Optimal estimation

Kenichi Kanatani, Okayama University

Synonyms

– Optimal parameter estimation

Related Concepts

– Algebraic fitting
– Bundle adjustment
– Geometric fitting
– Least squares
– Mahalanobis distance
– Maximum likelihood
– Reprojection error
– Sampson error

Definition

Optimal estimation in the computer vision context usually refers to estimating the parameters that describe the underlying problem from noisy observation. The estimation is done according to a given criterion of optimality, for which maximum likelihood is widely accepted. If Gaussian noise is assumed, it reduces to minimizing the Mahalanobis distance. If furthermore the Gaussian noise has a homogeneous and isotropic distribution, the procedure reduces to minimizing what is called the reprojection error.

Background

One of the central tasks of computer vision is the extraction of 2-D/3-D geometric information from noisy image data. Here, the term "image data" refers to values extracted from images by image processing operations such as edge filters and interest point detectors. Image data are said to be "noisy" in the sense that image processing operations for detecting them entail uncertainty to some extent.

For optimal estimation, a statistical model of observation needs to be introduced. Let $\boldsymbol{x}_1$, ..., $\boldsymbol{x}_N$ be the observed image data. The standard model is to view each datum $\boldsymbol{x}_\alpha$ as perturbed from its true value $\bar{\boldsymbol{x}}_\alpha$ by $\Delta\boldsymbol{x}_\alpha$, which is assumed to be independent Gaussian noise of mean $\boldsymbol{0}$ and covariance matrix $V[\boldsymbol{x}_\alpha]$. Then, maximum likelihood is equivalent to the minimization of the *Mahalanobis distance*

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

with respect to the true values $\bar{\boldsymbol{x}}_\alpha$ subject to given knowledge about them. Hereafter, $(\boldsymbol{a}, \boldsymbol{b})$ denotes the inner product of vectors $\boldsymbol{a}$ and $\boldsymbol{b}$.

If the noise is homogeneous and isotropic, in which case $V[\boldsymbol{x}_\alpha] = c\boldsymbol{I}$ for all α for some constant c and the unit matrix $\boldsymbol{I}$, the Mahalanobis distance I is equivalent to the sum of the squares of the geometric distances between the observations $\boldsymbol{x}_\alpha$ and their true values $\bar{\boldsymbol{x}}_\alpha$, often referred to as the *reprojection error*. That name originates from the following intuition: In inferring the 3-D structure of the scene from its projected images, maximum likelihood under homogeneous and isotropic Gaussian noise means "reprojecting" the inferred 3-D structure onto the images and minimizing the square distance between the "reprojection" of the solution and the projection of the scene. Reprojection error minimization is also referred to as *geometric fitting*.

Theory

The estimation procedure depends on the way the knowledge about true values $\bar{\boldsymbol{x}}_\alpha$ is represented. A typical approach is to introduce some function $\boldsymbol{g}(\boldsymbol{t}, \boldsymbol{\theta})$ to express $\bar{\boldsymbol{x}}_\alpha$ in a parametric form

$$\bar{\boldsymbol{x}}_\alpha = \boldsymbol{g}(\boldsymbol{t}_\alpha, \boldsymbol{\theta}), \tag{2}$$

where $\boldsymbol{t}_\alpha$ is a control variable that specifies the identity of the αth datum, and $\boldsymbol{\theta}$ is an unknown parameter that specifies the underlying structure. After (2) is substituted, the Mahalanobis distance I becomes a function of $\boldsymbol{\theta}$ alone, which is then minimized with respect to $\boldsymbol{\theta}$. This is the standard approach in the traditional statistic estimation framework and also known as *regression*.

This parametric approach, however, is quite limited in computer vision applications. Often, no such knowledge as (2) is available about the true values $\bar{\boldsymbol{x}}_\alpha$ except that that they satisfy some implicit equations of the form

$$F^{(k)}(\boldsymbol{x}, \boldsymbol{\theta}) = 0, \qquad k = 1, ..., L. \tag{3}$$

The unknown parameter $\boldsymbol{\theta}$ allows one to infer the 2-D/3-D shape and motion of the objects observed in the images.

This type of estimation leads to some theoretical problems. Usually, no restriction is imposed on the true values $\bar{\boldsymbol{x}}_\alpha$ except that they should satisfy (3). This is called the *functional model*. One could alternatively introduce some statistical model according to which the true values $\bar{\boldsymbol{x}}_\alpha$ are sampled. Then, the model is called *structural*. This distinction is crucial when one considers limiting processes in the following sense. Traditional statistical analysis mainly focuses on the asymptotic behavior as the number of observations increases to ∞. This is based on the reasoning that the mechanism underlying noisy observations would better reveal itself as the number of observations increases (the law of large numbers) while the number of available data is limited in practice. So, the estimation accuracy vs. the number of data is a major concern. In this light, efforts have been made to obtain a consistent estimator in the sense that the solution approaches its true value in the limit $N \to \infty$ of the number N of the data.

