Principles of Covariance Propagation

Robert M. Haralick

Intelligent Systems Laboratory
Department of Electrical Engineering
University of Washington
Seattle, WA 98195
U.S.A

ABSTRACT

This paper describes how to propagate approximately additive random perturbations through any kind of vision algorithm step in which the appropriate random perturbation model for the estimated quantity produced by the vision step is also an additive random perturbation. We assume that the vision algorithm step can be modeled as a calculation (linear or non-linear) that produces an estimate that minimizes an implicit scalar function of the input quantity and the calculated estimate. The only assumption is that the scalar function be non-negative, have finite first and second partial derivatives, that its value is zero for ideal data, and that the random perturbations are small enough so that the relationship between the scalar function evaluated at the ideal but unknown input and output quantities and evaluated at the observed input quantity and perturbed output quantity can be approximated sufficiently well by a first order Taylor series expansion.

The paper finally discusses the issues of verifying that the derived statistical behavior agrees with the experimentally observed statistical behavior.

Keywords: Error propagation, computer vision, covariance

1. INTRODUCTION

Each real computer vision problem begins with one or more noisy images and has many algorithmic steps. Development of the best algorithm requires understanding how the uncertainty due to the random perturbation affecting the input image(s) propagates through the different algorithmic steps and results in a perturbation on whatever quantities are finally computed. Perhaps a more accurate statement would be that the quantities finally computed must really be considered to be estimated quantities.

In contrast to this point of view is the view prevalent in digital geometry as can be seen from the papers in this proceedings. Digital geometry has developed methods to take care of the quantizing error arising from the spatial sampling of digital images. For example, there are digital geometry methods for determining whether a digital arc arises from a straight line, whether two or more lines are colinear, whether two or more lines are parallel, as well as methods for determining bounds for the error induced by the quantizing on computations such as centroids. And there are digital geometry methods for computing skeletons. Common to most of these methods is the perspective that the only error is the quantizing error. However, in most machine vision problems this assumption is certainly not true. There is measurement error, lens distortion error, feature extraction error due to noise, occlusion error, and quantizing error. In this case, the question of whether a digital arc could have arisen from a straight line should not be answered in the negative if it turns out that a straight line cannot pass through the ribbon determined by the pixels of the digital arc. This is because some of the pixels themselves may be in erroneous positions due to uncorrected lens distortion error or noise, which may in much of the image cause positional error that even exceed the quantizing error. In this case, we should regard the pixel positions on a given digital arc as having been randomly perturbed and questions such as whether two digital arcs could have arisen from parallel line segments should be answered statistically, where the basis of the statistical test is the covariance of the estimated quantities. In the case of the parallel line question these estimated quantities are the angles of the lines segments and a proper test would examine the squared modulo difference of the angle estimates normalized by the sum of the estimated angle variances.

E-mail: haralick@ee.washington.edu

Once we have the perspective that what we compute are estimates, then it becomes clear that even though the different ways of estimating the same quantity typically yield the same result if the input quantities are not affected by a random perturbation, it is certainly not the case that the different ways of estimating the same quantities yield an estimate with the same distribution when the input is perturbed by a random perturbation. It is clearly the case that the distribution of the estimate depends on the distribution of the input random perturbation and the method or type of estimate. An example of this is the typically poor performance of matching schemes in noisy data using perspective projection invariants over matching schemes that use the full perspective projective equations. However, both work perfectly with noiseless data.

With this in mind, it is then important to understand how to propagate a random perturbation through any algorithm step in a vision problem. The difficulty is that the steps are not necessarily linear computations, the random perturbations are not necessarily additive, and the appropriate kinds of perturbations change from algorithm step to algorithm step. Nevertheless, there are many computer vision and image analysis algorithm steps in which the appropriate kind of random perturbation is additive or approximately additive. For these kinds of steps one basic measure of the size of the random perturbation is given by the covariance matrix of the estimate.

In this paper, we describe how to propagate the covariance matrix of an input random perturbation through any kind of a calculation (linear or non-linear) that extremizes an implicit scalar function, with or without constraints, of the perturbed input quantity and the calculated output estimate. The only assumption is that the scalar function to be extremized have finite first and second order partial derivatives and that the random perturbations are small enough so that the relationship between the scalar function evaluated at the ideal but unknown input and output quantities and the observed input quantity and perturbed output quantity can be approximated sufficiently well by a first order Taylor series expansion. The propagation relationships do not depend on what algorithm is used to extremize the given scalar function.

As a related case, the given propagation relationships also show how to propagate the covariance of the coefficients of a function for which we wish to find a zero to the covariance of any zero we can find.

The analysis techniques of propagation of errors is well known in the photogrammetry literature. The Manual of Photogrammetry (Slama, 1980) has a section showing how to determine the variance of Y where $Y = F(X)$ from the variance of X. The generalization of this to find the covariance matrix for Y given the covariance matrix for X is rather straightforward. Just expand F around the mean of X in a first order Taylor expansion and consider that Y is a linear function T of X. Once the coefficients of the linear combination is known, so that the randomness of Y can be approximated by $Y - \mu_Y = T(X - \mu_X)$, then the covariance matrix Σ_{YY} of Y is easily seen to be given in terms of T and the covariance matrix Σ_{XX} of X by $\Sigma_{YY} = T\Sigma_{XX}T'$ (Mikhail, 1976; Koch, 1987). This only works well for cases where the function F can be given explicitly. The problem we discuss here is one in which the function F is not given explicitly, but Y is related to X in a specific way. The techniques we employ are well-known in statistical and engineering communities. There is nothing sophisticated in the derivation. However, this technique is perhaps not so well known in the computer vision community. There are many recent vision-related papers that could be cited to illustrate this.

The paper concludes with a discussion of how to validate that the software which we use to accomplish the calculation we desire actually works. We argue that this validation can be done by comparing the predicted statistical behavior with the experimentally observed statistical behavior in a set of controlled experiments.

2. THE ABSTRACT MODEL

The abstract model has three kinds of objects. The first kind of object relates to the measurable quantities or data. There is the unobserved $N \times 1$ vector X of the ideal unperturbed measurable quantities. We assume that each component of X is some real number. Added to this unobserved ideal unperturbed vector is an $N \times 1$ unobserved random vector $\triangle X$ of noise. The observed quantity is the randomly perturbed vector $X + \triangle X$.

The second kind of object relates to the unknown parameters. There is the unobserved $K \times 1$ vector Θ. We assume that each component of Θ is some real number. Added to this ideal unperturbed vector is a $K \times 1$ unobserved vector $\triangle\Theta$ that is the random perturbation on Θ induced by the random perturbation $\triangle X$ on X. The calculated quantity is the randomly perturbed parameter vector $\hat{\Theta} = \Theta + \triangle\Theta$.

The meaning of the data vector X and the parameter vector Θ is that there is a physical process which produces X on the basis of Θ. The law governing this production process is known. The third kind of object directly relates to this law. It is a continuous non-negative scalar valued function F which relates the unobserved vectors X and Θ: $F(X,\Theta)=0$. Since F is non-negative, this is the smallest value that F can take. Therefore, for a given X, the corresponding Θ must minimize $F(X,\Theta)$.

Of course neither X nor Θ are observed. Rather only $\hat{X}$, the randomly perturbed value of X is observed. From it we desire to infer the value for Θ. But because $\hat{X}=X+\triangle X$ is random, the inferred value $\hat{\Theta}=\Theta+\triangle\Theta$ that we compute for Θ will be random. It will not be the case that $F(\hat{X},\hat{\Theta})=0$. However, the estimation problem that we set up to infer a value $\hat{\Theta}$ for Θ will minimize $F(\hat{X},\hat{\Theta})$. Therefore, in this situation it is natural to require that the function F to have finite first and second partial derivatives with respect to each component of Θ and X, including all second mixed partial derivatives taken with respect to a component of Θ and with respect to a component of X.

The basic inference problem is: given $\hat{X}=X+\triangle X$, determine a $\hat{\Theta}=\Theta+\triangle\Theta$ to minimize $F(\hat{X},\hat{\Theta})$ given the fact that Θ minimizes $F(X,\Theta)$. For this estimate $\hat{\Theta}$ we want to compute its covariance matrix.

If $\hat{\Theta}$ is computed by an explicit function h, so that $\hat{\Theta}=h(\hat{X})$, the function F is just given by $F(X,\Theta)=(\Theta-h(X))'(\Theta-h(x))$. However, our development will handle as well the determining of the covariance of a $\hat{\Theta}$ which is known to minimize $F(\hat{X},\hat{\Theta})$, without requiring any knowledge of how the minimizing $\hat{\Theta}$ was computed.

