CISM JOURNAL ACSGC

NONLINEARITY AND LEAST SQUARES

P.J.G. Teunissen and E.H. Knickmeyer
The University of Calgary, Calgary, Alberta

Dedicated to Professor Dr. K.P. Schwarz
on the occasion of his 50th birthday.

Since almost all functional relations in our geodetic models are nonlinear, it is important, especially from a statistical inference point of view, to know how nonlinearity manifests itself at the various stages of an adjustment. In this paper particular attention is given to the effect of nonlinearity on the first two moments of least squares estimators. Expressions for the moments of least squares estimators of parameters, residuals and functions derived from parameters, are given. The measures of nonlinearity are discussed both from a statistical and differential geometric point of view. Finally, our results are applied to the 2D symmetric Helmert transformation with a rotational invariant covariance structure.

Puisque la plupart des relations fonctionnelles dans nos modèles géodésiques sont nonlinéaires, il est important, surtout au point de vue d'inférence statistique, de connaître la façon dont la non-linéarité se manifeste au long des différentes étapes d'une compensation. Dans le présent article, on porte une attention particulière aux effets de la non-linéarité sur les deux premiers moments des estimateurs par moindres carrés. On donne des expressions pour les moments des estimateurs par moindres carrés, pour les résiduelles et les fonctions dérivées des paramètres. On parle des mesures de non-linéarité des points de vue statistique et géométrique différentiel. Finalement, on applique nos résultats à la transformation d'Helmert symétrique à deux dimensions avec une structure de covariance invariante en rotation.

1. Introduction

Almost all functional relations in our geodetic models are nonlinear. Hence, one might question whether the use of the ideas, concepts and results from the theory of linear estimation is justifiable in all cases. Of course, it may be argued that probably most nonlinear models are only moderately nonlinear and thus permit the use of a linear(ized) model. This is true. Nevertheless, we need to have ways of assessing the amount of nonlinearity in nonlinear models and methods to prove whether a linear(ized) model is a sufficient approximation. We therefore need to know how nonlinearity manifests itself at the various stages of an adjustment.

A general theoretical and practical investigation into the various aspects of nonlinear adjustment using concepts of differential geometry was started in [*Krarup* 1982] and [*Teunissen* 1984, 1985a]. In this paper we will follow [*Teunissen*, 1985a] and in particular discuss the effect of nonlinearity on the first two moments of least squares estimators. More information on probabilistic properties of nonlinear estimators can be found in [*Baarda* 1967; *Bähr* 1985; *Jeudy* 1987; *Kubik* 1986; *Schaffrin* 1983; *Teunissen* 1985b and *Wolf* 1961].

This paper will refrain from discussing the methods which can be used for computing the nonlinear estimates. For more details on these numerical methods and their properties refer to [Blaha 1987; *Kelley* and *Thompson* 1978; *Kubik* 1967; *Pope* 1972, 1974; *Saito* 1973; *Schek* and *Maier* 1976; *Stark* and *Mikhail* 1973; *Teunissen* 1984, 1985a, 1987a, 1987b].

2. Linear versus nonlinear least squares

Consider the model

$$y = A(x) + e\,,\ \mathrm{Cov.}(y) = \sigma^2 Q_y\,, \tag{2.1}$$

where y is an m-vector of normally-distributed observational variates with mean $A(x)$ and covariance matrix $\sigma^2 Q_y$, σ^2 is the variance factor of unit weight, x is an n-vector of fixed but unknown parameters and $A(.)$ is a map which maps R^n into R^m. It is assumed that Q_y is positive-definite, that $A(.)$ is injective and sufficiently smooth and that $m>n$.

We will use the least squares criterion for obtaining estimates of x and the mean of y respectively. This is probabilistically justified by the fact that under the assumption of normality least squares estimators are identical to maximum likelihood estimators. The least squares criterion reads

In: CISM Journal ACSGC, Vol. 42, No. 4, 1988, pp. 321-330

$$\min_x \| y - A(x) \|^2, \qquad (2.2)$$

with the norm $\| . \|^2 = (.)^* Q_y^{-1} (.)$.

Since, for varying values of x, $A(x)$ traces locally an n-dimensional surface or manifold, say M, embedded in R^m, the scalar $\| y - A(x) \|$ can be seen to equal the distance from y to the point $A(x)$ on M. Hence, the solution to (2.2) is given by that point on M, say $\hat{y} = A(\hat{x})$, which has least distance to y. A necessary condition for having a least distance is that $\hat{y}$ equals the orthogonal projection of y onto M.

In case map $A(.)$ is *linear*, the n-dimensional surface M traced by $A(.)$ is flat and the condition of orthogonality becomes a sufficient condition. In this case we have the well known results (see Figure 1):

$$\hat{y} = P_A y ; \hat{x} = A^- \hat{y} , \qquad (2.3)$$

with the orthogonal projector
$P_A = A(A^* Q_y^{-1} A)^{-1} A^* Q_y^{-1}$

and an arbitrary inverse A^- of A.

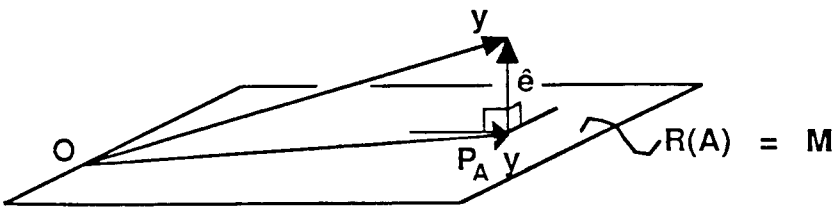

Figure 1: Linear least squares' orthogonal projection

In case map $A(.)$ is *nonlinear*, the n-dimensional surface M is curved and the condition of orthogonality fails to be sufficient. To ensure sufficiency we need in addition to orthogonality the condition $k \| y - A(\hat{x}) \| < 1$, where k is the largest curvature of M at $A(\hat{x})$ [*Teunissen* 1985a, p. 120]. If we assume sufficiency, the least squares estimators can formally be written as (see Figure 2):

$$\hat{y} = P_A(y) ; \hat{x} = A^-(\hat{y}) , \qquad (2.4)$$

where $P_A(.)$ and $A^-(.)$ will be nonlinear in general.

