{F}_\alpha)^\top}{\|\nabla_{\mathbf{x}}\bar{F}_\alpha\|^2} E[\boldsymbol{l}_\alpha \boldsymbol{l}_\beta] \frac{(\nabla_{\mathbf{x}}\bar{F}_\alpha)(\nabla_{\mathbf{u}}\bar{F}_\alpha)^\top}{\|\nabla_{\mathbf{x}}\bar{F}_\alpha\|^2} \\ &= \frac{1}{\varepsilon^2}\frac{(\nabla_{\mathbf{u}}\bar{F}_\alpha)(\nabla_{\mathbf{u}}\bar{F}_\alpha)^\top}{\|\nabla_{\mathbf{x}}\bar{F}_\alpha\|^2}. \end{aligned} \quad (92)$$

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. (79) and (84). 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. (91) is evidently positive semidefinite , so is the right-hand side. Hence, the following is also positive semidefinite:

$$\begin{aligned} &\begin{pmatrix} \boldsymbol{I} & \boldsymbol{M}^{-1} \\ & \boldsymbol{M}^{-1} \end{pmatrix}\begin{pmatrix} V[\hat{\boldsymbol{u}}] & -\boldsymbol{I} \\ -\boldsymbol{I} & \boldsymbol{M} \end{pmatrix}\begin{pmatrix} \boldsymbol{I} & \\ \boldsymbol{M}^{-1} & \boldsymbol{M}^{-1} \end{pmatrix} \\ &= \begin{pmatrix} V[\hat{\boldsymbol{u}}]-\boldsymbol{M}^{-1} & \\ & \boldsymbol{M}^{-1} \end{pmatrix}. \end{aligned} \quad (93)$$

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

The above proof is for the simplest case, but the same result holds for more general cases. 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 pseudoinverse and projection operators. 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, but the logic is the same [10].

D: Higher Order Terms of Renormalization

1. The vector $\boldsymbol{u}$ should be perturbed subject to the normalization constraint, so

$$\begin{aligned} &\|\boldsymbol{u}+\Delta_1\boldsymbol{u}+\Delta_2\boldsymbol{u}+\cdots\|^2 \\ &=(\boldsymbol{u}+\Delta_1\boldsymbol{u}+\Delta_2\boldsymbol{u}+\cdots, \boldsymbol{u}+\Delta_1\boldsymbol{u}+\Delta_2\boldsymbol{u}+\cdots)=1. \end{aligned} \quad (94)$$

Comparing terms of $O(1)$, $O(\varepsilon)$, and $O(\varepsilon^2)$ on both sides, we obtain

$$(\boldsymbol{u},\boldsymbol{u}) = \|\boldsymbol{u}\|^2 = 1, \qquad (\boldsymbol{u},\Delta_1\boldsymbol{u}) = 0, \quad (95)$$

$$(\boldsymbol{u},\Delta_2\boldsymbol{u}) = -(\Delta_1\boldsymbol{u},\Delta_1\boldsymbol{u}) = -\|\Delta_1\boldsymbol{u}\|^2. \quad (96)$$

2. Comparing terms of $O(1)$ on both sides of eq. (50), we obtain $\bar{\boldsymbol{M}}\boldsymbol{u} = \boldsymbol{0}$.

3. Comparing terms of $O(\varepsilon)$ on both sides of eq. (50), we obtain

$$\bar{\boldsymbol{M}}\Delta_1\boldsymbol{u} + \Delta_1\boldsymbol{M}\boldsymbol{u} = \Delta_1\lambda\boldsymbol{u}. \quad (97)$$

Computing inner product with $\boldsymbol{u}$ on both sides, we obtain

$$(\boldsymbol{u},\bar{\boldsymbol{M}}\Delta_1\boldsymbol{u}) + (\boldsymbol{u},\Delta_1\boldsymbol{M}\boldsymbol{u}) = \Delta_1\lambda(\boldsymbol{u},\boldsymbol{u}). \quad (98)$$

Noting that $(\boldsymbol{u},\bar{\boldsymbol{M}}\Delta_1\boldsymbol{u}) = (\bar{\boldsymbol{M}}\boldsymbol{u},\Delta_1\boldsymbol{u}) = 0$, $(\boldsymbol{u},\boldsymbol{u}) = \|\boldsymbol{u}\| = 1$, and eq. (48), we obtain

$$\begin{aligned} \Delta_1\lambda &= (\boldsymbol{u},\Delta_1\boldsymbol{M}\boldsymbol{u}) \\ &= \sum_{\alpha=1}^N \frac{(\boldsymbol{u},\Delta\boldsymbol{\xi}_\alpha)(\bar{\boldsymbol{\xi}}_\alpha,\boldsymbol{u}) + (\boldsymbol{u},\bar{\boldsymbol{\xi}}_\alpha)(\Delta\boldsymbol{\xi}_\alpha,\boldsymbol{u})}{(\boldsymbol{u},V_0[\boldsymbol{\xi}_\alpha]\boldsymbol{u})} = 0. \end{aligned} \quad (99)$$

