Numerical Aspects of Combination at the Observation Equation and Normal Equation Level

Michael Gerstl

German Geodetic Research Institute, Munich, Germany

1 Introduction

During combination work for an TRF, the DGFI obtained normal equations which caused numerical difficulties. We presume that these problems have their origin in reduced parameters (clock offsets in the case in point) which were given an insufficient approximate value of zero – most likely, as the observation equations are linear with respect to these parameters. The author of the normal equations was not aware of this problem, as he himself worked with the unreduced observation equations.

These circumstances drew our attention to the numerical differences between normal and observation equations, and particularly to the equilibrium of the parameter corrections.

2 Notation

Denote measurement and modelling by $m+\delta m = f(p)+\delta f(p)$, where $m, \delta m \in \mathbf{IR}^m$ are the vectors of measurement results and measurement errors, $f, \delta f \in \mathbf{IR}^m$ are the model function and model error, and $p \in \mathbf{IR}^n$ is the vector of the unknown parameters. A linearization with approximate value $p^{(0)}$ introduces a linearization error $\delta l(p, p^{(0)})$ of the order of $(p-p^{(0)})^2$.

All three types of errors are combined in the a priori error vector $e(p^{(0)}) = \delta f(p) - \delta m + \delta l(p, p^{(0)})$. The residuum "observed − computed" is loosely called the observation vector $b = [f(p) - f(p^{(0)})] + [\delta f(p) - \delta m]$. These elements comprise a system of observation equations

$$b - e(p^{(0)}) = f(p) - f(p^{(0)}) - \delta l(p, p^{(0)}) = \frac{\partial f}{\partial p}(p^{(0)})(p-p^{(0)}) = Ax \qquad (1)$$

The notation is summarized thus: the **linearized observation equations** or error equations denoted by Ax = b–e, the **normal equations** denoted by Nx = y with $N = A^T PA \in \mathbf{IR}^{n\times n}$ and $y = A^T Pb \in \mathbf{IR}^n$. Given any solution $\hat{x}$, we determine

- the estimated parameters: $p^{(1)} = p^{(0)} + \hat{x}$
- the a posteriori error vector:
  $e(p^{(1)})$ or $e(\hat{x}) = b - A\hat{x} = f(p) - f(p^{(1)}) + \delta f(p) - \delta m + \delta l(p^{(1)}, p^{(0)})$
- the weighted sum of squares of the residuals or, mathematically speaking, the **error norm**:
  $\| e(p^{(1)}) \|_P^2 = e(p^{(1)})^T P e(p^{(1)})$.

3 Sensitivity of the Least Squares Problem

Let $\tilde{x}$ and $\tilde{e}(\tilde{x})$ be the solution of the perturbed least squares problem

$$\| \tilde{e}(x) \|_P = \| \tilde{b} - \tilde{A}x \|_P = \min$$

with the components $\tilde{A} = A + \delta A$ and $\tilde{b} = b + \delta b$. How do the perturbations δA and δb influence the solution of the linear least squares problem?

Under the precondition of rank conservation, $rg(A + \delta A) = rg(A)$, it holds (Golub and Wilkinson, 1969):

$$\| \tilde{x} - x \|_2 \le \frac{cond(A)}{1 - \varepsilon} \cdot (\alpha \| x \|_2 + \beta) + \left(\frac{cond(A)}{1 - \varepsilon} \right)^2 \cdot \alpha \frac{\| e(x) \|_P}{\| A \|} \qquad (2)$$

$$\| \tilde{e} - e \|_P \le \| \delta A \| \cdot \| x \|_2 + \| \delta b \|_P + \frac{cond(A)}{1 - \varepsilon} \cdot \alpha \| e(x) \|_P \qquad (3)$$

$$\text{with } \alpha := \frac{\| \delta A \|}{\| A \|}, \quad \beta := \frac{\| \delta b \|}{\| A \|}, \quad \varepsilon := \frac{lub(\delta A)}{glb(A)} < 1. \qquad (4)$$

For $\| e(x) \|_P > 0$, it is the term with $cond(A)^2 \| e(x) \|_P$ in equation (2) which is the most important. In addition to the condition number of the weighted design matrix it depends on the consistency of the linearized observation equations.