Due to the nonlinearity of $A(.)$ it is very seldom that closed expressions can be found for $P_A(.)$ and $A^-(.)$ respectively. In practice it will therefore be necessary to recourse to methods which are iterative in nature. One starts with an initial guess x_0 and proceeds to generate a sequence x_0, $x_1, x_2, \ldots$ which under some conditions converges to the point $\hat{x}$.

Once the estimates $\hat{x}$ and $\hat{y}$ are computed it is of course not enough to just state that these are the estimated values of the unknown x and mean of y respectively. The step following the actual estimation process is equally important. That is, it is also necessary to know the distributional properties of the estimators involved.

In case map $A(.)$ is linear, it is not difficult to derive the precise distribution of the least squares estimators. The following distributional properties are well known:

$$\hat{y} \sim N(Ax, \sigma^2 P_A Q_y), \hat{x} \sim N(x, \sigma^2 (A^* Q_y^{-1} A)^{-1}). \qquad (2.5)$$

Furthermore the Gauss-Markov theorem states that these estimators are also minimum variance linear unbiased estimators.

Unfortunately these results do not carry over to the nonlinear case. Essential properties which are used repeatedly in the development of the linear theory break down completely in the nonlinear case. Take for instance the mathematical expectation operator $E\{.\}$. If θ is a random variable and F is a nonlinear map, then

$$E\{F(\theta)\} \neq F(E\{\theta\}) , \qquad (2.6)$$

i.e., the mean of the image differs generally from the image of the mean. Hence, we can hardly expect our least squares estimators to be unbiased in the nonlinear case. Consequently least squares estimation cannot be justified anymore by referring to the Gauss-Markov theorem. Of course this by no means implies that the least squares estimators should be done away with. Under the usual assumption of normality the least squares estimators are still maximum likelihood estimators. Besides, when the importance of exactly unbiased estimators is overemphasized, one can find one's self in an impossible situation. Very often, there is a natural estimator which is, however, slightly biased. For example, if θ is a good unbiased estimator of $\tilde{\theta}$, and if it is required to estimate $F(\tilde{\theta})$, then it seems natural to estimate $F(\tilde{\theta})$ by $F(\theta)$, although this estimator will nearly always be biased. It is important however to have a means of assessing the bias and more generally for statistical inference purposes, to have a means of computing the probability distributions, or approximations to it, of the estimators involved.

3. Mean and Variance of Nonlinear Estimators

Before discussing the probabilistic properties of least squares estimators, we first discuss in this section some methods of deriving the distributional properties of general nonlinear estimators.

We will start with the univariate case. The multivariate generalization then becomes straightforward.

Let θ be a random variable and $F(.)$ a nonlinear function. Our objective is to find some of the distributional properties of $F(\theta)$. Broadly speaking four approaches suggest themselves. Most of the exact methods are however difficult to apply in practice.

Firstly, if the density $D(\theta)$ of θ is given, then theoretically at least, the distribution of $F(\theta)$ can be found. This follows since the cumulative distribution $C(y)$ of $y = F(\theta)$ satisfies

$$C(y) = \text{Prob.}\left(F(\theta) \le y\right) = \int_{\{\theta \mid F(\theta) \le y\}} D(\theta)d\theta, \tag{3.1}$$

for fixed y. Since (3.1) describes the probability of an event in terms of θ, such a probability can theoretically be determined by integrating the density of θ over the region corresponding to the event. The problem with this method is that in general the desired probability for each y cannot easily be evaluated.

Secondly, under some restrictions on $F(.)$ equation (3.1) can be worked out to give the density $D(y)$ of $y = F(\theta)$ in terms of the density $D(\theta)$ of θ:

$$D(y) = D(\theta) / \left|d_\theta F(\theta)\right|, \text{ with } \theta = F^{-1}(y). \tag{3.2}$$

Unfortunately, because of the inversion of F, it is very seldom that such an exact method can be applied with success.

The next thing that one can try to do is to derive some of the moments of the distribution of $F(\theta)$, i.e., the mean and variance:

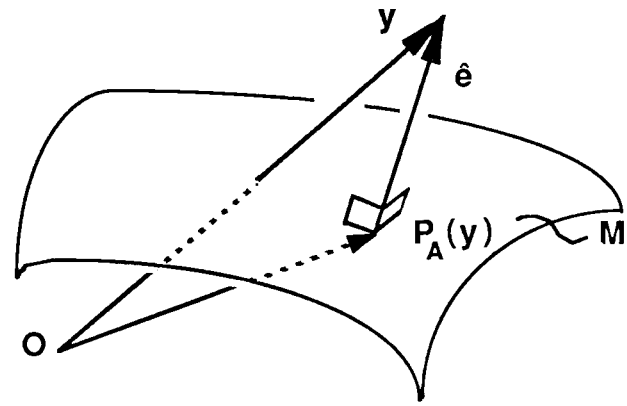

Figure 2: Nonlinear least squares' orthogonal projection

$$E\{F(\theta)\} = \int_{-\infty}^{\infty} F(\theta)D(\theta)d\theta; \quad \text{Var}\{F(\theta)\} = \int_{-\infty}^{\infty} [F(\theta) - E\{F(\theta)\}]^2 D(\theta)d\theta \tag{3.3}$$

The complexity of these computations depend very much on the nature of the functions $F(.)$ and $D(.)$. But in general they can become quite complicated, especially in the multivariate case.