It is not unusual for some computer vision problems to be constrained problems. In this case the parameter vector Θ satisfies some constraints which we represent as $s(\Theta)=0$. The unobserved ideal Θ, satisfying the constraints $s(\Theta)=0$, and the unobserved ideal X minimize the scalar function F. In the constrained problem, $\hat{X}=X+\triangle X$ is observed and the problem is to determine that $\hat{\Theta}=\Theta+\triangle\Theta$ satisfying the constraints $s(\hat{\Theta})=0$ which minimizes $F(\hat{X},\hat{\Theta})$.

We will see that the covariance matrix for $\hat{\Theta}$ will be a function of the unobserved unperturbed X and Θ, the covariance matrix for the perturbation $\triangle X$, and the partial derivatives of F evaluated at X and Θ. We will be able to develop estimates for this covariance matrix in terms of the observed $\hat{X}$, the inferred $\hat{\Theta}$, the covariance matrix for the perturbation $\triangle X$, and the partial derivatives of F evaluated at $\hat{X}$ and $\hat{\Theta}$.

Finally, we say what this abstract model is not. It is not a model for the general problem in which the covariance matrix for $\hat{X}$ is known and the inferred value for $\hat{\Theta}$ minimizes a non-negative $F(\hat{X},\hat{\Theta})$. It is not a model for this problem because this problem does not have the assumption that there is an ideal X and Θ and the ideal Θ minimizes $F(X,\Theta)$ and this minimum value of F is 0.

3. EXAMPLE COMPUTER VISION PROBLEMS

There is a rich variety of computer vision problems which fit the form of the abstract model. In this section we outline a few of them, specifically: curve fitting,(Koch,1987) coordinated curve fitting, local feature extraction, exterior orientation, and relative orientation. Other kinds of calculations in computer vision such as calculation of curvature, invariants, vanishing points, or points at which two or more curves intersect, or problems such as motion recovery (Jerian and Jain, 1984) are all examples of problems which can be put in the abstract form as given above.

3.1. Curve Fitting

In the general curve fitting scenario, there is the unknown free parameter vector, Θ, of the curve and the set of unknown ideal points on the curve $\{x_1,\ldots,x_N\}$. Each of the ideal points is then perturbed. If $\triangle x_n$ is the random noise perturbation of the n^{th} point, then the observed point n^{th} point is $\hat{x}_n=x_n+\triangle x_n$. The form of the curve is given by a known function f which relates a point on the curve to the parameters of the curve. That is, for each ideal point x_n we have $f(x_n,\Theta)=0$. We also assume that the parameters of the curve satisfy its own set of constraint equations: $h(\Theta)=0$. The curve fitting problem is then to find an estimate $\hat{\Theta}$ to minimize $\Sigma_{n=1}^{N}f^2(\hat{x}_n,\hat{\Theta})$ subject to $h(\hat{\Theta})=0$. To put this problem in the form of the abstract problem we let

$$\begin{aligned} X &= (x_1,\ldots,x_N) \\ \hat{X} &= (x_1+\triangle x_1,\ldots,x_n+\triangle x_N) \\ F(X,\Theta,\Lambda) &= \Sigma_{n=1}^{N}f^2(x_n,\psi)+h(\Theta)'\Lambda \end{aligned}$$

Then the curve fitting problem is to find $\hat{\Theta}$ and $\hat{\Lambda}$ to minimize $F(\hat{X},\hat{\Theta},\hat{\Lambda})$ where $F(X,\Theta,\Lambda)=0$.

3.2. Coordinated Curve Fitting

In the coordinated curve fitting problem, multiple curves have to be fit on independent data, but the fitted curves have to satisfy some joint constraint. We illustrate the discussion in this section with a coordinated fitting of two curves and a constraint that the two curves must have some common point at which they are tangent.

Let $(x_1, \ldots, x_I)$ be the ideal points which are associated with the first curve whose parameters are ψ_1 and whose constraint is $h_1(\psi_1) = 0$. Each point x_i satisfies $f_1(x_i, \psi_1) = 0,\ i = 1, \ldots, I$.

Likewise, let $(y_1, \ldots, y_J)$ be the ideal points which are associated with the second curve whose parameters are ψ_2 and whose constraint is $h_2(\psi_2) = 0$. Each point y_j satisfies $f_2(y_j, \psi_2) = 0,\ j = 1, \ldots, J$.

The coordinated constraint is that for some unknown z,

$$\begin{aligned} f_1(z, \psi_1) &= 0 \\ f_2(z, \psi_2) &= 0 \\ \frac{\partial f_1}{\partial z}(z, \psi_1) &= \frac{\partial f_2}{\partial z}(z, \psi_2) \end{aligned}$$

The observed points $\hat{x}_i$ and $\hat{y}_j$ are related to the corresponding ideal points by

$$\begin{aligned} \hat{x}_i &= x_i + \triangle x_i \\ \hat{y}_j &= y_j + \triangle y_j \end{aligned}$$

To put this problem in the framework of the abstract model, we take

$$\begin{aligned} \hat{X} &= (\hat{x}_1, \ldots, \hat{x}_I, \hat{y}_1, \ldots, \hat{y}_J) \\ \hat{\Theta} &= (\hat{\psi}_1, \hat{\psi}_2, \hat{z}) \\ \hat{\Lambda} &= (\hat{\lambda}_1, \hat{\lambda}_2, \hat{\lambda}_3, \hat{\lambda}_4, \hat{\lambda}_5) \end{aligned}$$

and define

$$\begin{aligned} F(\hat{X}, \hat{\Theta}, \hat{\Lambda}) &= \Sigma_{i=1}^{I} f_1^2(\hat{x}_i, \hat{\psi}_1) + \Sigma_{j=1}^{J} f_2^2(y_j, \psi_2) + \hat{\lambda}_1 h_1(\hat{\psi}_1) + \hat{\lambda}_2 h_2(\hat{\psi}_2) \\ &+ \hat{\lambda}_3 f_1(z, \hat{\psi}_1) + \hat{\lambda}_4 f_2(z, \hat{\psi}_2) + \hat{\lambda}_5 [\frac{\partial f_1}{\partial z}(z, \psi_1) - \frac{\partial f_2}{\partial z}(z, \psi_2)] \end{aligned}$$

The coordinated curve fitting problem is then to determine a $\hat{\Theta}$ and $\hat{\Lambda}$ to minimize $F(\hat{X}, \hat{\Theta}, \hat{\Lambda})$, where the perturbed $\hat{\Theta}$ is considered related to the ideal Θ by $\hat{\Theta} = \Theta + \triangle\Theta$.

3.3. Local Feature Extraction

There are a variety of local features that can be extracted from an image. Examples include edges, corners, ridges, valleys, flats, saddles, slopes, hillsides, saddle hillsides, etc. Each local feature involves the calculation of some quantities assuming that the neighborhood has the feature and then a detection is performed based on the calculated quantities. For example, in the simple gradient edge feature, the quantity calculated is the gradient magnitude and the edge feature is detected if the calculated gradient magnitude is high enough. Here we concentrate on the calculation of the quantities associated with the feature and not the detection of the feature itself.

To put this problem in the setting of the abstract problem, we let Θ be the vector of unknown free parameters of the feature and X be the unobserved neighborhood array of noiseless brightness values. We let $\hat{X}$ be the perturbed observed neighborhood array of brightness values, $\hat{X} = X + \triangle X$, and $\hat{\Theta}$ be the calculation of the required quantities from the perturbed brightness values $\hat{X}$. The form the of feature is given by the known function f which satisfies that $f(X, \Theta) = 0$. The feature extraction problem is then to find the estimate $\hat{\Theta}$ to minimize $F(\hat{X}, \Theta) = f^2(\hat{X}, \hat{\Theta})$.

3.4. Exterior Orientation

In the exterior orientation problem, there is a known 3D object model having points $(x_n, y_n, z_n), n = 1, \ldots, N$. The unobserved noiseless perspective projection of the point (x_n, y_n, z_n) is given by (u_n, v_n). The relationship between a 3D model point and its corresponding perspective projection is given by a rotation and translation of the object model point, to put it in the reference frame of the camera, followed by a perspective projection. So if ψ represents the triple of tilt angle, pan angle, and swing angle of the rotation, t represents the x-y-z-translation vector, and k represents the camera constant (the focal length of the camera lens), we can write:

$$\begin{aligned}(u_n, v_n)' &= \frac{k}{r_n}(p_n, q_n)' \text{ where} \\ (p_n, q_n, r_n)' &= R(\psi)(x_n, y_n, z_n)' + t\end{aligned}$$

and where $R(\psi)$ is the 3×3 rotation matrix corresponding to the rotation angle vector ψ.