Let λ_1, ..., λ_{n-1} be the nonzero eigenvalues of $\bar{\boldsymbol{M}}$, and $\boldsymbol{u}_1$, ..., $\boldsymbol{u}_{n-1}$ the corresponding orthonormal system of eigenvectors (recall that the unit eigenvector for eigenvalue 0 is $\boldsymbol{u}$ itself). Computing inner product with $\boldsymbol{u}_i$ on both sides of eq. (97), we obtain

$$(\boldsymbol{u}_i,\bar{\boldsymbol{M}}\Delta_1\boldsymbol{u}) + (\boldsymbol{u}_i,\Delta_1\boldsymbol{M}\boldsymbol{u}) = \Delta_1\lambda(\boldsymbol{u}_i,\boldsymbol{u}). \quad (100)$$

Since $(\boldsymbol{u}_i,\bar{\boldsymbol{M}}\Delta_1\boldsymbol{u}) = (\bar{\boldsymbol{M}}\boldsymbol{u}_i,\Delta_1\boldsymbol{u}) = (\lambda_i\boldsymbol{u}_i,\Delta_1\boldsymbol{u})$ and $(\boldsymbol{u}_i,\boldsymbol{u}) = 0$, the above equation is rewritten as

$$\lambda_i(\boldsymbol{u}_i,\Delta_1\boldsymbol{u}) + (\boldsymbol{u}_i,\Delta_1\boldsymbol{M}\boldsymbol{u}) = 0. \quad (101)$$

Since $\Delta_1\boldsymbol{u}$ is orthogonal to $\boldsymbol{u}$ by the second of eqs. (95), it can be expressed as a linear combination of the orthonormal system $\boldsymbol{u}_1$, ..., $\boldsymbol{u}_{n-1}$ in the form

$$\begin{aligned} \Delta_1\boldsymbol{u} &= \sum_{i=1}^{n-1}(\boldsymbol{u}_i,\Delta_1\boldsymbol{u})\boldsymbol{u}_i = -\sum_{i=1}^{n-1}\frac{\boldsymbol{u}_i(\boldsymbol{u}_i,\Delta_1\boldsymbol{M}\boldsymbol{u})}{\lambda_i} \\ &= -\Big(\sum_{i=1}^{n-1}\frac{\boldsymbol{u}_i\boldsymbol{u}_i^\top}{\lambda_i}\Big)\Delta_1\boldsymbol{M}\boldsymbol{u} = -\bar{\boldsymbol{M}}^-\Delta_1\boldsymbol{M}\boldsymbol{u}. \end{aligned} \quad (102)$$

4. Comparing terms of $O(\varepsilon^2)$ on both sides of eq. (50), we obtain

$$\bar{\boldsymbol{M}}\Delta_2\boldsymbol{u}+\Delta_2\boldsymbol{M}\boldsymbol{u}+\Delta_1\boldsymbol{M}\Delta_1\boldsymbol{u} = \Delta_1\lambda\Delta_1\boldsymbol{u}+\Delta_2\lambda\boldsymbol{u}. \quad (103)$$

Computing inner product with $\boldsymbol{u}$ on both sides, we have

$$\begin{aligned} &(\boldsymbol{u},\bar{\boldsymbol{M}}\Delta_2\boldsymbol{u}) + (\boldsymbol{u},\Delta_2\boldsymbol{M}\boldsymbol{u}) + (\boldsymbol{u},\Delta_1\boldsymbol{M}\Delta_1\boldsymbol{u}) \\ &\qquad = \Delta_1\lambda(\boldsymbol{u},\Delta_1\boldsymbol{u}) + \Delta_2\lambda(\boldsymbol{u},\boldsymbol{u}). \end{aligned} \quad (104)$$