As a last resort one can try to compute approximations to the mean and variance. This can be done by using a suitable Taylor expansion. If $F(.)$ is sufficiently smooth and θ is a random variable with mean $\tilde{\theta}$ and variance σ^2, expansion of $F(\theta)$ at $\tilde{\theta}$ gives:

$$F(\theta) = F(\tilde{\theta}) + d_\theta F(\tilde{\theta})(\theta - \tilde{\theta}) + \frac{1}{2} d_\theta^2 F(\tilde{\theta})(\theta - \tilde{\theta})^2 + \ldots \tag{3.4}$$

Now if σ is small enough then by Chebychev's inequality, Prob. $\left(|\theta - \tilde{\theta}| \ge \delta\right) \le \sigma^2/\delta^2$ for any $\delta > 0$, θ will depart only a little from $\tilde{\theta}$ except on rare occasions and therefore $(\theta - \tilde{\theta})$ will typically be small. The higher order terms in (3.4) can therefore be neglected. Taking the expectations of both sides of (3.4) then gives for the mean of $F(\theta)$:

$$E\{F(\theta)\} \doteq F(\tilde{\theta}) + \frac{1}{2}\sigma^2 d_\theta^2 F(\tilde{\theta}). \tag{3.5}$$

In a similar way we find for the variance of $F(\theta)$, if terms of σ^4 and higher are neglected, that

$$\text{Var.}\{F(\theta)\} \doteq \sigma^2 \left[d_\theta F(\tilde{\theta})\right]^2. \tag{3.6}$$

These approximations can easily be improved upon by including more terms of the Taylor expansion.

In the multivariate case we can take θ to be a random vector with n components θ^α, $\alpha = 1, \ldots, n$. The above approximations then generalize to:

$$\text{a)}\quad E\{F(\theta) - F(\tilde{\theta})\} \doteq \frac{1}{2}\sigma^2\partial^2_{\alpha\beta}F(\tilde{\theta})\,g^{\alpha\beta}$$

$$\text{b)}\quad \text{Var.}\{F(\theta)\} \doteq \sigma^2\partial_\alpha F(\tilde{\theta})\,g^{\alpha\beta}\,\partial_\beta F(\tilde{\theta}) \qquad (3.7)$$

where $\alpha, \beta = 1, \ldots, n$, $\partial_{\alpha\beta} F$ is the Hessian matrix of F, $\sigma^2 g^{\alpha\beta}$ is the covariance matrix of θ in index notation and Einstein's summation convention is used for repeated indices. Note that in (3.7a) one needs to compute the trace of the product of the Hessian matrix of F and the covariance matrix of θ.

Example 1

Consider the distance function $\ell(x,y) = (x^2 + y^2)^{1/2}$. Assume that x and y are uncorrelated random variables with variance σ^2 and mean $\tilde{x}$ and $\tilde{y}$ respectively. Since x and y are uncorrelated, $g^{\alpha\beta}$ is a 2x2 unit matrix. Hence we only need to compute the trace of the Hessian of $\ell(x,y)$ to find the difference between $E\{\ell(x,y)\}$ and $\ell(\tilde{x},\tilde{y}.)$.
The Hessian reads

$$\partial^2_{\alpha\beta}\,\ell(\tilde{x},\tilde{y}) = \frac{1}{\ell^3(\tilde{x},\tilde{y})}\begin{bmatrix} \tilde{y}^2 & -\tilde{x}\tilde{y} \\ -\tilde{x}\tilde{y} & \tilde{x}^2 \end{bmatrix}.$$

Thus

$$E\{\ell(x,y) - \ell(\tilde{x},\tilde{y})\} \doteq \frac{\sigma^2}{2\,\ell(\tilde{x},\tilde{y})}. \qquad (3.8)$$

Example 2

Consider the function $x = \ell \sin A$, where ℓ stands for distance and A for azimuth. Assume that ℓ and A are uncorrelated random variables with variance σ^2_ℓ and σ^2_A, and mean $\tilde{\ell}$ and $\tilde{A}$ respectively. A Taylor expansion around $\tilde{\ell}$ and $\tilde{A}$ gives

$$x = \tilde{\ell}\sin\tilde{A} + \begin{pmatrix}\sin\tilde{A} & \ell\cos\tilde{A}\end{pmatrix}\begin{bmatrix}\ell-\tilde{\ell}\\ A-\tilde{A}\end{bmatrix} + \frac{1}{2}\begin{bmatrix}\ell-\tilde{\ell}\\ A-\tilde{A}\end{bmatrix}^*\begin{bmatrix}0 & \cos\tilde{A}\\ \cos\tilde{A} & -\tilde{\ell}\sin\tilde{A}\end{bmatrix}\begin{bmatrix}\ell-\tilde{\ell}\\ A-\tilde{A}\end{bmatrix} + \ldots$$

Hence the Hessian reads

$$\partial^2_{\alpha\beta}\,x = \begin{bmatrix}0 & \cos\tilde{A}\\ \cos\tilde{A} & -\tilde{\ell}\sin\tilde{A}\end{bmatrix}.$$

Since the covariance matrix of ℓ and A is given by

$$g^{\alpha\beta} = \begin{bmatrix}\sigma^2_\ell & 0\\ 0 & \sigma^2_A\end{bmatrix},$$

it follows that the trace of the product of the Hessian and the covariance matrix equals $-\sigma^2_A\tilde{\ell}\sin\tilde{A}$. Thus:

$$E\{x - \tilde{\ell}\sin\tilde{A}\} \doteq -\frac{1}{2}\sigma^2_A\tilde{\ell}\sin\tilde{A} \qquad (3.9)$$

By inserting realistic values $\sigma_A = 10^{-5}$ and $\ell = 1$ km into (3.9) it can be seen that the difference between $E\{x\}$ and $\tilde{\ell}\sin\tilde{A}$ is completely negligible. This is typical of many geodetic situations. Note however that it is not so much the absence of severe nonlinearity which makes bias negligible, but the relatively high precision of the measurements.