In computer vision applications, in contrast, one cannot "repeat" observations. One makes an inference given a single set of images, and how many times one applies image processing operations, the result is always the same, because standard image processing algorithms are deterministic and no randomness is involved. This is in a stark contrast to conventional statistical problems where observations are viewed as "samples" from potentially infinitely many possibilities and could obtain, by repeating observations, different values originating from unknown, uncontrollable, or unmodeled causes, which is called "noise" as a whole.

In vision problems, the accuracy of inference deteriorates as the uncertainty of image processing operations increases. Thus, the inference accuracy vs. the uncertainty of image operations, which is called "noise" for simplicity, is a major concern. Usually, the noise is very small, often subpixel levels. In light of this observation, it has been pointed out that in image domains the "consistency" of estimators should more appropriately be defined by the behavior in the limit $\sigma \to 0$ of the noise level σ [1, 3]. The functional model suits this purpose. If the error behavior in the limit of $N \to \infty$ were to be analyzed, one needs to assume some structural model that specifies how the statistical characteristics of the data depend on N. However, it is difficult to predict the noise characteristics for different N. Image processing filters usually output a list of points or lines or their correspondences along with their confidence values, from which only those with high confidence are used. If a lot of data are to be collected, those with low confidence need to be included, but their statistical properties are hard to estimate, since such data are possibly misdetections. This is the most different aspect of image processing from laboratory experiments, in which any number of data can be collected by repeated trials.

Maximum likelihood with implicit constraints

Maximum likelihood based on the functional model is to minimize (1) subject to implicit constraints in the form of (3). In statistics, maximum likelihood is criticized for its lack of consistency. In fact, estimation of the true values $\bar{\boldsymbol{x}}_\alpha$, called *nuisance parameters* when viewed as parameters, is not consistent as $N \to \infty$ in the maximum likelihood framework [6]. However, the lack of consistency has no realistic meaning in vision applications as explained above. On the contrary, maximum likelihood has very desirable properties in the limit $\sigma \to 0$ of the noise level σ: the solution is "consistent" in the sense that it converges to the true value as $\sigma \to 0$ and "efficient" in the sense that its covariance matrix approaches a theoretical lower bound as $\sigma \to 0$ [1, 3].

According to the experience of many vision researchers, maximum likelihood is known to produce highly accurate solutions, and no necessity is felt for further accuracy improvement. Rather, a major concern is its computational burden, because maximum likelihood usually requires complicated nonlinear optimization. The standard approach is to introduce some auxiliary parameters to express each of $\bar{\boldsymbol{x}}_\alpha$ explicitly in terms of $\boldsymbol{\theta}$ and the auxiliary parameters. After they are substituted back into (1), the Mahalanobis distance I becomes a function of $\boldsymbol{\theta}$ and the auxiliary parameters. Then, this joint parameter space, which usu-

ally has very high dimensions, is searched for the minimum. This approach is called *bundle adjustment*, a term originally used by photogrammetrists. This is very time consuming, in particular if one seeks a globally optimal solution by searching the entire parameter space exhaustively.

Linear reparameterization

In many important vision applications, the problem can be reparameterized to make the functions $F^{(k)}(\boldsymbol{x}, \boldsymbol{\theta})$ linear in $\boldsymbol{\theta}$ (but generally nonlinear in $\boldsymbol{x}$), allowing one to write (3) as

$$(\boldsymbol{\xi}^{(k)}(\boldsymbol{x}), \boldsymbol{\theta}) = 0, \qquad k = 1, ..., L, \tag{4}$$

where $\boldsymbol{\xi}^{(k)}(\boldsymbol{x})$ represents a nonlinear mapping of $\boldsymbol{x}$. This formalism covers many fundamental problems of computer vision including fitting a parametric curve such as a line, an ellipse, and a polynomial curve to a noisy 2-D point sequence or a parametric surface such as a plane, an ellipsoid, and a polynomial surface to a noisy 3-D point sets and computing the fundamental matrix or the homography from noisy point correspondences over two images. For this type of problem, a popular alternative to bundle adjustment is minimization of a function of $\boldsymbol{\theta}$ alone, called the *Sampson error*. Let us abbreviate $\boldsymbol{\xi}^{(k)}(\boldsymbol{x}_\alpha)$ to $\boldsymbol{\xi}_\alpha^{(k)}$. The first order variation of $\boldsymbol{\xi}_\alpha^{(k)}$ by noise is