The function to be minimized can then be written as:

$$\begin{aligned}f_n(u_n, v_n, \psi, t) &= f(u_n, v_n, x_n, y_n, z_n, \psi, t) \text{ where} \\ f(u_n, v_n, x_n, y_n, z_n, \psi, t) &= [u_n - k\frac{(1,0,0)(R(\psi)(x_n, y_n, z_n)' + t)}{(0,0,1)(R(\psi)(x_n, y_n, z_n)' + t)}]^2 \\ &+ [v_n - k\frac{(0,1,0)(R(\psi)(x_n, y_n, z_n)' + t)}{(0,0,1)(R(\psi)(x_n, y_n, z_n)' + t)}]^2\end{aligned}$$

To put this problem in the form of the abstract description we take

$$\begin{aligned}X &= (u_1, v_1, \ldots, u_n, v_n) \\ \hat{X} &= (\hat{u}_1, \hat{v}_1, \ldots \hat{u}_n, \hat{v}_n) \\ \Theta &= (\psi, t) \\ \hat{\Theta} &= (\hat{\psi}, \hat{t})\end{aligned}$$

and define

$$F(\hat{X}, \hat{\Theta}) = \Sigma_{n=1}^{N} f_n^2(\hat{u}_n, \hat{v}_n, \hat{\Theta})$$

The exterior orientation problem is then to find a $\hat{\Theta}$ to minimize $F(\hat{X}, \hat{\Theta})$, given that $F(X, \Theta) = 0$. Because F is non-negative it must be that Θ minimizes $F(X, \Theta)$.

3.5. Relative Orientation

The relative orientation problem can be put into the form of the abstract problem in a similar way to the exterior orientation problem. We let the perspective projection of the n^{th} point on the left image be (u_{nL}, v_{nL}) and the perspective projection of the n^{th} point on the right image be (u_{nR}, v_{nR}). Then we can write that

$$\begin{aligned}(u_{nL}, v_{nL})' &= \frac{k}{z_n}(x_n, y_n)' \text{ and that} \\ (u_{nR}, v_{nR})' &= \frac{k}{r_n}(p_n, q_n)\end{aligned}$$

where (p_n, q_n, r_n) is the rotated and translated model point as given in the description of the exterior orientation problem.

The observed perspective projection of the n^{th} model point is noisy and represented as $(\hat{u}_n, \hat{v}_n) = (u_n + \triangle u_n, v_n + \triangle v_n)$. Then taking

$$X = (u_{1L}, v_{1L}, u_{1R}, v_{1R}, \ldots, u_{NL}, v_{NL}, u_{NR}, v_{NR})$$

$$
\begin{aligned}
\hat{X} &= (\hat{u}_{1L}, \hat{v}_{1L}, \hat{u}_{1R}, \hat{v}_{1R}, \ldots, \hat{u}_{NL}, \hat{v}_{NL}, \hat{u}_{NR}, \hat{v}_{NR}) \\
\Theta &= (x_1, y_1, z_1, \ldots, x_N, y_N, z_N, \psi, t) \\
\hat{\Theta} &= (\hat{x}_1, \hat{y}_1, \hat{z}_1, \ldots, \hat{x}_N \hat{y}_N, \hat{z}_N, \hat{\psi}, \hat{t})
\end{aligned}
$$

the relative orientation problem is to find $\hat{\Theta}$ to minimize

$$F(\hat{X}, \hat{\Theta}) = \Sigma_{n=1}^{N} f(u_{nR}, v_{nR}, x_n, y_n, z_n, \psi, t) + f(u_{nl}, v_{nL}, x_n, y_n, z_n, 0, 0)$$

4. ZERO FINDING

Zero finding such as finding the zero of a polynomial in one or more variables occurs in a number of vision problems. Two examples are the three point perspective resection problem and some of the techniques for motion recovery. The zero finding problem is precisely in the form required for computing the covariance matrix $\Sigma_{\triangle\Theta}$ as described in the solution section. Let X be the ideal input vector and $\hat{X}$ be the observed perturbed input vector. Let Θ be a $K \times 1$ vector zeroing the $K \times 1$ function $g(X, \Theta)$; that is, $g(X, \Theta) = 0$. Finally, let $\hat{\Theta}$ be the computed vector zeroing $g(\hat{X}, \hat{\Theta})$; that is, $g(\hat{X}, \hat{\Theta}) = 0$.

5. SOLUTION: UNCONSTRAINED CASE

For the purpose of covariance determination of the computed $\hat{\Theta} = \Theta + \triangle\Theta$, the technique used to solve the extremization problem is not important, provided that there are no singularities or near singularities in the numerical computation procedure itself.

To understand how the random perturbation $\triangle X$ acting on the unobserved vector X to produce the observed vector $\hat{X} = X + \triangle X$ propagates to the random perturbation $\triangle\Theta$ on the true but unknown parameter vector Θ to produce the computed parameter vector $\hat{\Theta} = \Theta + \triangle\Theta$, we can take partial derivatives of F with respect to each of the K components of Θ forming the gradient vector g of f. The gradient g is a $K \times 1$ vector function.

$$g(X, \Theta) = \frac{\partial F}{\partial \Theta}(X, \Theta)$$

The solution $\hat{\Theta} = \Theta + \triangle\Theta$ extremizing $F(X + \triangle X, \Theta + \triangle\Theta)$, however it is calculated, must be a zero of $g(X + \triangle X, \Theta + \triangle\Theta)$. Now taking a Taylor series expansion of g around (X, Θ) we obtain to a first order approximation:

$$g^{K\times 1}(X + \triangle X, \Theta + \triangle\Theta) = g^{K\times 1}(X, \Theta) + \frac{\partial g}{\partial X}'^{\,K\times N}(X, \Theta)\triangle X^{N\times 1} + \frac{\partial g}{\partial \Theta}'^{\,K\times K}(X, \Theta)\triangle\Theta^{K\times 1}$$

But since $\Theta + \triangle\Theta$ extremizes $F(X + \triangle X, \Theta + \triangle\Theta)$, $g(X + \triangle X, \Theta + \triangle\Theta) = 0$. Also, since Θ extremizes $F(X, \Theta)$, $g(X, \Theta) = 0$. Thus to a first order approximation,

$$0 = \frac{\partial g}{\partial X}'(X, \Theta)\triangle X + \frac{\partial g}{\partial \Theta}'(X, \Theta)\triangle\Theta$$

Since the relative extremum of F is a relative minimum, the $K \times K$ matrix

$$\frac{\partial g}{\partial \Theta}(X, \Theta) = \frac{\partial f^2}{\partial^2 \Theta}(X, \Theta)$$

must be positive definite for all (X, Θ). This implies that $\frac{\partial g}{\partial \Theta}(X, \Theta)$ is non-singular. Hence $(\frac{\partial g}{\partial \Theta})^{-1}$ exists and since it is symmetric we can write:

$$\triangle\Theta = -\{\frac{\partial g}{\partial \Theta}(X, \Theta)\}^{-1} \frac{\partial g}{\partial X}'(X, \Theta)\triangle X$$

This relation states how the random perturbation $\triangle X$ on X propagates to the random perturbation $\triangle\Theta$ on Θ. If the expected value of $\triangle X$, $E[\triangle X]$, is zero, then from this relation we see the $E[\triangle\Theta]$ will also be zero, to a first order approximation.

This relation also permits us to calculate the covariance of the random perturbation $\triangle\Theta$.

$$\begin{aligned}
\Sigma_{\triangle\Theta} &= E[\triangle\Theta\triangle\Theta'] \\
&= E[-(\frac{\partial g}{\partial\Theta})^{-1}\frac{\partial g}{\partial X}'\triangle X(-(\frac{\partial g}{\partial\Theta})^{-1}\frac{\partial g}{\partial X}'\triangle X)'] \\
&= (\frac{\partial g}{\partial\Theta})^{-1}\frac{\partial g}{\partial X}'E[\triangle X\triangle X']\frac{\partial g}{\partial X}(\frac{\partial g}{\partial\Theta})'^{-1} \\
&= (\frac{\partial g}{\partial\Theta})^{-1}\frac{\partial g}{\partial X}'\Sigma_{\triangle X\triangle X}\frac{\partial g}{\partial X}(\frac{\partial g}{\partial\Theta})^{-1}
\end{aligned}$$

Thus to the extent that the first order approximation is good, (i.e. $E[\triangle\Theta]=0$), then

$$\Sigma_{\hat\Theta\hat\Theta} = \Sigma_{\triangle\Theta\triangle\Theta}$$

The way in which we have derived the covariance matrix for $\triangle\Theta$ based on the covariance matrix for $\triangle X$ requires that the matrices

$$\frac{\partial g}{\partial\Theta}(X,\Theta) \text{ and } \frac{\partial g}{\partial X}(X,\Theta)$$

be known. But X and Θ are not observed. $X+\triangle X$ is observed and by some means $\Theta+\triangle\Theta$ is then calculated. So if we want to determine an estimate $\hat\Sigma_{\hat\Theta\hat\Theta}$ for the covariance matrix $\Sigma_{\hat\Theta\hat\Theta}$, we can proceed by expanding $g(X,\Theta)$ around $g(X+\triangle X,\Theta+\triangle\Theta)$.