4. *Mean of Nonlinear Least-Squares Estimators*

Let us now return to our model (2.1). To obtain an approximation to the mean of the least squares estimators of model (2.1) we have to modify our method of the previous section somewhat. The reason being that in general we do not have closed expressions available which express the estimators as known functions of the observables. We therefore have to invert, in some way, the Taylor expansion. The method we will use is given in [*Teunissen*, 1985a] and can be described as follows:

First assume that the least squares estimator $\hat{x}$ of x in model (2.1) can be written as a smooth enough function of the random m-vector y. Then Taylor expand this function at the mean $E\{y\}$ of y. This gives an expansion in $e = y - E\{y\}$. The problem is now to find the coefficients of this expansion, i.e., to find the partial derivatives of various order of the function relating $\hat{x}$ to y. Once these coefficients are known formula (3.7) of the previous section can be applied.

The coefficients are found in the following way. We start from the orthogonality condition $0 = \partial_x A(\hat{x})^* Q_y^{-1} e(\hat{x})$, where $\partial_x A(\hat{x})$ is the Jacobian of $A(.)$ evaluated at $\hat{x}$ and $e(\hat{x}) = y - A(\hat{x})$. Expansion of right-hand side of the orthogonality condition at x gives an expansion in $\hat{x}$-x. We now substitute our first expansion in e in the above expansion in $\hat{x}$-x. The result is a new expansion in e, which is identical to zero for all e. Hence we may collect terms of the same order and set them to zero. In this way we can recursively determine all the coefficients sought. Although the derivations

become quite lengthy and tedious it is in principle possible to obtain expressions for the mean of the least squares estimators to any desired approximation. Note however that our expansions are evaluated at the "true" unknown x. The approximations so obtained can therefore only be evaluated by substituting the estimated parameters for the true ones.

Neglecting terms of order higher than σ^2, our results for the biases in $\hat{x}$ and $\hat{e}$ read:

$$\begin{aligned} &\text{a)}\quad E\{\hat{x}-x\} \doteq Q_x\, \partial_x A(x)^* Q_y^{-1}\, b_y\,, \\ &\text{b)}\quad E\{\hat{e}\} = E\{y-\hat{y}\} \doteq P^{\perp}_{\partial_x A(x)}\, b_y\,, \text{ with} \\ &\text{c)}\quad b_y = -\frac{1}{2}\sigma^2\, \partial^2_{\alpha\beta} A(x)\; g^{\alpha\beta}\,(x)\,. \end{aligned} \tag{4.1}$$

where $\alpha, \beta = 1, \ldots, n$; Q_x equals the inverse of $\left(\partial_x A(x)^* Q_y^{-1} \partial_x A(x)\right)$, $\partial_x A(x)$ is the Jacobian of $A(.)$ evaluated at x, $P^{\perp}_{\partial_x A}(x)$ is the orthogonal projector projecting onto the orthogonal complement of the range space of $\partial_x A(x)$, i.e., $P^{\perp}_{\partial_x A(x)} = I - P_{\partial_x A(x)}$, $\partial^2_{\alpha\beta} A(x)$ is the m-vector of $n \times n$ Hessian matrices of the observation equations, and $g^{\alpha\beta}$ is Q_x written in index notation. Note that the components of the m-vector b_y in (4.1) are computed from the trace of the products of the Hessian matrices of the observation equations and the matrix $g^{\alpha\beta}(x)$.

It is interesting to observe that the biases in the least squares estimators $\hat{x}$ and $\hat{e}$ are computed from the m-vector b_y, just like in the linearized least squares case the estimators $\hat{x}$ and $\hat{e}$ themselves are computed from the m-vector y. One may interpret the m-vector b_y as describing what the bias in the m-vector y would have been if it were computed from x as $y = A(x)$, assuming now that x is a random vector with covariance matrix $\sigma^2 g^{\alpha\beta}$. Compare with (3.7).

Although (4.1) completely describes the bias in all the components of the parameter vector and residual vector respectively, it is useful to have *scalar* bias-measures available which summarize the bias present in the nonlinear model. To discern the significance of the biases we propose to weight the biases in the parameters and residuals with the inverses of $\sigma^2 Q_x$ and $\sigma^2 Q_y$ respectively. Our *scalar* bias-measures read therefore:

$$\begin{aligned} &\text{a)}\quad \sigma^{-2}\, E\{\hat{x}-x\}^*\, Q_x^{-1}\, E\{\hat{x}-x\} \doteq \sigma^{-2} \| P_{\partial_x A(x)} b_y \|^2 \\ &\text{b)}\quad \sigma^{-2}\, E\{\hat{e}\}^*\, Q_y^{-1}\, E\{\hat{e}\} \doteq \sigma^{-2} \| P^{\perp}_{\partial_x A(x)}\, b_y \|^2 \end{aligned} \tag{4.2}$$

Expressions (4.1) and (4.2) show that the bias in the parameters is zero if either the model is linear, i.e., $b_y = 0$, or if b_y is orthogonal to the range space of $\partial_x A(x)$. The bias in the residuals is zero if either the model is linear, b_y is orthogonal to the orthogonal complement of the range space of $\partial_x A(x)$ or if $m = n$, i.e., if there is no redundancy.

Note that the effect of b_y on the bias in $\hat{e}$ increases for increasing values of the so-called redundancy numbers, i.e., the diagonal elements of the projector $P^{\perp}_{\partial_x A}(x)$. This is comparable with the effect of the redundancy numbers on reliability in linear models [*Teunissen* 1985c, p. 541]. Drawing this parallel with reliability theory, we could call (4.2a) a scalar measure of *external bias* and (4.2b) a scalar measure of *internal bias*.

Due to orthogonality (see Figure 3) we have according to Pythagoras:

$$\| b_y \|^2 = \| P_{\partial_x A(x)} b_y \|^2 + \| P^{\perp}_{\partial_x A(x)} b_y \|^2 \,. \tag{4.3}$$

This shows that the scalar bias-measures of (4.2) are bounded from above by $\sigma^{-2} \| b_y \|^2$. Hence, to decide whether bias is small it may in some cases be sufficient to evaluate only the m-vector b_y.