$$\Delta\boldsymbol{\xi}_\alpha^{(k)} = \boldsymbol{T}_\alpha^{(k)} \Delta\boldsymbol{x}_\alpha, \qquad \boldsymbol{T}_\alpha^{(k)} \equiv \left.\frac{\partial \boldsymbol{\xi}^{(k)}(\boldsymbol{x})}{\partial \boldsymbol{x}}\right|_{\boldsymbol{x}=\bar{\boldsymbol{x}}_\alpha}. \tag{5}$$

Define the covariance matrices of $\boldsymbol{\xi}_\alpha^{(k)}$, $k = 1, ..., L$, by

$$V^{(kl)}[\boldsymbol{\xi}_\alpha] = E[\Delta\boldsymbol{\xi}_\alpha^{(k)} \Delta\boldsymbol{\xi}_\alpha^{(l)\top}] = \boldsymbol{T}_\alpha^{(k)} E[\Delta\boldsymbol{x}_\alpha \Delta\boldsymbol{x}_\alpha^\top] \boldsymbol{T}_\alpha^{(l)\top} = \boldsymbol{T}_\alpha^{(k)} V[\boldsymbol{x}_\alpha] \boldsymbol{T}_\alpha^{(l)\top}, \tag{6}$$

where $E[\,\cdot\,]$ denotes expectation. The Sampson error that approximates the minimum of the Mahalanobis distance I subject to the constraints in (4) has the form

$$K = \sum_{\alpha=1}^{N} \sum_{k,l,=1}^{L} W_\alpha^{(kl)} (\boldsymbol{\xi}_\alpha^{(k)}, \boldsymbol{\theta}) (\boldsymbol{\xi}_\alpha^{(l)}, \boldsymbol{\theta}), \tag{7}$$

where $W_\alpha^{(kl)}$ is the (kl) element of $(\boldsymbol{V}_\alpha)_r^-$. Here, $\boldsymbol{V}_\alpha$ is the matrix whose (kl) element is

$$\boldsymbol{V}_\alpha = \Big((\boldsymbol{\theta}, V^{(kl)}[\boldsymbol{\xi}_\alpha] \boldsymbol{\theta}) \Big), \tag{8}$$

where the true data values $\bar{\boldsymbol{x}}_\alpha$ in the definition of $V^{(kl)}[\boldsymbol{\xi}_\alpha]$ are replaced by their observations $\boldsymbol{x}_\alpha$. The operation $(\,\cdot\,)_r^-$ denotes the pseudoinverse of truncated rank r, (i.e., with all eigenvalues except the largest r replaced by 0 in the spectral decomposition), and r is the rank (the number of independent equations) of (4). The name Sampson error stems from the classical ellipse fitting scheme [8].

The Sampson error (7) can be minimized by various means including the *FNS* (*Fundamental Numerical Scheme*) [2], the *HEIV* (*Heteroscedastic Errors-in-Variable*) [5]. It can be shown that the exact maximum likelihood solution can

be obtained by repeating Sampson error minimization, each time modifying the Sampson error so that in the end the modified Sampson error coincides with the Mahalanobis distance [4]. It turns out that in many practical applications the solution that minimizes the Sampson error coincides with the exact maximum likelihood solution up to several significant digits; usually, two or three rounds of Sampson error modification are sufficient.

It can be shown that the covariance matrix $V[\hat{\boldsymbol{\theta}}]$ of any unbiased estimator $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$ satisfies under some general conditions the inequality

$$V[\hat{\boldsymbol{\theta}}] \succ \Big(\sum_{\alpha=1}^{N} \sum_{k,l=1}^{L} \bar{W}_{\alpha}^{(kl)} \bar{\boldsymbol{\xi}}_{\alpha}^{(k)} \bar{\boldsymbol{\xi}}_{\alpha}^{(l)\top} \Big)_{r}^{-}, \tag{9}$$

where $\bar{\boldsymbol{\xi}}_{\alpha}^{(k)}$ are the trues value of $\boldsymbol{\xi}_{\alpha}^{(k)}$, and $\bar{W}_{\alpha}^{(kl)}$ is the value of $W_{\alpha}^{(kl)}$ defined earlier evaluated for the true values of $\boldsymbol{\xi}_{\alpha}^{(k)}$ and $\boldsymbol{\theta}$. The symbol $\succ$ means that the left-hand side minus the right-hand side is positive semidefinite. The right-hand side of (9) is called the *KCR* (*Kanatani-Cramer-Rao*) *lower bound* [1, 3]. It can be shown that the covariance matrix of Sampson error minimization solution coincides with this bound in the leading order in the noise level [1, 3].