$$g(X,\Theta) = g(X+\triangle X,\Theta+\triangle\Theta) - \frac{\partial g}{\partial X}'(X+\triangle X,\Theta+\triangle\Theta)\triangle X - \frac{\partial g}{\partial\Theta}'(X+\triangle X,\Theta+\triangle\Theta)\triangle\Theta$$

Here we find in a similar manner,

$$\triangle\Theta = -(\frac{\partial g}{\partial\Theta}(X+\triangle X,\Theta+\triangle\Theta))^{-1}\frac{\partial g}{\partial X}(X+\triangle X,\Theta+\triangle\Theta)\triangle X$$

This motivates the estimator $\hat\Sigma_{\triangle\Theta\triangle\Theta}$ for $\Sigma_{\triangle\Theta\triangle\Theta}$ defined by

$$\hat\Sigma_{\triangle\Theta\triangle\Theta} = (\frac{\partial g}{\partial\Theta}(\hat X,\hat\Theta)^{-1}\frac{\partial g}{\partial X}'(\hat X,\hat\Theta)\Sigma_{\triangle X\triangle X}\frac{\partial g}{\partial X}(\hat X,\hat\Theta)(\frac{\partial g}{\partial\Theta}(\hat X,\hat\Theta)^{-1}$$

So to the extent that the first order approximation is good, $\hat\Sigma_{\hat\Theta\hat\Theta} = \hat\Sigma_{\triangle\Theta\triangle\Theta}$.

The relation giving the estimate $\hat\Sigma_{\hat\Theta\hat\Theta}$ in terms of the computable

$$\frac{\partial g}{\partial\Theta}(\hat X,\hat\Theta) \text{ and } \frac{\partial g}{\partial X}(\hat X,\hat\Theta)$$

means that an estimated covariance matrix for the computed $\hat\Theta=\Theta+\triangle\Theta$ can also be calculated at the same time that the estimate $\hat\Theta$ of Θ is calculated.

5.1. Bayesian Mean Estimation

Consider the case when we observe a random vector $\hat X$ which is known to come from a Normal distribution with unknown mean Θ and known covariance matrix $\Sigma_{\hat X\hat X}$. The prior distribution on Θ has mean 0 and known covariance matrix $\Sigma_{\theta\theta}$. From the observation $\hat X$ we are to find $\hat\Theta$, the most probable value for the mean Θ. In this case the function F to be minimized by choice of $\hat\Theta$ is:

$$F(\hat X,\hat\Theta) = (\hat X-\hat\Theta)'\Sigma^{-1}_{\hat X\hat X}(\hat X-\hat\Theta)+\hat\Theta'\Sigma^{-1}_{\Theta\Theta}\hat\Theta$$

In this case we can compute

$$\begin{aligned}
g(\hat X,\hat\Theta) &= \frac{\partial F}{\partial\Theta} \\
&= -2\Sigma_{\hat X\hat X}(\hat X-\Theta)'+2\Sigma_{\Theta\Theta}\Theta
\end{aligned}$$

We can find the optimal value for $\hat{\Theta}$ by solving for that $\hat{\Theta}$ that makes $g(\hat{X}, \hat{\Theta}) = 0$. We find that

$$\hat{\Theta} = (\Sigma_{\hat{X}\hat{X}} + \Sigma_{\Theta\Theta})^{-1}\Sigma_{\hat{X}\hat{X}}\hat{X}$$

From this it is easy to explicitly determine $\Sigma_{\hat{\Theta}\hat{\Theta}}$.

$$\Sigma_{\hat{\Theta}\hat{\Theta}} = (\Sigma_{\hat{X}\hat{X}} + \Sigma_{\Theta\Theta})^{-1}\Sigma_{\hat{X}\hat{X}}(\Sigma_{\hat{X}\hat{X}} + \Sigma_{\Theta\Theta})^{-1}$$

Proceeding to compute the covariance matrix of $\hat{X}$ implicitly, we have

$$\frac{\partial g}{\partial \Theta} = 2\Sigma_{\hat{X}\hat{X}} + 2\Sigma_{\Theta\Theta}$$

and

$$\frac{\partial g}{\partial X} = -2\Sigma_{\hat{X}\hat{X}}$$

Now substituting into the equation for the implicit computation of $\Sigma_{\hat{\Theta}\hat{\Theta}}$ there results

$$\begin{aligned} \Sigma_{\hat{\Theta}\hat{\Theta}} &= (2\Sigma_{\hat{X}\hat{X}} + 2\Sigma_{\Theta\Theta})^{-1}(-2\Sigma_{\hat{X}\hat{X}})\Sigma_{\hat{X}\hat{X}}(-2\Sigma_{\hat{X}\hat{X}})(2\Sigma_{\hat{X}\hat{X}} + 2\Sigma_{\Theta\Theta})^{-1} \\ &= (\Sigma_{\hat{X}\hat{X}} + \Sigma_{\Theta\Theta})^{-1}\Sigma_{\hat{X}\hat{X}}(\Sigma_{\hat{X}\hat{X}} + \Sigma_{\Theta\Theta})^{-1} \end{aligned}$$

Notice that in this case the covariance matrix for the estimate $\hat{\Theta}$ does not depend on the ideal, non-observed value for X. But this is not always the case as our next example shows.

5.2. Regression

As a special and classic case of the unconstrained optimization, we consider the regression problem of finding Θ to minimize $F(X, \Theta) = (X - J\Theta)'\Sigma_{XX}^{-1}(X - J\Theta)$. For this F,

$$g(X, \Theta) = \frac{\partial F}{\partial \Theta} = -2J'\Sigma_{XX}^{-1}J\Theta$$

Hence,

$$\frac{\partial g}{\partial \Theta} = 2J'\Sigma_{XX}^{-1}J$$

and

$$\frac{\partial g}{\partial X} = -2\Sigma_{XX}^{-1}J$$

Then,

$$\begin{aligned} \Sigma_{\Theta\Theta} &= (2J'\Sigma_{XX}^{-1}J)^{-1}(-2\Sigma_{XX}^{-1}J)\Sigma_{XX}(-2\Sigma_{XX}^{-1}J)'(2J'\Sigma_{XX}^{-1}J)^{-1} \\ &= (J'\Sigma_{XX}^{-1}J)^{-1} \end{aligned}$$

5.3. Line Fitting

Another special case of the unconstrained optimization problem is the general line-fitting problem, which we illustrate for two-dimensional data. Assume that the unobserved points unperturbed points (x_n, y_n), $n = 1, \ldots, N$, lie on a line $x_n \cos\theta + y_n \sin\theta - \rho = 0$. In the line-fitting problem, we observe $(\hat{x_n}, \hat{y_n})$, noisy instances of (x_n, y_n). $(\hat{x_n}, \hat{y_n})$ are related to (x_n, y_n) by the noise model:

$$\begin{pmatrix} \hat{x_n} \\ \hat{y_n} \end{pmatrix} = \begin{pmatrix} x_n \\ y_n \end{pmatrix} + \xi_n \begin{pmatrix} \cos\theta \\ \sin\theta \end{pmatrix}$$

where ξ_n are independent and identically distributed as $N(0, \sigma^2)$.

To estimate the best fitting line parameters $(\hat{\theta}, \hat{\rho})$ using the least squares method, we use the criterion function which is the sum of the squared distances between the observed points and the fitted line:

$$F(X,\Theta) = \sum_{n=1}^{N} (x_n \cos\theta + y_n \sin\theta - \rho)^2$$

where $X = (x_1, y_1, \ldots, x_N, y_N)$ and $\Theta = (\theta, \rho)$.

Now,

$$g^{2\times 1}(X,\Theta) = \frac{\partial F}{\partial \Theta} = \begin{pmatrix} \frac{\partial F}{\partial \theta} \\ \\ \frac{\partial F}{\partial \rho} \end{pmatrix}$$

Letting

$$\begin{aligned}
\mu_x &= \frac{1}{N}\sum_{n=1}^{N} x_n \\
\mu_y &= \frac{1}{N}\sum_{n=1}^{N} y_n \\
S_x^2 &= \sum_{n=1}^{N}(x_n - \mu_x)^2 \\
S_y^2 &= \sum_{n=1}^{N}(y_n - \mu_y)^2 \\
S_{xy} &= \sum_{n=1}^{N}(x_n - \mu_x)(y_n - \mu_y);
\end{aligned}$$

we can compute

$$\frac{\partial F}{\partial \theta} = (S_y^2 - S_x^2 + N(\mu_y^2 - \mu_x^2))\sin 2\theta + 2(S_{xy} + N\mu_x\mu_y)\cos 2\theta + 2N\rho(\mu_x \sin\theta - \mu_y cos\theta)$$

$$\frac{\partial F}{\partial \rho} = -2N(\mu_x \cos\theta + \mu_y \sin\theta - \rho)$$