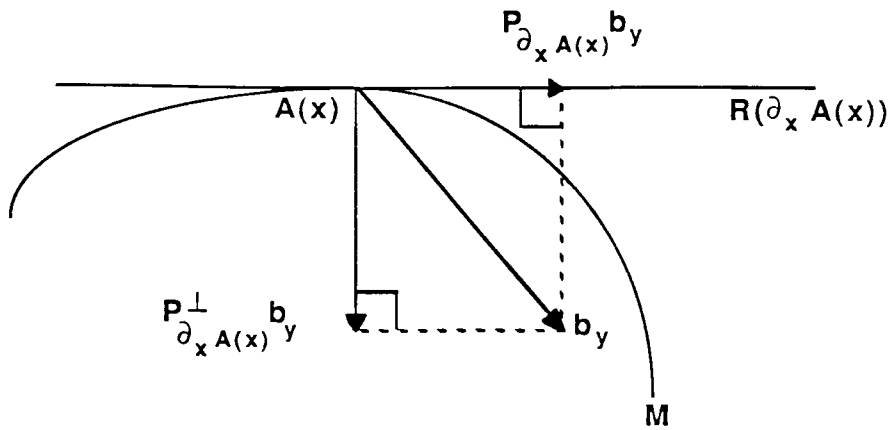

Figure 3: Orthogonal decomposition of b_y

Note that, if only a few parameters per observation equation are involved (this is typical of many geodetic situations), the Hessian matrix of the observation equation becomes sparse. This may simplify the computation of the components of b_y considerably.

Since the components of b_y of (4.1c) are computed from the traces of the products of the Hessian matrices of the observation equations with $g^{\alpha\beta}$ we can bound the absolute value of the i^{th} component of b_y from above as:

$$|b_y^i| \leq \frac{1}{2}\sigma^2 n\, \rho\left(\partial^2_{\alpha\beta}A^i\right)\rho\left(g^{\alpha\beta}\right), \tag{4.4}$$

where $\partial^2_{\alpha\beta}A^i$ is the Hessian matrix of the i^{th} observation equation and $\rho(.)$ stands for spectral radius. The spectral radius, $\rho(A)$, of a matrix A is defined to be $\rho(A) = \max|\lambda|$, where λ is an eigenvalue of A.

5. A Differential Geometric Interpretation

The differential geometry of nonlinear adjustment was introduced in *Krarup* [1982] and *Teunissen* [1984, 1985a, p. 118]. Here we will briefly comment on the differential geometric interpretation of (4.1). Using the Christoffel symbols of the second kind $\Gamma^{\alpha}_{\beta\gamma}$ and the mean curvature vector $\overline{N}$ of manifold M we can write (4.1) as

$$\begin{aligned} &\text{a)}\quad E\{\hat{x}^\alpha - x^\alpha\} \doteq -\frac{1}{2}\sigma^2 g^{\beta\gamma}\Gamma^{\alpha}_{\beta\gamma} \\ &\text{b)}\quad E\{\hat{e}\} = E\{y-\hat{y}\} \doteq -\frac{1}{2}\sigma^2 n\,\overline{N} \end{aligned} \tag{5.1}$$

where $\alpha, \beta, \gamma = 1, \ldots, n$.

It is obvious that the bias in $\hat{x}^\alpha$ depends on the parametrization. This is also reflected by the fact that the Christoffel symbols fail to be tensors. The bias can be made to vanish locally by choosing a suitable parametrization. The corresponding parameters are known in differential geometry as *geodetic polar coordinates*. Since the Christoffel symbols describe the turning and twisting of the coordinate lines in manifold M, equation (5.1a) also shows that the bias in $\hat{x}$ is caused by the nonlinear behavior of these coordinate lines.

As (5.1b) shows, the bias in the residual vector $\hat{e}$ is determined by the mean curvature of manifold M. It is therefore invariant for reparametrizations. The bias in $\hat{e}$ can vanish identically despite nonlinearity in the observation equations. If this is the case, one has a flat manifold M with nonlinear coordinate lines in it, or a manifold with vanishing mean curvature.

6. A Dual Formulation

As is well known, an adjustment problem can be formulated in terms of either observation equations or condition equations. So far we have used observation equations. We will now formulate the bias in $\hat{e}$ in terms of quantities which can be computed from the condition equations.

Model (2.1) in terms of condition equations reads

$$U^*(y) = t, \tag{6.1}$$

where t is an $(m\text{-}n)$-vector of misclosure variates and $U^*(.)$ is a map which maps R^m into R^{m-n}. The relation between the maps $A(.)$ and $U^*(.)$ is given by

$$U^*(A(x)) = 0 \quad \text{for all } x. \tag{6.2}$$

This relation enables us to re-express the bias in $\hat{e}$ in terms of quantities derivable from the condition equations. Our result reads

$$E\{\hat{e}\} = E\{y-\hat{y}\} \doteq Q_y\,\partial_y U(\tilde{y})\,Q_t^{-1} b_t, \text{ with}$$
$$b_t = \frac{1}{2}\sigma^2 \partial^2_{ij} U^*(\tilde{y})\, g^{ij}(\tilde{y}) \tag{6.3}$$

where $i,j = 1, \ldots, m$, $\tilde{y}$ stands for the expectation of y, $\partial_y U(\tilde{y})$ is the Jacobian of $U(.)$ evaluated at $\tilde{y}$, $Q_t = \partial_y U^*(\tilde{y})\, Q_y\, \partial_y U(\tilde{y})$, $\partial^2_{ij} U^*(\tilde{y})$ is the $(m\text{-}n)$-vector of mxm Hessian matrices of the condition equations and $g^{ij}(\tilde{y})$ is the matrix $P\partial_x A(x) Q_y = \partial_x A(x) Q_x \partial_x A(x)^*$ in index notation.

Note that the bias in $\hat{e}$ is computed from the $(m\text{-}n)$-vector b_t, just like in the linear(ized) least

squares case, the estimator $\hat{e}$ itself is computed from the vector of misclosures t. One may interpret the $(m\text{-}n)$-vector b_t, as describing what the bias in t would have been if it were computed from y as t= U^* (y), assuming now that y is a random vector with covariance matrix $\sigma^2 g^{ij}$.