Noting that $(\boldsymbol{u}, \bar{\boldsymbol{M}}\Delta_2\boldsymbol{u}) = (\bar{\boldsymbol{M}}\boldsymbol{u}, \Delta_2\boldsymbol{u}) = 0$, $(\boldsymbol{u}, \Delta_1\boldsymbol{u}) = 0$, $(\boldsymbol{u}, \boldsymbol{u}) = \|\boldsymbol{u}\| = 1$, and eq. (102), we obtain

$$\begin{aligned}\Delta_2\lambda &= (\boldsymbol{u}, \Delta_2\boldsymbol{M}\boldsymbol{u}) + (\boldsymbol{u}, \Delta_1\boldsymbol{M}\Delta_1\boldsymbol{u}) \\ &= (\boldsymbol{u}, \Delta_2\boldsymbol{M}\boldsymbol{u}) - (\boldsymbol{u}, \Delta_1\boldsymbol{M}\bar{\boldsymbol{M}}^-\Delta_1\boldsymbol{M}\boldsymbol{u}). \end{aligned} \tag{105}$$

Computing inner product with $\boldsymbol{u}_i$ on both sides of eq. (103), we obtain

$$\begin{aligned}&(\boldsymbol{u}_i, \bar{\boldsymbol{M}}\Delta_2\boldsymbol{u}) + (\boldsymbol{u}_i, \Delta_2\boldsymbol{M}\boldsymbol{u}) + (\boldsymbol{u}_i, \Delta_1\boldsymbol{M}\Delta_1\boldsymbol{u}) \\ &\quad = \Delta_1\lambda(\boldsymbol{u}_i, \Delta_1\boldsymbol{u}) + \Delta_2\lambda(\boldsymbol{u}_i, \boldsymbol{u}). \end{aligned} \tag{106}$$

Noting that $(\boldsymbol{u}_i, \bar{\boldsymbol{M}}\Delta_2\boldsymbol{u}) = (\bar{\boldsymbol{M}}\boldsymbol{u}_i, \Delta_2\boldsymbol{u}) = (\lambda_i\boldsymbol{u}_i, \Delta_2\boldsymbol{u})$, $(\boldsymbol{u}_i, \boldsymbol{u}) = 0$, and eqs. (99), (102), we obtain

$$\lambda_i(\boldsymbol{u}_i, \Delta_2\boldsymbol{u}) + (\boldsymbol{u}_i, \Delta_2\boldsymbol{M}\boldsymbol{u}) - (\boldsymbol{u}_i, \Delta_1\boldsymbol{M}\bar{\boldsymbol{M}}^-\Delta_1\boldsymbol{M}\boldsymbol{u}) = 0. \tag{107}$$

From this and eq. (96), we obtain

$$\begin{aligned}\Delta_2\boldsymbol{u} &= \sum_{i=1}^{n-1}(\boldsymbol{u}_i, \Delta_2\boldsymbol{u})\boldsymbol{u}_i + (\boldsymbol{u}, \Delta_2\boldsymbol{u})\boldsymbol{u} \\ &= -\sum_{i=1}^{n-1}\frac{\boldsymbol{u}_i(\boldsymbol{u}_i, \Delta_2\boldsymbol{M}\boldsymbol{u})}{\lambda_i} \\ &\quad + \sum_{i=1}^{n-1}\frac{\boldsymbol{u}_i(\boldsymbol{u}_i, \Delta_1\boldsymbol{M}\bar{\boldsymbol{M}}^-\Delta_1\boldsymbol{M}\boldsymbol{u})}{\lambda_i} - \|\Delta_1\boldsymbol{u}\|^2\boldsymbol{u} \\ &= -\Big(\sum_{i=1}^{n-1}\frac{\boldsymbol{u}_i\boldsymbol{u}_i^\top}{\lambda_i}\Big)\Delta_2\boldsymbol{M}\boldsymbol{u} - \|\bar{\boldsymbol{M}}^-\Delta_1\boldsymbol{M}\boldsymbol{u}\|^2\boldsymbol{u} \\ &\quad + \Big(\sum_{i=1}^{n-1}\frac{\boldsymbol{u}_i\boldsymbol{u}_i^\top}{\lambda_i}\Big)\Delta_1\boldsymbol{M}\bar{\boldsymbol{M}}^-\Delta_1\boldsymbol{M}\boldsymbol{u} \\ &= -\bar{\boldsymbol{M}}^-\Delta_2\boldsymbol{M}\boldsymbol{u} + \bar{\boldsymbol{M}}^-\Delta_1\boldsymbol{M}\bar{\boldsymbol{M}}^-\Delta_1\boldsymbol{M}\boldsymbol{u} \\ &\quad -\|\bar{\boldsymbol{M}}^-\Delta_1\boldsymbol{M}\boldsymbol{u}\|^2\boldsymbol{u}. \end{aligned} \tag{108}$$

E: Evaluation of $E[\|\bar{\boldsymbol{M}}^-\Delta_1\boldsymbol{M}\boldsymbol{u}\|^2]$