Then,

$$\frac{\partial g}{\partial \Theta}^{2\times 2} = \begin{pmatrix} \frac{\partial g}{\partial \theta} \\ \\ \frac{\partial g}{\partial \rho} \end{pmatrix} = \begin{pmatrix} \frac{\partial^2 F}{\partial \theta^2} & \frac{\partial^2 F}{\partial \theta \partial \rho} \\ \\ \frac{\partial^2 F}{\partial \rho \partial \theta} & \frac{\partial^2 F}{\partial \rho^2} \end{pmatrix}$$

where

$$\frac{\partial^2 F}{\partial \theta^2} = 2[S_y^2 - S_x^2 + N(\mu_y^2 - \mu_x^2)]\cos 2\theta - 4(S_{xy} + N\mu_x\mu_y)\sin 2\theta + 2N\rho(\mu_x \cos\theta + \mu_y \sin\theta)$$

$$\frac{\partial^2 F}{\partial \rho^2} = 2N$$

$$\frac{\partial^2 F}{\partial \theta \partial \rho} = \frac{\partial^2 F}{\partial \rho \partial \theta} = 2N(\mu_x \sin\theta - \mu_y \cos\theta)$$

and

$$\frac{\partial g}{\partial X}^{\prime\ 2\times 2N} = \underbrace{\begin{pmatrix} \frac{\partial^2 F}{\partial \theta \partial x_1} & \frac{\partial^2 F}{\partial \theta \partial y_1} & \frac{\partial^2 F}{\partial \theta \partial x_2} & \frac{\partial^2 F}{\partial \theta \partial y_2} & \frac{\partial^2 F}{\partial \theta \partial x_n} & \cdots & \frac{\partial^2 F}{\partial \theta \partial y_n} \\ \\ \frac{\partial^2 F}{\partial \rho \partial x_1} & \frac{\partial^2 F}{\partial \rho \partial y_1} & \frac{\partial^2 F}{\partial \rho \partial x_2} & \frac{\partial^2 F}{\partial \rho \partial y_2} & \cdots & \frac{\partial^2 F}{\partial \rho \partial x_N} & \frac{\partial^2 F}{\partial \rho \partial y_N} \end{pmatrix}}_{2\times 2N} \quad (1)$$

where

$$\begin{aligned}
\frac{\partial^2 F}{\partial\theta\partial x_n} &= 2[-x_n \sin 2\theta + y_n \cos 2\theta + \rho \sin\theta] \\
\frac{\partial^2 F}{\partial\theta\partial y_n} &= 2[x_n \cos 2\theta + y_n \sin 2\theta - \rho \cos\theta] \\
\frac{\partial^2 F}{\partial\rho\partial x_n} &= -2\cos\theta \\
\frac{\partial^2 F}{\partial\rho\partial y_n} &= -2\sin\theta
\end{aligned}$$

Since the parametric equation of the line is given by

$$\begin{pmatrix} x_n \\ y_n \end{pmatrix} = \rho \begin{pmatrix} \cos\theta \\ \sin\theta \end{pmatrix} + \lambda_n \begin{pmatrix} -\sin\theta \\ \cos\theta \end{pmatrix}$$

substituting the above expressions for x_n and y_n into the partial derivatives, we obtain

$$\begin{aligned}
\frac{\partial^2 F}{\partial\theta\partial x_n} &= 2\lambda_n \cos\theta \\
\frac{\partial^2 F}{\partial\theta\partial y_n} &= 2\lambda_n \sin\theta \\
\frac{\partial^2 F}{\partial\rho\partial x_n} &= -2\cos\theta \\
\frac{\partial^2 F}{\partial\rho\partial y_n} &= -2\sin\theta
\end{aligned}$$

For the given noise model, the covariance matrix Σ_{XX} is given by:

$$\Sigma_{XX} = \sigma^2 \begin{pmatrix}
\cos^2\theta & \sin\theta\cos\theta & \dots & 0 & 0 & 0 \\
\sin\theta\cos\theta & \sin^2\theta & 0 & \dots & 0 & 0 \\
0 & 0 & \cos^2\theta & \sin\theta\cos\theta & \dots & 0 \\
0 & 0 & \sin\theta\cos\theta & \sin^2\theta & \dots & 0 \\
\vdots & & & & & \\
0 & 0 & \dots & 0 & \cos^2\theta & \sin\theta\cos\theta \\
0 & 0 & 0 & \dots & \sin\theta\cos\theta & \sin^2\theta
\end{pmatrix}$$

Now we can easily do the required multiplications.

$$\frac{\partial g}{\partial X}' \Sigma_{XX} \frac{\partial g}{\partial X} = 4\sigma^2 \begin{pmatrix} \sum_{n=1}^N \lambda_n^2 & -\sum_{n=1}^N \lambda_n \\ -\sum_{n=1}^N \lambda_n & N \end{pmatrix}$$

Define

$$
\begin{aligned}
\mu_\lambda &= \frac{1}{N}\sum_{n=1}^{N}\lambda_n \\
S_\lambda^2 &= \sum_{n=1}^{N}(\lambda_n - \mu_\lambda)^2
\end{aligned}
$$

Then we have that

$$
\begin{aligned}
\mu_x &= \rho\cos\theta - \mu_\lambda \sin\theta \\
\mu_y &= \rho\sin\theta + \mu_\lambda \cos\theta \\
S_x^2 &= \sin^2\theta S_\lambda^2 \\
S_y^2 &= \cos^2\theta S_\lambda^2 \\
S_{xy} &= -\sin\theta\cos\theta S_\lambda^2
\end{aligned}
$$

Thus,

$$
\begin{aligned}
\mu_y^2 - mu_x^2 &= (\mu_\lambda^2 - \rho^2)\cos 2\theta + 2\rho\mu_\lambda \sin 2\theta \\
\mu_x\mu_y &= \frac{\rho^2 - \mu_\lambda^2}{2}\sin 2\theta + \rho\mu_\lambda\cos 2\theta \\
\mu_x\cos\theta + \mu_y\sin\theta &= \rho
\end{aligned}
$$

Then, after substituting these expressions and simplifying,

$$
\begin{pmatrix} \frac{\partial^2 F}{\partial\theta^2} & \frac{\partial^2 F}{\partial\theta\rho} \\ \frac{\partial^2 F}{\partial\rho\theta} & \frac{\partial^2 F}{\partial\rho^2} \end{pmatrix} = \begin{pmatrix} 2(S_\lambda^2 + \mu_\lambda^2) & -2N\mu_\lambda \\ -2N\mu_\lambda & 2N \end{pmatrix}
$$

Hence,

$$
\begin{pmatrix} \frac{\partial^2 F}{\partial\theta^2} & \frac{\partial^2 F}{\partial\theta\rho} \\ \frac{\partial^2 F}{\partial\rho\theta} & \frac{\partial^2 F}{\partial\rho^2} \end{pmatrix}^{-1} = \frac{1}{2NS_\lambda^2}\begin{pmatrix} N & N\mu_\lambda \\ N\mu_\lambda & S_\lambda^2 + N\mu_\lambda^2 \end{pmatrix}
$$

Using these expressions, the covariance matrix of Θ, $\Sigma_{\Theta\Theta}$, can be computed as:

$$
\begin{aligned}
\Sigma_\Theta^{2\times 2} &= \begin{pmatrix} \sigma_{\theta\theta} & \sigma_{\theta\rho} \\ \sigma_{\rho\theta} & \sigma_{\rho\rho} \end{pmatrix} \\
&= \frac{\partial g}{\partial\Theta}^{-1}(X,\Theta)\frac{\partial g}{\partial X}'(X,\Theta)\Sigma_{XX}\frac{\partial g}{\partial X}(X,\Theta)\frac{\partial g}{\partial\Theta}^{-1}(X,\Theta)
\end{aligned}
$$

We will find that

$$
\Sigma_{\Theta\Theta} = \begin{pmatrix} \frac{1}{S_\lambda^2} & \frac{\mu_\lambda}{S_\lambda^2} \\ \frac{\mu_\lambda}{S_\lambda^2} & \frac{1}{N} + \frac{\mu_\lambda^2}{S_\lambda^2} \end{pmatrix}
$$

This result has a simple geometric interpretation. In the coordinate system of the line where 0 is the point on the line closest to the origin, μ_λ is the mean position of the points and S_λ^2 is the scatter of the points. The value of μ_λ acts like a length of an arm relative to a moment calculation. If the mean position of the points on the line is a distance of $|\mu_\lambda|$ from the origin on the line, then the variance of the estimated ρ increases by $\mu_\lambda^2\sigma^2/S_\lambda^2$. This says that the variance of the estimate ρ is not invariant to the translation of the coordinate system, a fact that is typically overlooked.

An immediate application of having the covariance of the estimated parameters of fitted lines is for grouping. One of the grouping questions is whether or not two fitted line segments should be grouped together because they are part of the same line. Depending on the grouping application, it may make a difference how far apart the line segments are. However the issue of whether the fitted line segments could have arisen from the same underlying line can in either case be answered using the covariance of the fitted parameters.