7. Variance of Nonlinear Least Squares Estimators

The derivation of the covariance matrix of the least squares estimator $\hat{x}$ of x follows along the same lines as our derivation of the mean of $\hat{x}$. Although the result becomes rather unwieldy, some structure can be discerned by using differential geometric quantities. Our approximation to the covariance matrix of $\hat{x}$, neglecting terms of orders higher than σ^4 reads:

$$E\{(\hat{x}^\alpha - E\{\hat{x}^\alpha\})(\hat{x}^\beta - E\{\hat{x}^\beta\})\} \doteq \sigma^2 g^{\alpha\beta} +$$
$$+\frac{1}{2}\sigma^4\Gamma^\alpha_{\gamma\delta}\Gamma^\beta_{\varepsilon\rho g}\gamma\varepsilon g\,\delta\rho + \sigma^4 g^{\alpha\varepsilon}\left(P^\perp_{\partial_x A}\partial^2{}_{\varepsilon\gamma}A\right)^* Q_y^{-1} g_{\gamma\delta}\left(P^\perp_{\partial_x A}\partial^2{}_{\varepsilon\gamma}A\right) g^{\rho\beta}$$
$$+\sigma^4 g^{\alpha\gamma}\partial_\gamma\left(-\frac{1}{2}\Gamma^\beta_{\varepsilon\delta g}\varepsilon\delta\right) + \sigma^4 g^{\beta\gamma}\partial_\gamma\left(-\frac{1}{2}\Gamma^\alpha_{\varepsilon\delta g}\varepsilon\delta\right)$$

(7.1)

where $\alpha,\beta,\ldots,\varepsilon = 1,\ldots,n$.

Note that the first term on the right-hand side of (7.1) equals the approximation which is customarily used in linearized least squares. This approximation thus suffices if one may neglect terms of order σ^4 and higher. The second and last two terms on the right-hand side of (7.1) depend on the parametrization, whereas the third term depends on the curvature of manifold M. Since this third term depends on curvature there is a limit in the reduction that can be achieved through reparametrizations. The last two terms describe the change of bias in $\hat{x}$. Compare with (5.1a). The third term vanishes if the manifold is flat and the last four terms vanish if the observation-equations are linear.

8. On the Propagation of Bias

In many practical applications the purpose of the estimation process is not so much the computation of the estimator of x in model (2.1), but the computation of an estimator of a quantity derived from x, say $f = F(x)$, where $F(.)$ is a nonlinear function. In this case the computation of $\hat{x}$ constitutes an intermediate step in the estimation process.

Since $f = F(x)$, it is natural to compute the estimator of f as $\hat{f} = F(\hat{x})$ where $\hat{x}$ is the least squares estimator or x. To find the bias in the estimator $\hat{f}$ of f, we apply a Taylor expansion to $\hat{f} = F(\hat{x})$ at x, and take the expectation. This gives:

$$E\{\hat{f}-F(x)\} = \partial_x F(x)^* E\{\hat{x}-x\} + \frac{1}{2}\partial^2_{\alpha\beta}F(x)E\{(\hat{x}^\alpha-x^\alpha)(\hat{x}^\beta-x^\beta)\} + \ldots$$

(8.1)

If terms of order σ^4 and higher are neglected, substitution of (4.1a) and (7.1) into (8.1) gives for the bias in $\hat{f}$:

$$E\{\hat{f}-F(x)\} \doteq \partial_x F(x)^* Q_x \partial_x A(x)^* Q_y^{-1} b_y + \frac{1}{2}\sigma^2\partial^2_{\alpha\beta}F(x)g^{\alpha\beta}(x)$$

(8.2)

Note that the quadratic term of (8.1) must be included in the bias propagation, since both terms on the right-hand side of (8.1) are of the order σ^2. The second term on the right-hand side of (8.2) only vanishes if the function $F(.)$ is linear. The first term vanishes if x happens to be an unbiased estimator of x.

9. On the Bias in the 2D Symmetric Helmert Transformation

The nonlinear model of the 2D Symmetric Helmert transformation with a rotational invariant covariance structure was introduced in *Teunissen* [1985a, p. 141] as:

$$E\{\begin{pmatrix} u' \\ v' \end{pmatrix}\} = \begin{pmatrix} \lambda R(\theta)\otimes I_n & I_2\otimes e \\ I_{2n} & 0 \end{pmatrix}\begin{pmatrix} v \\ t \end{pmatrix};\ \begin{pmatrix} I_2\otimes Q_u & 0 \\ 0 & I_2\otimes Q_v \end{pmatrix},$$

(9.1)

where u' and v' are the random $2n$-vectors of 2D cartesian coordinates of the n network points in the two coordinate systems respectively; λ is scale; $R(\theta)$ is a 2x2 rotation matrix with rotation angle θ; v is the $2n$-vector of unknown coordinates in the second coordinate system; t is the two dimensional translation vector; $e = (1, \ldots, 1)^*$; Q_u is a scaled version of Q_v, i.e. $Q_u = s^2 Q_v$, where Q_v is an nxn positive definite matrix; and $\otimes$ denotes the Kronecker product.

With $u' = (\ldots, x1_i, y1_i, \ldots)^*$, $v' = (\ldots, x2_i, y2_i, \ldots)^*$ and $v = (\ldots, x_i, y_i, \ldots)^*$, the functional part of (9.1) can be written in components as:

$$
\begin{aligned}
E\{x1_i\} &= \lambda\cos\theta\, x_i + \lambda\sin\theta\, y_i ,\\
E\{y1_i\} &= -\lambda\sin\theta\, x_i + \lambda\cos\theta\, y_i ,\\
E\{x2_i\} &= x_i ,\\
E\{y2_i\} &= y_i , \qquad i = 1, \ldots, n.
\end{aligned}
$$

We have called the model the *Symmetric* Helmert transformation, since it assumes that *both* coordinate sets u' and v' are random. In the classical Helmert transformation only one of the two coordinate sets is assumed random. The remaining coordinate set is then assumed fixed, see e.g., [*Helmert* 1893; *Köchle* 1982; *Krarup* 1985; *Teunissen* 1985a]. As a consequence the classical 2D Helmert transformation constitutes a linear model, whereas the 2D Symmetric Helmert transformation constitutes a nonlinear model. That is, contrary to the classical 2D Helmert transformation, there does not exist a parametrization for the Symmetric Helmert transformation which results in linear observation equations.