Algebraic methods

Sampson error minimization schemes such as FNS and HEIV rely on local search, and the iterations do not always converge unless started from a value sufficiently close to the solution. For accurate initialization of the iterations, various types of algebraic method have been studied. *Algebraic methods* refer to minimizing the *algebraic distance*

$$J = \sum_{\alpha=1}^{N} \sum_{k=1}^{L} (\boldsymbol{\xi}_{\alpha}^{(k)}, \boldsymbol{\theta})^2 = \sum_{\alpha=1}^{N} \sum_{k=1}^{L} \boldsymbol{\theta}^{\top} \boldsymbol{\xi}_{\alpha}^{(k)} \boldsymbol{\xi}_{\alpha}^{(k)\top} \boldsymbol{\theta} = (\boldsymbol{\theta}, \boldsymbol{M}\boldsymbol{\theta}), \tag{10}$$

where the matrix $\boldsymbol{M}$ is defined by

$$\boldsymbol{M} = \sum_{\alpha=1}^{N} \sum_{k=1}^{L} \boldsymbol{\xi}_{\alpha}^{(k)} \boldsymbol{\xi}_{\alpha}^{(k)\top}. \tag{11}$$

Note that if $W_{\alpha}^{(kl)}$ in (7) is replaced by the Kronecker delta δ_{kl}, the Sampson error K coincides with the algebraic distance J. Algebraic distance minimization is also referred to as *algebraic fitting* or simply *least squares*.

The algebraic distance J is trivially minimized by $\boldsymbol{\theta} = \boldsymbol{0}$ unless some scale normalization is imposed on $\boldsymbol{\theta}$. The standard normalization is $\|\boldsymbol{\theta}\| = 1$. A more general class of normalization is the form of $(\boldsymbol{\theta}, \boldsymbol{N}\boldsymbol{\theta}) = 1$, where $\boldsymbol{N}$ is some positive definite or semidefinite symmetric matrix. It is easily seen that the solution is given by solving the generalized eigenvalue problem

$$\boldsymbol{M}\boldsymbol{\theta} = \lambda \boldsymbol{N}\boldsymbol{\theta}, \tag{12}$$

for the smallest generalized eigenvalue λ. The choice of $\boldsymbol{N} = \boldsymbol{I}$ corresponds to the standard normalization. However, the solution depends on $\boldsymbol{N}$, and it has been reported by many researchers that the choice $\boldsymbol{N} = \boldsymbol{I}$ leads to large statistical bias. For example, the ellipse thus fitted to a point sequence is almost alway smaller than the true one.

One naturally asks: What $\boldsymbol{N}$ will maximize the accuracy of the solution? It is widely recognized that the scheme due to Taubin [9] produces a fairly accurate estimate. Recently, it has been found that if $\boldsymbol{N}$ is allowed to be nondefinite, i.e., neither positive or negative (semi)definite, the statistical bias can be eliminated up to second order noise terms, resulting in a method called *HyperLS* with slightly better performance than the Taubin method [7].

Recommended Readings

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

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

[3] Kanatani, K. (2008). Statistical optimization for geometric fitting: Theoretical accuracy analysis and high order error analysis. Int. J. Comput. Vis., 80(2), 167–188.

[4] Kanatani, K., Sugaya, Y. (2010). Unified computation of strict maximum likelihood for geometric fitting. J. Math. Imaging Vis., 38(1), 1–13.

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

[6] Neyman, J., Scott, E.L. (1948). Consistent estimates based on partially consistent observations. Econometrica, 16(1), 1–32.

[7] P. Rangarajan, K. Kanatani, H. Niitsuma, Y. Sugaya (2010). Hyper least squares and its applications. Proc. 20th Int. Conf. Pattern Recognition, August 2010, Istanbul, Turkey.

[8] Sampson, P.D. (1982). Fitting conic sections to "very scattered" data: An iterative refinement of the Bookstein algorithm. Comput. Graphics Image Process., 18(1), 97–108.

[9] Taubin, G. (1991). Estimation of planar curves, surfaces, and non-planar space curves defined by implicit equations with applications to edge and range image segmentation. IEEE Trans. Patt. Anal. Mach. Intell., 13(11), 1115–1138.