Let $\hat{\Theta}_1$ be the (θ, ρ) estimated line parameters from the first line segment and let $\hat{\Theta}_2$ be the estimated line parameters from the second line segment. Let $\Sigma_{\hat{\Theta}_1\hat{\Theta}_1}$ be the covariance matrix associated with $\hat{\Theta}_1$ and let $\Sigma_{\hat{\Theta}_2\hat{\Theta}_2}$ be the covariance matrix associated with $\hat{\Theta}_2$. The hypothesis to be tested is that $\Theta_1 = \Theta_2$. A test statistic for this hypothesis is

$$\chi^2 = (\hat{\Theta}_1 - \hat{\Theta}_2)'(\Sigma_{\hat{\Theta}_1\hat{\Theta}_1} + \Sigma_{\hat{\Theta}_2\hat{\Theta}_2})^{-1}(\hat{\Theta}_1 - \hat{\Theta}_2)$$

Under the null hypothesis, χ^2 has a Chi-square distribution with 1 degree of freedom. We can reject the null hypothesis that $\Theta_1 = \Theta_2$ at the α significance level if $\chi^2 < T_\alpha$ where the probability that a χ^2 variate with 1 degree of freedom is less than α is T_α.

6. SOLUTION: CONSTRAINED CASE

The constrained problem is: given $\hat{X}$, determine that $\hat{\Theta}$ satifying the constraints $s(\hat{\Theta}) = 0$ which minimizes the function $F(\hat{X}, \hat{\Theta})$. Using the Lagrange multiplier method, the function to be minimized is $F(\hat{X}, \hat{\Theta}) + s(\hat{\Theta})'\hat{\Lambda}$. As before, we define $g(X, \Theta) = \frac{\partial}{\partial\Theta}F(X, \Theta)$. We must have at the minimizing $(\hat{X}, \hat{\Theta})$,

$$\frac{\partial}{\partial\Theta}(F(\hat{X}, \hat{\Theta}) + s(\hat{\Theta})'\hat{\Lambda}) = 0$$

In the case of no noise with the squared criterion function as we have been considering, $F(X, \Theta) = 0$. This is certainly the smallest F can be given that F is a squared criterion function. Hence it must be that $g(X, \Theta) = \frac{\partial F}{\partial\Theta}(X, \Theta) = 0$. This implies that $\frac{\partial s}{\partial\Theta}(\Theta)\Lambda = 0$, which will only happen when Λ=0 since we expect $\frac{\partial s}{\partial\Theta}$, a $K \times L$ matrix where $K > L$, to be of full rank.

Define

$$S(X, \Theta, \Lambda) = \begin{pmatrix} g(X, \Theta) + \frac{\partial s}{\partial\Theta}\Lambda \\ s(\Theta) \end{pmatrix}$$

Taking a Taylor series expansion of S,

$$S(X, \Theta, \Lambda) = S(X + \triangle X, \Theta + \triangle\Theta, \Lambda + \triangle\Lambda) - \frac{\partial S}{\partial X}'\triangle X - \frac{\partial S}{\partial\Theta}'\triangle\Theta - \frac{\partial S}{\partial\Lambda}'\triangle\Lambda$$

Because Θ satifies the constraints $s(\Theta) = 0$ and the pair (X, Θ) minimizes $F(x, \Theta)$, it follows that $S(X, \Theta, \Lambda) = 0$. Furthermore, at the computed $\hat{\Theta} = \Theta + \triangle\Theta$ and $\hat{\Lambda} = \Lambda + \triangle\Lambda$, $S(X + \triangle X, \Theta + \triangle\Theta, \Lambda + \triangle\Lambda) = 0$. Hence,

$$-\frac{\partial S}{\partial X}'\triangle X = \frac{\partial S}{\partial\Theta}'\triangle\Theta + \frac{\partial S}{\partial\Lambda}'\triangle\Lambda$$

Writing this equation out in terms of g and s, and using the fact that $\Lambda = 0$, there results

$$\begin{pmatrix} \frac{\partial g}{\partial\Theta} & \frac{\partial s}{\partial\Theta} \\ \frac{\partial s'}{\partial\Theta} & 0 \end{pmatrix} \begin{pmatrix} \triangle\Theta \\ \triangle\Lambda \end{pmatrix} = \begin{pmatrix} -\frac{\partial g'}{\partial X} \\ 0 \end{pmatrix} \triangle X$$

From this it follows that

$$\Sigma_{\triangle\Theta,\triangle\Lambda} = A^{-1}B\Sigma_{XX}B'A$$

where

$$A = \begin{pmatrix} \frac{\partial g}{\partial \Theta}, & \frac{\partial s}{\partial \Theta} \\ \frac{\partial s}{\partial \Theta}' & 0 \end{pmatrix}$$

and

$$B = -\begin{pmatrix} \frac{\partial g'}{\partial X} \\ 0 \end{pmatrix}$$

and all functions are evaluated at Θ and X. For the estimated value $\hat{\Sigma}_{\Delta\Theta\Delta\Lambda}$ of $\Sigma_{\Delta\Theta\Delta\Lambda}$, we evaluate all functions at $\hat{\Theta}$ and $\hat{\Lambda}$.

As a special but classic case of this consider the constrained regression problem to find Θ minimizing

$$F(X,\Theta) = (X - J\Theta)'(X - J\Theta)$$

subject to $H'\Theta = 0$. In this case,

$$A = \begin{pmatrix} 2J'J & H \\ H' & 0 \end{pmatrix}$$

and

$$B = -\begin{pmatrix} 2J' \\ 0 \end{pmatrix}$$

Then

$$A^{-1} = \begin{pmatrix} (2J'J)^{-1}[I - H(H'(2J'J)^{-1}H)^{-1}H'(2JJ')^{-1}] & (2J'J)^{-1}H(H'(2J'J)^{-1}H)^{-1} \\ (H'(2J'J)^{-1}H)^{-1}H'(2J'J)^{-1} & -(H'(2J'J)^{-1}H)^{-1} \end{pmatrix}$$

and

$$A^{-1}B = -\begin{pmatrix} (2J'J)^{-1}[I - H(H'(2J'J)^{-1}H)^{-1}H'(2JJ')^{-1}]2J' \\ (H'(2J'J)^{-1}H)^{-1}H'(2J'J)^{-1}2J' \end{pmatrix}$$

From this it directly follows that if $\Sigma_{XX} = \sigma^2 I$, then

$$\Sigma_{\Theta\Theta} = \sigma^2(J'J)^{-1}[I - H(H'(JJ')^{-1}H)^{-1}H'(J'J)^{-1}]$$

7. METHOD LIMITATIONS

In this section we briefly discuss when the derivations provided in the previous section are valid. They are valid when first order approximations in the gradient expansion are valid, meaning that the second derivatives of the gradient of the criterion function must be uniformly much smaller than the first derivatives of the gradient.

The following is an example invented by Georgy Gimelfarb of a criterion function of scaler variables for which there are some locations in which the second derivative of the gradient of the criterion function is not much smaller than the first derivatives of the gradient.

$$F(X,\Theta) = \begin{cases} (X-\Theta)^2 + K(X_0 - \Theta)^2 & \text{if } \Theta \geq X_0 \\ (X-\Theta)^2 & \text{if } -X_0 < \Theta < X_0 \\ (X-\Theta)^2 + K(X_0 - \Theta)^2 & \text{if } \Theta \leq -X_0 \end{cases}$$

For this criterion function, the gradient is

$$g(X,\Theta) = \begin{cases} -2(X-\Theta) - 2K(X_0 - \Theta) & \text{if } \Theta \geq X_0 \\ -2(X-\Theta) & \text{if } -X_0 < \Theta < X_0 \\ -2(X-\Theta) - 2K(X_0 - \Theta) & \text{if } \Theta \leq -X_0 \end{cases}$$

And the first partial derivative of the gradient is

$$\frac{\partial g}{\partial \Theta} = \begin{cases} 2+2K & \text{if } \Theta \geq X_0 \\ 2\text{if } -X_0 < \Theta < X_0 & \\ 2+2K & \text{if } \Theta \leq -X_0 \end{cases}$$

From this it is clear that when $K > 1$, the second partial of g with respect to Θ will be infinitely large at the points $X = X_0$ and $X = -X_0$. In this case the variance of Θ as computed by the method of the previous section will be wrong.