As was shown in *Teunissen* [1985a, 1987a and b], it is possible to derive an *exact* nonlinear least squares solution of model (9.1). Hence no linearization of model (9.1) is needed and the computation of approximate values and iterations can be avoided.

Although the covariance structure of model (9.1) is of a simplified form, it is felt that this structure is still sufficiently general for many practical applications. When digitizing maps, the covariance matrix of the digitized coordinates can often even be simplified to a scaled unit matrix. Also in case of geodetic networks the assumption of a rotational invariant covariance structure often suffices. The rotational invariant Baarda-Alberda substitute matrix [*Teunissen*, 1985c] for instance, describes the precision of many geodetic networks to a sufficient degree.

Since the exact nonlinear least squares solution of model (9.1) is known, the biases in the estimators can be derived through either the use of (3.7), (4.1) or (5.1). As was mentioned in section four, the vector b_y of (4.1) already gives valuable information on the bias of the estimators. For model (9.1) this vector reads

$$
b_y = \frac{1}{2}\lambda^{-1}\sigma^2_{\hat\lambda}\begin{pmatrix} R(\theta)\otimes I_n \\ 0 \end{pmatrix} v , \tag{9.2}
$$

where

$$
\bar v = \left(I_2 \otimes P^{\perp}_e\right) v , \quad P^{\perp}_e = I_n - P_e , \quad P_e = e\left(e^* Q_u^{-1} e\right)^{-1} e^* Q_u^{-1}
$$

and the variance of $\hat\lambda$ is given by

$$
\sigma^2_{\hat\lambda} = \frac{1 + \lambda^2 s^{-2}}{\bar v^*\left(I_2 \otimes Q_u^{-1}\right)\bar v} . \tag{9.3}
$$

Equation (9.2) shows that the last $2n$ components of b_y are zero. This is due to the absence of nonlinearity in the corresponding observation equations. The first $2n$ components of b_y can, however, still be considerable. In particular, the components of the vector v in (9.2) can be arbitrarily large, depending on the location of the origin of the coordinate system. Thus it cannot be decided on the basis of (9.2) whether bias will be small. We therefore apply (4.1) to our model (9.1) to obtain the biases in the parameters and the residuals of the estimated coordinates. The result is surprisingly, that all biases, except the one for scale, equal zero. The explanation reads as follows. The bias in the translation is zero because t occurs linearly in our model. The bias in the rotation angle vanishes, since parametrization with θ reduces the Christoffel symbols to zero. Finally, the fact that the bias in the least squares residuals of the estimated coordinates is zero, can be explained as follows. One can show [*Teunissen* 1985a, p. 168] that the manifold described by model (9.1) has only two non-zero curvatures per normal direction. Moreover, the curvatures only differ in their signs. Hence the mean curvature of the manifold is zero. Therefore, according to (5.1b), the bias in the residual vector must vanish.

The bias in the least squares estimator $\hat\lambda$ of scale is given by the simple formula

$$
E\left\{\hat\lambda - \lambda\right\} \doteq \frac{1}{2}\sigma_{\hat\lambda}\left(\frac{\sigma_{\hat\lambda}}{\lambda}\right) \tag{9.4}
$$

This shows that the mean of the test statistic $(\hat\lambda - 1) / \sigma_{\hat\lambda}$ which is customarily used for testing whether $\lambda = 1$, differs from zero by $\frac{1}{2}\sigma_{\hat\lambda}$.

To get some indications of how the bias in scale depends on the number of network points and network precision, we assume as in [*Teunissen* 1986, p. 221], that the points in the second coordinate system are distributed over a square grid of spacing d. We also assume that $Q_u = \sigma_u^2 I_n$ and $Q_v = \sigma_v^2 I_n$. With these assumptions, equation (9.4) yields

$$E\{\hat{\lambda}-\lambda\} \doteq \frac{3\left(\left(\frac{\sigma_u}{d}\right)^2 + \lambda^2\left(\frac{\sigma_v}{d}\right)^2\right)}{\lambda n(n-1)}, \qquad (9.5)$$

where n equals the number of network points.

This shows that for most practical applications the bias in scale can be neglected. For $(\sigma_u/d) = (\sigma_v/d) = 10^{-5}$, $\lambda = 1$ and $n = 4$ we have namely $E\{\hat{\lambda}-\lambda\} = 1/2 \cdot 10^{-10}$. Again note that it is primarily the high precision of the measurements which make bias negligible.

10. Conclusions

In this paper we have discussed the effect of nonlinearity on the probabilistic properties of least squares estimators. Some exact methods of deriving the probability distribution of nonlinear estimators were given. These methods are, however, difficult to apply in practice. Hence, the best one can do in most cases is to obtain approximations using suitable Taylor expansions. We have shown how to derive these approximations for general nonlinear estimators and for least squares estimators. In the case of least squares estimators, we had to overcome the difficulty of not knowing the exact relation between the estimator and the observables.

Two different expressions for the bias in the parameter vector and residual vector respectively, were derived. The second set of expressions gives a differential geometric interpretation and shows how curvature of the manifold and nonlinearity of the coordinate lines effect bias.

We also presented a dual expression for the bias in the residual vector based on condition equations and showed how bias in the parameter vector propagates into the bias of estimators derived from the parameters.

An approximation of the second moment of the parameters was given to the order of σ^4 and expressed in terms of differential geometric quantities. The first term of this expression, being of the order σ^2, corresponds to the approximation customarily made.

Finally, we applied our results to the 2D Symmetric Helmert transformation. To an approximation of the order σ^4 all biases, except the one for scale, turn out to be zero. A geometric explanation was given. Although the bias in scale does not vanish, it is of the order σ^2 and can, therefore, for most practical cases be considered insignificant.