It is well known that the solution of a least squares problem by means of normal equations is not as stable as a direct inversion of the design matrix by orthogonalization methods. Moreover, the calculation of the error norm ($\| e(\hat{x}) \|_P^2$) from the norm of the observation vector ($b^T P b$ or "$l^T P l$") is a technical error. These statements do not hold in an unlimited way!

The relation between the norms of the observation vector and the a posteriori error vector can be written as

$$\| e(p^{(1)}) \|_P^2 = e(p^{(1)})^T P e(p^{(1)}) = b^T P b - y^T \hat{x}.$$

This subtraction produces numerical cancellation of mantissa digits! If observation equations are available, the error norm can be calculated by inserting the solution $\hat{x}$ into the equations and summing up – a time-consuming but sure procedure. If normal equations only are the basis for calculation, then the latter formula is unavoidable. Such is the case for the ITRF combination, where only normal equations or solutions, which allow the reconstruction of the normal equations only, are available. Thus, we want to know the conditions for the mantissa being numerically cancelled in the latter difference. For this reason, we investigate the effect of poor approximate values on the solution and its error bounds.

4 Effects of Unbalanced Approximate Values

Let a subset of the unknown parameters $\{ p_j : j \in J \}$ with index $J \subset_{\neq} \{1, 2, \ldots, n\}$, and a significant influence on the system of equations, have approximate values substantially worse than the remaining parameters. It will be shown that, for this case, the formulas for

- the calculation of the error norm from the norm of observation vector,

- a transformation to suitable approximate values, and
- the reduction of the parameters with poor approximate values

are subject to numerical cancelling of the mantissa. The presentation method used consists in comparing a well-balanced least squares problem with a perturbed problem, which results from the good one in such a way that the approximate values of the parameters $\{p_j : j \in J\}$ are tampered with a relatively large value Δp_j:

$$\tilde{p}^{(0)} = p^{(0)} - \Delta p \qquad \text{with} \qquad \left\{ \begin{array}{ll} |\Delta p_i| > \max_k |x_k| & \text{for } i \in J \\ \Delta p_i = 0 & \text{for } i \notin J \end{array} \right\}$$
$$\tilde{x} = x + \Delta p$$

The unperturbed system of equations, $Ax = b - e$, is assumed to be consistent, balanced in x and b, and near-linear with respect to the parameters $\{p_j : j \in J\}$. For ease of notation, these equations, for which the parameters $\{p_j : j \in J\}$ provide a numerically not vanishing contribution, are placed at the beginning of the system.

$$\begin{array}{ll} k = 1, 2, \ldots, m_1 \le m: & \left| \sum_{j \in J} A_{kj} \Delta p_j \right| \ge \varepsilon > 0 \\ k = m_1 + 1, \ldots, m: & \left| \sum_{j \in J} A_{kj} \Delta p_j \right| \approx 0 \end{array} \tag{5}$$

The first m_1 equations are written as $A_1 x = b_1 - e_1$, and the remaining $m_2 = m - m_1$ equations as $A_2 x = b_2 - e_2$. In addition, we assume b_1 and b_2 to be uncorrelated.

$$A = \begin{pmatrix} A_1 \\ A_2 \end{pmatrix} \begin{matrix} \} m_1 \\ \} m_2 \end{matrix}, \qquad b = \begin{pmatrix} b_1 \\ b_2 \end{pmatrix}, \qquad P = \begin{pmatrix} P_1 & 0 \\ 0 & P_2 \end{pmatrix} \tag{6}$$

Now, the terms of the perturbed system shall be represented by those of the unperturbed system of equations. For indices $k = 1, \ldots, m_1$ it holds that

$$\tilde{b}_k = b_k + \sum_{i \in J} A_{ki} \Delta p_i = \left(1 + \frac{b_k}{\sum_{\nu \in J} A_{k\nu} \Delta p_\nu} \right) \sum_{i=1}^{n} A_{ki} \Delta p_{i,}$$

and for the remaining indices, that $\tilde{b}_k = b_k$. Let us introduce δ_k $(k = 1, \ldots, m_1)$ by

$$\delta_k := b_k \Big/ \left(\sum_{\nu=1}^{n} A_{k\nu} \Delta p_\nu \right) \quad \text{and} \quad \Delta := \operatorname{diag}\left(\delta_1, \ldots, \delta_{m_1} \right)$$