8. CLOSEST DISTANCE

In the case that the criterion function being minimized is composed of a sum of squares where each term in the sum represents the Mahalnobis distance between an observed point and the closest point on a curve of fixed form, there is considerable simplification of the formula for the covariance of the estimated parameters of the curve. Suppose the curve has the form $h(x) = 0$, where we have supressed the free parameters of the curve. Suppose that $\hat{x}$ is the observed noisy point having covariance Σ. Then the term in the criterion function corresponding to $\hat{x}$ will be

$$\min\{(\hat{x}-x)'\Sigma^{-1}(\hat{x}-x) \mid h(x) = 0\}$$

Using first order approximations, it is not hard to show that this minimizing value is

$$\frac{h^2(\hat{x})}{\frac{\partial h(\hat{x})}{\partial x}'\Sigma\frac{\partial h(\hat{x})}{\partial x}}$$

In this case the criterion function can be written

$$\begin{aligned} F(X,\Theta) &= F(x_1,\ldots,x_n,\Theta) \\ &= \sum_{n=1}^{N} \frac{h^2(\hat{x}_n)}{\frac{\partial h(\hat{x}_n)}{\partial x_n}'\Sigma_n\frac{\partial h(\hat{x}_n)}{\partial x}} \end{aligned}$$

where Σ_n is the covariance of $\hat{x}_n$. Proceeding with this form of the criterion function, we take its partial derivative with respect to Θ.

$$\begin{aligned} g &= \frac{\partial F}{\partial \Theta} \\ &= \sum_{n=1}^{N} \frac{\partial}{\partial \Theta} \frac{f^2(x_n,\Theta)}{\frac{\partial f}{\partial x}(x_n,\Theta)'\Sigma_n\frac{\partial f}{\partial x}(x_n,\Theta)} \\ &= \sum_{n=1}^{N} \frac{\frac{\partial f}{\partial x}(x_n,\Theta)'\Sigma_n\frac{\partial f}{\partial x}(x_n,\Theta)2f(x_n,\Theta)\frac{\partial f}{\partial \Theta}(x_n,\Theta)}{\left(\frac{\partial f}{\partial x}(x_n,\Theta)'\Sigma_n\frac{\partial f}{\partial x}(x,\Theta)\right)^2} - \\ &\qquad \frac{f^2(x_n,\Theta)\frac{\partial}{\partial \Theta}\left(\frac{\partial f}{\partial x}(x_n,\Theta)'\Sigma_n\frac{\partial f}{\partial x}(x_n,\Theta)\right)}{\left(\frac{\partial f}{\partial x}(x_n,\Theta)'\Sigma_n\frac{\partial f}{\partial x}(x_n,\Theta)\right)^2} \\ &= \sum_{n=1}^{N} 2f(x_n,\Theta)\frac{\frac{\partial f}{\partial \Theta}(x_n,\Theta)}{\frac{\partial f}{\partial x}(x_n,\Theta)'\Sigma_n\frac{\partial f}{\partial x}(x,\Theta)} - \\ &\qquad f^2(x_n,\Theta)\frac{\frac{\partial}{\partial \Theta}\left(\frac{\partial f}{\partial x}(x_n,\Theta)'\Sigma_n\frac{\partial f}{\partial x}(x_n,\Theta)\right)}{\left(\frac{\partial f}{\partial x}(x_n,\Theta)'\Sigma_n\frac{\partial f}{\partial x}(x_n,\Theta)\right)^2} \end{aligned}$$

The next step requires us to take more partial derivatives. But the general form that we are taking a partial derivative is of a product of a scaler with a vector. Suppose that f is a scaler function of a $K \times 1$ vector variable Θ and that v is a $M \times 1$ vector function of a vector variable Θ. Then,

$$\frac{\partial}{\partial \Theta} f(\Theta) v(\Theta)$$

is a $K \times M$ matrix defined by

$$\frac{\partial}{\partial \Theta} f(\Theta) v(\Theta) \quad = \quad f(\Theta) \frac{\partial}{\partial \Theta} v(\Theta)' + \frac{\partial f}{\partial \Theta} v(\Theta)'$$

Using this formula, we can proceed.

$$\begin{aligned}
\frac{\partial g}{\partial \Theta} \quad = \quad & \sum_{n=1}^{N} 2 f(x_n, \Theta) \frac{\partial}{\partial \Theta} \frac{\frac{\partial f}{\partial \Theta}(x_n, \Theta)}{\frac{\partial f}{\partial x}(x_n, \Theta)' \Sigma_n \frac{\partial f}{\partial x}(x, \Theta)} + \\
& 2 \frac{\partial f}{\partial \Theta}(x_n, \Theta) \left(\frac{\frac{\partial f}{\partial \Theta}(x_n, \Theta)}{\frac{\partial f}{\partial x}(x_n, \Theta)' \Sigma_n \frac{\partial f}{\partial x}(x, \Theta)} \right)' - \\
& f^2(x_n, \Theta) \frac{\partial}{\partial \Theta} \frac{\frac{\partial}{\partial \Theta} \left(\frac{\partial f}{\partial x}(x_n, \Theta)' \Sigma_n \frac{\partial f}{\partial x}(x_n, \Theta) \right)}{\left(\frac{\partial f}{\partial x}(x_n, \Theta)' \Sigma_n \frac{\partial f}{\partial x}(x_n, \Theta) \right)^2} - \\
& 2 f(x_n, \Theta) \frac{\partial f}{\partial \Theta}(x_n, \Theta) \left(\frac{\frac{\partial}{\partial \Theta} \left(\frac{\partial f}{\partial x}(x_n, \Theta)' \Sigma_n \frac{\partial f}{\partial x}(x_n, \Theta) \right)}{\left(\frac{\partial f}{\partial x}(x_n, \Theta)' \Sigma_n \frac{\partial f}{\partial x}(x_n, \Theta) \right)^2} \right)'
\end{aligned}$$

We evaluate the partial derivative at $(x_1, \ldots, x_N)$ where

$$f(x_n, \Theta) = 0, \; n = 1, \ldots, N$$

Therefore,

$$\begin{aligned}
\frac{\partial g}{\partial \Theta} \quad & = \quad 2 \sum_{n=1}^{N} \frac{\partial f}{\partial \Theta}(x_n, \Theta) \left(\frac{\frac{\partial f}{\partial \Theta}(x_n, \Theta)}{\frac{\partial f}{\partial x}(x_n, \Theta)' \Sigma_n \frac{\partial f}{\partial x}(x_n, \Theta)} \right)' \\
& = \quad 2 \sum_{n=1}^{N} \frac{\partial f}{\partial \Theta}(x_n, \Theta) \frac{\partial f}{\partial \Theta}(x_n, \Theta)' \left(\frac{1}{\frac{\partial f}{\partial x}(x_n, \Theta)' \Sigma_n \frac{\partial f}{\partial x}(x_n, \Theta)} \right) \\
& = \quad 2 \sum_{n=1}^{N} \frac{\frac{\partial f}{\partial \Theta}(x_n, \Theta) \frac{\partial f}{\partial \Theta}(x_n, \Theta)'}{\frac{\partial f}{\partial x}(x_n, \Theta)' \Sigma_n \frac{\partial f}{\partial x}(x_n, \Theta)}
\end{aligned}$$

$$\begin{aligned}
\frac{\partial g}{\partial x_n} \quad = \quad & 2 f(x_n, \Theta) \frac{\partial}{\partial x_n} \frac{\frac{\partial f}{\partial \Theta}(x_n, \Theta)}{\frac{\partial f}{\partial x}(x_n, \Theta)' \Sigma_n \frac{\partial f}{\partial x}(x, \Theta)} + \\
& 2 \frac{\partial f}{\partial x_n}(x_n, \Theta) \left(\frac{\frac{\partial f}{\partial \Theta}(x_n, \Theta)}{\frac{\partial f}{\partial x_n}(x_n, \Theta)' \Sigma_n \frac{\partial f}{\partial x}(x, \Theta)} \right)' -
\end{aligned}$$

$$f^2(x_n,\Theta)\frac{\partial}{\partial x_n}\frac{\frac{\partial}{\partial\Theta}\left(\frac{\partial f}{\partial x}(x_n,\Theta)'\Sigma_n\frac{\partial f}{\partial x}(x_n,\Theta)\right)}{\left(\frac{\partial f}{\partial x}(x_n,\Theta)'\Sigma_n\frac{\partial f}{\partial x}(x_n,\Theta)\right)^2} -$$

$$2f(x_n,\Theta)\frac{\partial f}{\partial x_n}(x_n,\Theta)\left(\frac{\frac{\partial}{\partial\Theta}\left(\frac{\partial f}{\partial x}(x_n,\Theta)'\Sigma_n\frac{\partial f}{\partial x}(x_n,\Theta)\right)}{\left(\frac{\partial f}{\partial x}(x_n,\Theta)'\Sigma_n\frac{\partial f}{\partial x}(x_n,\Theta)\right)^2}\right)'$$