Acknowledgement

The financial assistance for the first author was provided by the Netherlands organization for scientific research (N.W.O.).

References

Baarda, W. 1967. Statistical concepts in geodesy, *Netherlands Geodetic Commission, Publications on Geodesy, New Series*, Vol. 2, No. 4, Delft.

Bähr, H.G. 1985. Second order effects in the Gauss-Helmert model, *Proceedings of the 7th International Symposium on Geodetic Computations*, Cracow, Poland.

Blaha, G. 1987. Nonlinear parametric least squares adjustment, *Nova University Oceanographic Center, Scientific Report No. 1*, March.

Helmert, F.R. 1893. *Die Europäische Längengradmessung in 52 Grad Breite von Greenwich bis Warschau*, I. Heft. Veröff. des Königl. Preussischen Geodätischen Instituts und Centralbureaus der Int. Erdmessung. Berlin, pp. 47-50.

Jeudy, L.M.A. 1987. Generalized variance-covariance propagation law formulae and application to explicit least squares adjustments, *Proceedings of the XIX General Assembly of the International Union of Geodesy and Geophysics* in Vancouver, 9 - 22 August.

Kelley, R.P. and W.A. Thompson jr. 1978. Some results on nonlinear and constrained least squares, *Manuscripta Geodaetica*, Vol. 3, pp. 299-320.

Köchle, T. 1982. *Die Räumliche Helmertransformation in Algebraischer Darstellung*, Vermessung, Photogrammetrie, Kulturtechnik, Vol. 9, pp. 292-297.

Krarup, T. 1982. Nonlinear adjustment and curvature, *Forty Years of Thought*, Delft, pp. 145-159.

Krarup, T. 1985. *Contribution to the Geometry of the Helmert Transformation*, Geodetic Institute, Copenhagen, Denmark.

Kubik, K.K. 1967. Iterative Methoden zur Lösung des nichtlinearen Ausgleichsproblemes (Iterative methods for the solution of the nonlinear adjustment problem) *Zeitschrift für Vermessungswesen* 91, pp. 214-225.

Kubik, K.K. 1968. On the efficiency of least squares estimators in nonlinear models, *Statistica Neerlandica* 22, 1, pp. 33-36.

Pope, A. 1972. Some pitfalls to be avoided in the iterative adjustment of nonlinear problems, *Proceedings of the 38th Annual Meeting of the American Society of Photogrammetry.*

Pope, A. 1974. Two approaches to nonlinear least squares adjustments, *The Canadian Surveyor*, Vol. 28, No. 5, pp. 663-669.

Saito, T. 1973. The nonlinear least squares of condition equations, *Bulletin Geodesique*, Vol. 110, pp. 367-395.

Schaffrin, B. 1973. A note on linear prediction within a Gauss-Markoff model linearized with respect to a random approximation, *Proceedings of First Tampere Seminan of Linear Models*, Department of Mathematics, University of Tampere, Finland.

Schek, H.J. and P. Maier 1976. Nichtlineare Normalgleichungen zur Bestimmung der Unbekannten and deren Kovarianzmatrix, *Zeitschrift für Vermessungswesen*, Vol. 101, pp. 140-159.

Stark, E. and E. Mikhail 1973. Least squares and nonlinear functions, *Photogrammetric Engineering*, pp. 405-412.

Teunissen, P.J.G. 1984. A note on the use of Gauss Formula in nonlinear geodetic adjustments, *Statistics & Decisions, Symposium* No. 2, pp. 455-466, R. Oldenburg Verlag, München.

Teunissen, P.J.G. 1985a. The geometry of geodetic inverse linear mapping and nonlinear adjustment, *Netherlands Geodetic Commission, Publications on Geodesy, New Series*, Vol. 8, No. 1, Delft, 177 pp.

Teunissen, P.J.G. 1985b. Nonlinear adjustment: An introductory discussion and some new results, *Proceedings of SSG 4.56 Workshop Meeting*, Ghania, Greece, pp. 10-12.

Teunissen, P.J.G. 1985c. Quality control in geodetic networks, *Optimization and design of Geodetic Networks*, Springer, Berlin, E.W. Grafarend and F. Sanso editors.

Teunissen, P.J.G. 1986. Adjusting and testing with models of the affine and similarity transformation, *Manuscripta Geodaetica* Vol. 11, pp. 214-225.

Teunissen, P.J.G.1987a. The 1D and 2D symmetric Helmert transformation: Exact nonlinear least squares solutions, *Reports of the Department of Geodesy, Section Mathematical and Physical Geodesy*, No. 87.1, Delft, 25 pp.

Teunissen, P.J.G. 1987b. The nonlinear 2D symmetric Helmert transformation: An exact least squares solution, Delft, accepted for *Bulletin Geodesique*, May.

Wolf, H. 1961. Das Fehlerfortpflanzungsgesetz mit Gliedern II. Ordnung, *Zeitschrift für Vermessungswesen* 86, pp. 86-88.

MS rec'd 88/02/15

Authors

Dr.Ir. P.J.G. Teunissen is employed as a geodetic project researcher by the Netherlands Organization for the Advancement of Pure Research (Z.W.O.). He is presently a guest of the Department of Surveying Engineering of The University of Calgary. He graduated cum laude in 1980 at the Delft University of Technology and obtained his Ph.D. in 1985, with distinction, from the same university. He received the Steven Hoogendyk Prize in 1986 from the Batavian Society of Experimental Philosophy and the Brigadier Guy Bomford Prize in 1987 from the International Association of Geodesy (IAG). He is president of IAG's Special Study Group on Nonlinear Adjustment.

Dr.-Ing. E.H. Knickmeyer is a Post-Doctorate Fellow at the Department of Surveying Engineering of The University of Calgary. He graduated at the Technical University Berlin, Federal Republic of Germany, in 1978 and obtained the degree Dr.-Ing. from the Stuttgart University, F.R.G., in 1984. He is a State Survey and Land Consolidation Assessor.❑