The perturbation must have a significant impact on the system of equations. For that, it is not sufficient to require $|\tilde{x}_j| \approx |p_j| \gg \max_k |x_k|$ but

$$|b_k| = \left| \sum_{i=1}^{n} A_{ki} x_i + e_k \right| \ll \left| \sum_{i=1}^{n} A_{ki} p_i \right| \qquad (k = 1, 2, \ldots, m_1), \tag{7}$$

where "$\ll$" means a difference in size of several powers of ten. As to the normal equations mentioned in the introduction, a clock offset of 10 microseconds corresponds to a range correction of 3 km, while the station coordi-

nate corrections lay in the range of a few centimetres. Direct consequence of (7) is $|\delta_k| \ll 1$.

Now we have it. The terms of the perturbed observation and normal equations in a decomposition conformable to (6) are written as

$$\begin{aligned}
\tilde{b} &= \begin{pmatrix} (I+\Delta)A_1\Delta p \\ b_2 \end{pmatrix} \\
\tilde{y} &= A^T P\tilde{b} = A_1^T P_1 (I+\Delta)A_1\Delta p + A_2^T P_2 b_2 \\
\tilde{b}^T P\tilde{b} &= \Delta p^T A_1^T (I+\Delta)P_1(I+\Delta)A_1\Delta p + b_2^T P_2 b_2 \tilde{A} = A = \begin{pmatrix} A_1 \\ A_2 \end{pmatrix} \\
\tilde{N} &= N \\
\tilde{e}(\tilde{x}) &= e(x)
\end{aligned} \tag{8}$$

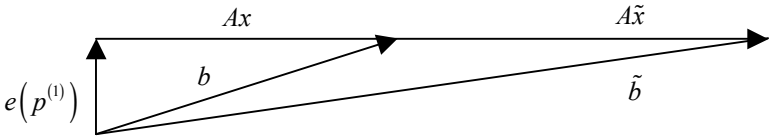
4.1 Calculation of the error norm (with normal equations)

We compare the perturbed with the unperturbed problem. From $\tilde{e}(\tilde{x}) = e(x)$ we get

$$\tilde{y}^T\tilde{x} - y^T x = \tilde{b}^T P\tilde{b} - b^T Pb = (\Delta p)^T A_1^T (I+2\Delta) P_1 A_1 \Delta p.$$

Thereby we have shown that the perturbation increases the left as the right term in $b^T Pb - y^T x$ by the same large amount which cancels on performing the difference.

Poor approximate values are the only cause of cancellation. This could be geometrically recognized by



One is tempted to improve the system of equations by transforming it to better approximate values.

4.2 Transformation to suitable approximate values

Let us write the formulas of the transformation of variables $\tilde{x}_j \mapsto x_j$ ($j \in J$), using the representation (8):

$$\begin{aligned}
b &= \tilde{b} - A\Delta p = \begin{pmatrix} (I+\Delta)A_1\Delta p \\ b_2 \end{pmatrix} - \begin{pmatrix} A_1\Delta p \\ 0 \end{pmatrix} \\
y &= \tilde{y} - N\Delta p = A_1^T P_1 (I+\Delta)A_1\Delta p - A_1^T P_1 A_1 \Delta p + A_2^T P_2 b_2 \\
b^T Pb &= \tilde{b}^T P\tilde{b} - [\,2(A\Delta p)^T P\tilde{b} - (A\Delta p)^T P(A\Delta p)] = \\
&= \tilde{b}^T P\tilde{b} - (\Delta p)^T A_1^T P_1 (I+2\Delta)A_1\Delta p = \\
&= (\Delta p)^T A_1^T (I+\Delta)P_1(I+\Delta)A_1\Delta p - (\Delta p)^T A_1^T P_1 (I+2\Delta)A_1\Delta p + b_2^T P_2 b_2
\end{aligned}$$

Since the components of $A_1\Delta p$ are large compared to those of b (see (7)), the right-hand sides of both normal and observation equations likewise suffer from numerical cancellation. For the norm $b^T Pb$, the cancelled digits are twice as many.