We evaluate the partial derivative at $(x_1,\ldots,x_N)$ where

$$f(x_n,\Theta)=0,\ n=1,\ldots,N$$

Therefore,

$$\begin{aligned}
\frac{\partial g}{\partial x_n} &= 2\frac{\partial f}{\partial x_n}(x_n,\Theta)\left(\frac{\frac{\partial f}{\partial\Theta}(x_n,\Theta)}{\frac{\partial f}{\partial x}(x_n,\Theta)'\Sigma_n\frac{\partial f}{\partial x}(x_n,\Theta)}\right)' \\
&= 2\frac{\partial f}{\partial x_n}(x_n,\Theta)\frac{\partial f}{\partial\Theta}(x_n,\Theta)'\left(\frac{1}{\frac{\partial f}{\partial x}(x_n,\Theta)'\Sigma_n\frac{\partial f}{\partial x}(x_n,\Theta)}\right)
\end{aligned}$$

Let,

$$\begin{aligned}
\frac{\partial g}{\partial X}' &= \left(\frac{\partial g'}{\partial x_1}\ \frac{\partial g'}{\partial x_2}\ \cdots\ \frac{\partial g'}{\partial x_N}\right) \\
&= 2\left(\frac{\frac{\partial f}{\partial\Theta}(x_1,\Theta)\frac{\partial f}{\partial x_1}(x_1,\Theta)'}{\frac{\partial f}{\partial x}(x_1,\Theta)'\Sigma_1\frac{\partial f}{\partial x}(x_1,\Theta)}\ \cdots\ \frac{\frac{\partial f}{\partial\Theta}(x_N,\Theta)\frac{\partial f}{\partial x_N}(x_N,\Theta)'}{\frac{\partial f}{\partial x}(x_N,\Theta)'\Sigma_N\frac{\partial f}{\partial x}(x_N,\Theta)}\right)
\end{aligned}$$

Then,

$$\begin{aligned}
\frac{\partial g}{\partial X}'\Sigma\frac{\partial g}{X} &= 4\sum_{n=1}^{N}\frac{\frac{\partial f}{\partial\Theta}(x_n,\Theta)\frac{\partial f}{\partial x_n}(x_n,\Theta)'\Sigma_n\frac{\partial f}{\partial x_n}(x_n,\Theta)\frac{\partial f}{\partial\Theta}(x_n,\Theta)'}{\left[\frac{\partial f}{\partial x}(x_n,\Theta)'\Sigma_n\frac{\partial f}{\partial x}(x_n,\Theta)\right]^2} \\
&= 4\sum_{n=1}^{N}\frac{\frac{\partial f}{\partial\Theta}(x_n,\Theta)\frac{\partial f}{\partial\Theta}(x_n,\Theta)'}{\frac{\partial f}{\partial x}(x_n,\Theta)'\Sigma_n\frac{\partial f}{\partial x}(x_n,\Theta)} \\
&= 2\frac{\partial g}{\partial\Theta}(X,\Theta)
\end{aligned}$$

Finally,

$$\begin{aligned}
\Sigma_{\Theta\Theta} &= \left(\frac{\partial g}{\partial\Theta}\right)^{-1}\frac{\partial g}{\partial X}'\Sigma_{XX}\frac{\partial g}{\partial X}\left(\frac{\partial g}{\partial\Theta}\right)^{-1} \\
&= \left(\frac{\partial g}{\partial\Theta}\right)^{-1}\frac{2\partial g}{\partial\Theta}\left(\frac{\partial g}{\partial\Theta}\right)^{-1} \\
&= 2\left(\frac{\partial g}{\partial\Theta}\right)^{-1} \\
&= 2\left(2\sum_{n=1}^{N}\frac{\frac{\partial f}{\partial\Theta}(x_n,\Theta)\frac{\partial f}{\partial\Theta}(x_n,\Theta)'}{\frac{\partial f}{\partial x}(x_n,\Theta)'\Sigma_n\frac{\partial f}{\partial x}(x_n,\Theta)}\right)^{-1} \\
&= \left(\sum_{n=1}^{N}\frac{\frac{\partial f}{\partial\Theta}(x_n,\Theta)\frac{\partial f}{\partial\Theta}(x_n,\Theta)'}{\frac{\partial f}{\partial x}(x_n,\Theta)'\Sigma_n\frac{\partial f}{\partial x}(x_n,\Theta)}\right)^{-1}
\end{aligned}$$

9. VALIDATION

There are two levels of validation. One level of validation is for the software. This can be tested by a large set of Monte-Carlo experiments off-line where we know what the correct answers are.

Another level of validation is on-line reliability. Here all that we have is the computed estimate and estimated covariance matrix for the estimate.

9.1. Software and Algorithm Validation

Software for performing the optimization required to compute the estimate $\hat{\Theta}$ is often complicated and it is easy for there to be errors that are not immediately observable. For example, there we may have optimization software that produces correct answers on a few known examples but fails in a significant fraction of more difficult cases that we are not specifically trying out. One approach in testing that the software is producing the right answers is to test the statistical properties of the answers. That is, we can statistically test whether the statistical properties of its answers are similar to the statistical properties we expect. These expectations are whether the mean of the computed estimates is sufficiently close to the population mean and whether the estimated covariance matrix of the estimates is sufficiently close to the population covariance matrix. Rephrasing this more precisely the test is whether the computed estimates could have arisen from a population with given mean and covariance matrix.

Consider what happens in a hypothesis test: a significance level, α, is selected. When the test is run, a test statistic, say $\hat{\phi}$, is computed. The test statistic is typically designed so that in the case that the hypothesis is true, the test statistic will tend to have its values distributed around zero, in accordance with a known distribution. If the test statistic has a value say higher than a given ϕ_0, we reject the hypothesis that the computed estimate is statistically behaved as we expected it to be. If we do not reject, then in effect, we are tentatively accepting the hypothesis. The value of ϕ_0 is chosen so that the probability that we reject the hypothesis, given that is the hypothesis is true is less than the significance level α.

The key in using this kind of testing is that we can set up an experiment in which we know what the correct answer for the no noise ideal case would be. Then we can additively perturb the input data by a normally distributed vector from a population having zero mean and given covariance matrix. Then using the analytic propagation results derived earlier in the paper, we can derive the covariance matrix of the estimates produced by software.

If we repeat this experiment many times just changing the perturbed realizations and leaving everything else the same, the experiment produces estimates $\theta_1, \ldots, \theta_N$ that will come from a normal population having mean θ, the correct answer for the ideal no noise case, and covariance matrix Σ, computed from the propagation equations. Now the hypothesis test is whether the observations $\theta_1, \ldots, \theta_N$ come fron a Normal population with mean θ and covariance matrix Σ. For this hypothesis test, there is a uniformly most powerful test. Let

$$B = \Sigma_{n=1}^{N}(\theta_n - \bar{\theta})(\theta_n - \bar{\theta})'$$

Define

$$\begin{aligned}\lambda &= (e/N)^{pN/2}|B\Sigma^{-1}|^{N/2} \\ &\quad \times exp(-\frac{1}{2}[tr(B\Sigma^{-1}) + N(\bar{\theta} - \theta)'\Sigma^{-1}(\bar{\theta} - \theta)])\end{aligned}$$

The test statistic is:

$$T = -2log\lambda$$

Under the hypothesis, T is distributed as:

$$\chi^2_{p(p+1)/2+p}$$

where p is the dimension of θ.

So to perform a test that the program's behavior is as expected we repeatedly generate the T statistic and compute its empirical distribution function. Then we test the hypothesis that T is distributed as the χ^2 variate using a Kolmogorov-Smirnov test.

9.2. On-line Reliability

For the on-line reliablity testing, the estimate is computed by minimizing the scalar objective function. Then based on the given covariance matrix of the input data, an estimated covariance matrix of the estimate is computed using the linearization around the estimate itself. An on-line reliablity test can be done by testing whether the each of the diagonal entries of the estimated covariance matrix is sufficiently small.

10. CONCLUSION

Making a successful vision system for any particular application typically requires many steps, the optimal choice of which is not always apparent. To understand how to do the optimal design, a synthesis problem, requires that we first understand how to solve the analysis problem: given the steps of a particular algorithm, determine how to propagate the parameters of the perturbation process from the input to the parameters describing the perturbation process of the computed output. The first basic case of this sort of uncertainty propagation is the propagation of the covariance matrix of the input to the covariance matrix of the output. This is what this paper has described.

This work does not come near to solving what is required for the general problem, because the general problem involves perturbations which are not additive. That is, in mid and high-level vision, the appropriate kinds of perturbations are perturbations of structures. Now, we are in the process of understanding some of the issues with these kinds of perturbations and expect to soon have some results in this area.

REFERENCES

1. Jerian, C. and Jain R., (1984) "Determining Motion Parameters for Scenes with Translation and Rotation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. PAMI-6, pp.523-530.
2. Koch, K. (1987) *Parameter Estimation and Hypothesis Testing in Linear Models*, Springer-Verlag, Berlin, p117, p121.
3. Mikhail, E. (1976) *Observations and Least Squares*, IEP – A Dun-Donnelley Publisher, New York, p72-90.
4. Slama, C. (Ed.), (1980) *The Manual of Photogrammetry* The American Society of Photogrammetry, Falls Church, VA 22046, p73-74.