Since $(\boldsymbol{\xi}_\alpha, \boldsymbol{u}) = 0$, eq. (48) implies

$$\Delta_1\boldsymbol{M}\boldsymbol{u} = \sum_{\alpha=1}^{N}\frac{(\Delta\boldsymbol{\xi}_\alpha, \boldsymbol{u})\bar{\boldsymbol{\xi}}_\alpha}{(\boldsymbol{u}, V_0[\boldsymbol{\xi}_\alpha]\boldsymbol{u})}. \tag{109}$$

Hence,

$$\begin{aligned}&E[\|\bar{\boldsymbol{M}}^-\Delta_1\boldsymbol{M}\boldsymbol{u}\|^2] = E[(\bar{\boldsymbol{M}}^-\Delta_1\boldsymbol{M}\boldsymbol{u}, \bar{\boldsymbol{M}}^-\Delta_1\boldsymbol{M}\boldsymbol{u})] \\ &= E[(\Delta_1\boldsymbol{M}\boldsymbol{u}, (\bar{\boldsymbol{M}}^-)^2\Delta_1\boldsymbol{M}\boldsymbol{u})] \\ &= E[(\sum_{\alpha=1}^{N}\frac{\bar{\boldsymbol{\xi}}_\alpha(\Delta\boldsymbol{\xi}_\alpha, \boldsymbol{u})}{(\boldsymbol{u}, V_0[\boldsymbol{\xi}_\alpha]\boldsymbol{u})}, (\bar{\boldsymbol{M}}^-)^2\sum_{\beta=1}^{N}\frac{\bar{\boldsymbol{\xi}}_\beta(\Delta\boldsymbol{\xi}_\beta, \boldsymbol{u})}{(\boldsymbol{u}, V_0[\boldsymbol{\xi}_\beta]\boldsymbol{u})})] \\ &= \sum_{\alpha,\beta=1}^{N}\frac{E[(\Delta\boldsymbol{\xi}_\alpha, \boldsymbol{u})(\Delta\boldsymbol{\xi}_\beta, \boldsymbol{u})](\bar{\boldsymbol{\xi}}_\alpha, (\bar{\boldsymbol{M}}^-)^2\bar{\boldsymbol{\xi}}_\beta)}{(\boldsymbol{u}, V_0[\boldsymbol{\xi}_\alpha]\boldsymbol{u})(\boldsymbol{u}, V_0[\boldsymbol{\xi}_\beta]\boldsymbol{u})} \\ &= \sum_{\alpha,\beta=1}^{N}\frac{(\boldsymbol{u}, E[\Delta\boldsymbol{\xi}_\alpha\Delta\boldsymbol{\xi}_\beta^\top]\boldsymbol{u})(\bar{\boldsymbol{\xi}}_\alpha, (\bar{\boldsymbol{M}}^-)^2\bar{\boldsymbol{\xi}}_\beta)}{(\boldsymbol{u}, V_0[\boldsymbol{\xi}_\alpha]\boldsymbol{u})(\boldsymbol{u}, V_0[\boldsymbol{\xi}_\beta]\boldsymbol{u})} \\ &= \sum_{\alpha,\beta=1}^{N}\frac{(\boldsymbol{u}, \varepsilon^2\delta_{\alpha\beta}V_0[\boldsymbol{\xi}_\alpha]\boldsymbol{u})(\bar{\boldsymbol{\xi}}_\alpha, (\bar{\boldsymbol{M}}^-)^2\bar{\boldsymbol{\xi}}_\beta)}{(\boldsymbol{u}, V_0[\boldsymbol{\xi}_\alpha]\boldsymbol{u})(\boldsymbol{u}, V_0[\boldsymbol{\xi}_\beta]\boldsymbol{u})} \\ &= \varepsilon^2\sum_{\alpha=1}^{N}\frac{(\boldsymbol{u}, V_0[\boldsymbol{\xi}_\alpha]\boldsymbol{u})(\bar{\boldsymbol{\xi}}_\alpha, (\bar{\boldsymbol{M}}^-)^2\bar{\boldsymbol{\xi}}_\alpha)}{(\boldsymbol{u}, V_0[\boldsymbol{\xi}_\alpha]\boldsymbol{u})^2} \\ &= \varepsilon^2\sum_{\alpha=1}^{N}\frac{(\bar{\boldsymbol{\xi}}_\alpha, (\bar{\boldsymbol{M}}^-)^2\bar{\boldsymbol{\xi}}_\alpha)}{(\boldsymbol{u}, V_0[\boldsymbol{\xi}_\alpha]\boldsymbol{u})} \\ &= \varepsilon^2\mathrm{tr}(\sum_{\alpha=1}^{N}\frac{\bar{\boldsymbol{\xi}}_\alpha\bar{\boldsymbol{\xi}}_\alpha^\top}{(\boldsymbol{u}, V_0[\boldsymbol{\xi}_\alpha]\boldsymbol{u})}(\bar{\boldsymbol{M}}^-)^2) = \varepsilon^2\mathrm{tr}(\bar{\boldsymbol{M}}(\bar{\boldsymbol{M}}^-)^2) \\ &= \varepsilon^2\mathrm{tr}(\bar{\boldsymbol{M}}^-\bar{\boldsymbol{M}}\bar{\boldsymbol{M}}^-) = \varepsilon^2\mathrm{tr}(\bar{\boldsymbol{M}}^-). \end{aligned} \tag{110}$$