4.3 Reduction of parameters with poor approximate values

The variables to be reduced, $\{ x_j : j \in J\}$, are placed at the beginning of the vector:

$$x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \quad with \begin{cases} x_1 \in \mathrm{IR}^{n_1} : & \text{variables to be reduced,} \\ x_2 \in \mathrm{IR}^{n_2} : & \text{variables to be preserved.} \end{cases}$$

The corresponding partitioning of columns of the design matrix leads to a refinement of the partition (6) :

$$A = \begin{pmatrix} A_{11} & A_{12} \\ \underbrace{0}_{n_1} & \underbrace{A_{22}}_{n_2} \end{pmatrix} \begin{matrix} \} & m_1 \\ \} & m_2 \end{matrix}, \quad \Delta p = \begin{pmatrix} \Delta p_1 \\ 0 \end{pmatrix} \begin{matrix} \} n_1 \\ \} n_2 \end{matrix}$$

The submatrix A_{21} vanishes, because parameters from $\{ p_j : j \in J \}$ are per definition not included in these observation equations. For x_1 being reducible, the existence of the pseudoinverse $A_{11}^{+} = \left(A_{11}^T P_1 A_{11} \right)^{-1} A_{11}^T P_1$ is a necessary condition. This representation is to be applied on the perturbed systems of equations. Using (8), we get

$$\begin{pmatrix} \tilde{b}_1 \\ \tilde{b}_2 \end{pmatrix} = \begin{pmatrix} (I+\Delta) A_{11} \Delta p_1 \\ b_2 \end{pmatrix}, \begin{pmatrix} \tilde{y}_1 \\ \tilde{y}_2 \end{pmatrix} = \begin{pmatrix} A_{11}^T P_1 (I+\Delta) A_{11} \Delta p_1 \\ A_{12}^T P_1 (I+\Delta) A_{11} \Delta p_1 + A_{22}^T P_2 b_2 \end{pmatrix}$$

and $\tilde{b}^T P \tilde{b} = (\Delta p_1)^T A_{11}^T (I+\Delta) P_1 (I+\Delta) A_{11} \Delta p_1 + b_2^T P_2 b_2$.

Finally, the equations to reduce parameters p_j with dominant corrections x_j are written to show the numerical cancellation feature,:

$$\tilde{b}^{red} = \begin{pmatrix} (I+\Delta) A_{11} \Delta p_1 - \left(I + A_{11} A_{11}^{+} \Delta \right) A_{11} \Delta p_1 \\ b_2 \end{pmatrix}$$

$$\begin{aligned} \tilde{y}^{red} &= A_{12}^T P_1 (I+\Delta) A_{11} \Delta p_1 + A_{22}^T P_2 b_2 - \\ &\quad - A_{12}^T P_1 \left(I + A_{11} A_{11}^{+} \Delta \right) A_{11} \Delta p_1 \\ \left(\tilde{b}^T P \tilde{b} \right)^{red} &= (\Delta p_1)^T A_{11}^T (I+\Delta) P_1 (I+\Delta) A_{11} \Delta p_1 + b_2^T P_2 b_2 - \\ &\quad - (\Delta p_1)^T A_{11}^T (I+\Delta) P_1 \left(I + A_{11} A_{11}^{+} \Delta \right) A_{11} \Delta p_1 \end{aligned}$$

5 Conclusions

Let us come to a decision between normal equations and observation equations.

At the moment, we can live with normal equations as long as the design matrix A and the observation vector b are sufficiently balanced. The balance depends on the equilibrium of the approximate values. Seriously unbalanced normal equations cannot be repaired afterwards; they must be rebuild from the beginning.

There are still some problems left which might call in question the usability of normal equations. They concern the behaviour of reduced variables and the variance component estimation and the robust estimation methods.