F: Higher Order Terms of ML

1. From the constraint that $\boldsymbol{u}$ be a unit vector, eq. (94) holds for the perturbations $\Delta_1\boldsymbol{u}$ and $\Delta_2\boldsymbol{u}$.

2. Comparing terms of $O(1)$ on both sides of eq. (60), we obtain $\bar{\boldsymbol{M}}\boldsymbol{u} = \boldsymbol{0}$.

3. Comparing terms of $O(\varepsilon)$ on both sides of eq. (60), we obtain

$$\bar{\boldsymbol{M}}\Delta_1\boldsymbol{u} + \Delta_1\boldsymbol{M}\boldsymbol{u} = \boldsymbol{0}. \tag{111}$$

Multiplying both sides by $\bar{\boldsymbol{M}}^-$ from left, we obtain

$$\boldsymbol{P}_\mathbf{u}\Delta_1\boldsymbol{u} + \bar{\boldsymbol{M}}^-\Delta_1\boldsymbol{M}\boldsymbol{u} = \boldsymbol{0}, \tag{112}$$

where $\boldsymbol{P}_\mathbf{u} = \boldsymbol{I} - \boldsymbol{u}\boldsymbol{u}^\top$ is the projection operator along $\boldsymbol{u}$ (note that $\bar{\boldsymbol{M}}^-\bar{\boldsymbol{M}} = \boldsymbol{P}_\mathbf{u}$ [10]). Since $\Delta_1\boldsymbol{u}$ is orthogonal to $\boldsymbol{u}$ by the second of eqs. (95), we have $\boldsymbol{P}_\mathbf{u}\Delta_1\boldsymbol{u} = \Delta_1\boldsymbol{u}$. Hence, we obtain eq. (61).

4. Comparing terms of $O(\varepsilon^2)$ on both sides of eq. (60), we obtain

$$\bar{\boldsymbol{M}}\Delta_2\boldsymbol{u} + \Delta_1\boldsymbol{M}\Delta_1\boldsymbol{u} + \Delta_2\boldsymbol{M}\boldsymbol{u} = \Delta_2\boldsymbol{L}\boldsymbol{u}. \tag{113}$$

Multiplying both sides by $\bar{\boldsymbol{M}}^-$ from left, we obtain after some rearrangements

$$\begin{aligned}\boldsymbol{P}_\mathbf{u}\Delta_2\boldsymbol{u} = &-\bar{\boldsymbol{M}}^-\Delta_2\boldsymbol{M}\boldsymbol{u} + \bar{\boldsymbol{M}}^-\Delta_1\boldsymbol{M}\bar{\boldsymbol{M}}^-\Delta_1\boldsymbol{M}\boldsymbol{u} \\ &+\bar{\boldsymbol{M}}^-\Delta_2\boldsymbol{L}\boldsymbol{u}. \end{aligned} \tag{114}$$

This is the component of $\Delta_2\boldsymbol{u}$ orthogonal to $\boldsymbol{u}$. The component along $\boldsymbol{u}$ is $-\|\Delta_1\boldsymbol{u}\|^2\boldsymbol{u}$ from eq. (96). Hence,

$$\Delta_2\boldsymbol{u} = \boldsymbol{P}_\mathbf{u}\Delta_2\boldsymbol{u} - \|\Delta_1\boldsymbol{u}\|^2\boldsymbol{u}. \tag{115}$$

The first order perturbation $\Delta_1\boldsymbol{u}$ is the same as in the case of renormalization (see eqs. (51) and (61)). Hence, $\|\Delta_1\boldsymbol{u}\|^2 = -\|\bar{\boldsymbol{M}}^-\Delta_1\boldsymbol{M}\boldsymbol{u}\|^2\boldsymbol{u}$, as we have already shown. Thus, we obtain eq. (62).