



Advanced Estimation Methods and Analysis

ENGO 629

Y. Gao

Department of Geomatics Engineering

CONTENTS

	PAGE
1. INTRODUCTION.....	1
1.1 Concept of Estimation	1
1.2 Major Issues in Estimation	2
2. LEAST SQUARES ESTIMATION	9
2.1 The Least Squares Problem.....	9
2.2 The Least Squares Normal Equations	12
2.3 Derivation of Variance-Covariance Matrices	14
2.4 Generation of the Standard Cases.....	20
3. KALMAN FILTERING	27
3.1 Random process and Properties	27
3.2 Gauss-Markov Process	32
3.3 Kinematic Modeling and Transition Matrix.....	34
3.4 A Geomatics Example of Kinematic Modeling.....	38
3.5 Kalman Filtering.....	42
3.6 Implementation Aspects of Kalman Filtering	49
4. ROBUST ESTIMATION	57
4.1 Concept of Robust Statistics.....	55

4.2	Influence Function.....	57
4.3	Robust Criteria Based on the Influence Function	59
4.4	Min-Max M-Estimator	62
4.5	Robust Least Squares	68
4.6	Robust Kalman Filtering	70
5.	COLLOCATION	81
5.1	Collocation Mathematical Model.....	79
5.2	Least Squares Collocation	81
5.3	Least Squares Collocation Equations	83
5.4	Stepwise Collocation.....	88
6.	STATISTICAL TESTING AND ANALYSIS	93
6.1	Statistical Testing in Least Squares.....	91
6.2	Statistical Testing in Kalman Filtering	95
6.2	Reliability Analysis	99
6.3	Robust Statistical Testing	102

1

INTRODUCTION

1.1 Concept of Estimation

Estimation is defined as the drawing inferences regarding parameters of probability distributions on the basis of sample statistics. More specifically estimation is the obtaining a (unique) set of values for a set of unknown parameters and their properties from a redundant (non-deterministic) set of observations.

Two other definitions related to estimation are also introduced here. They are estimator and estimate and an understanding of their difference is necessary when reading literature on estimation. An estimator is defined as the sample statistic used for estimating the corresponding parameters or an algorithm that processes measurements to deduce a minimum error estimate of the parameters of a system by utilizing knowledge of the system and measurement dynamics, assumed statistics of system noises and measurement errors, and initial condition information. So estimator defines the method of estimation. An estimate is defined as the computed value of the parameters. So an estimate is the actual value obtained from estimation.

In order to estimate the unknown parameters from given observations, the functional relationship between the unknown parameters and the observed qualities must be established. Let x represent the unknown parameter vector ($m \times 1$) and l the observation vector ($n \times 1$), the following function can be given as a general expression for a non-linear system:

$$f(x, l) = 0, \quad (1.1)$$

where f denotes the vector of individual functions. If the observation vector can be explicitly expressed by a function of x , equation 1.1 reduces to

$$l = f(x), \quad (1.2)$$

The non-linear system in equation 1.2 can be linearized into a linear algebraic system as follows:

$$\mathbf{Ax} = \mathbf{b} \quad (1.3a)$$

or

$$\sum_{j=1}^m a_{ij}x_j = b_i \quad 1 \leq i \leq n \quad (1.3b)$$

where \mathbf{A} is $n \times m$ matrix, \mathbf{x} is a vector of $m \times 1$, \mathbf{b} is a vector of $n \times 1$. Estimation with system equation 1.3 is to find a solution to \mathbf{x} from the observation \mathbf{l} .

Many solutions are possible dependent on the algebraic properties of the linear system and they are outlined in Table 1. In the table, $r = \text{rank}(\mathbf{A}) = r(\mathbf{A})$.

1.2 Major Issues In Estimation

There are three major issues that must be dealt with in estimation, namely







- 1) Mathematical modeling
- 2) Estimation principle selection
- 3) Analysis of estimation results

The above three aspects will be discussed in the following as they are the three major tasks in optimal estimation and analysis.

1.2.1 Mathematical Modeling

Given in equation 1.3 is only the functional relationship between the unknown parameters and the observed qualities for a linear system. In reality, most functional relationships are not fully satisfied without errors due to the imperfections of observing instrumentation and procedures resulting in errors in the observed quantities. To describe this aspect, a stochastic model, typically represented by a variance and co-variance matrix \mathbf{C} , should be introduced to describe the randomness or un-deterministic component of the observed qualities in addition to the function model given in equation 1.1. As a result, the mathematical model related to estimation has to be expanded to the form of

Table 1.1: Summary of Solutions of $Ax = b$

	Dimension and rank of A n. m. r	Type of A	Solutions of $Ax = 0$	Existence of Solution(s)	Number of Solutions
1	$n = m = r$		$x = 0$ ($r = m$)	$r([A \ b]) = r(A)$ ($r = n$) Always	1
2	$r = n < m$		$x = 0 \ \& \ x \neq 0$ ($r < m$)	$r([A \ b]) = r(A)$ ($r = n$) Always	∞
3	$n > m = r$		$x = 0$ ($r = m$)	$r([A \ b]) > r(A)$ ($r < n$) $r([A \ b]) = r(A)$	0 1
4	$r < n = m$		$x = 0 \ \& \ x \neq 0$ ($r < m$)	$r([A \ b]) > r(A)$ ($r < n$) $r([A \ b]) = r(A)$	0 ∞
5	$n > m > r$		$x = 0 \ \& \ x \neq 0$ ($r < m$)	$r([A \ b]) > r(A)$ ($r < n$) $r([A \ b]) = r(A)$	0 ∞
6	$r < n < m$		$x = 0 \ \& \ x \neq 0$ ($r < m$)	$r([A \ b]) > r(A)$ ($r < n$) $r([A \ b]) = r(A)$	0 ∞

$$Ax = b + \epsilon \quad C \quad (1.4)$$

where ϵ is the observation error vector with the following typically properties: $E(\epsilon) = 0$ (mean value of the observation errors) and $E(\epsilon\epsilon^T) = C$ (variance-covariance matrix of the observations). In summary, a functional model describes the mathematical relationship between the measurements such as distance and angle measurements and the unknown parameters such as position and attitude parameters. A functional model $Ax = b$ can be considered as the deterministic component in mathematical modeling. The stochastic model $E(\epsilon) = 0$ and $E(\epsilon\epsilon^T) =$

C describe the uncertainty or random component of a functional model. From statistical point of view, it describes the statistical properties of the underlying functional model, namely the measurements, the unknowns as well as the model equation itself. The stochastic model is necessary because neither measurements nor the functional model are considered perfect.

Time is an important parameter in real-time parameter estimation. Given this, the functional model in equation 1.1 and 1.2 can be extended into the following forms, respectively

$$f(x, l, t) = 0 \quad (1.4)$$

$$l = f(x, t) \quad (1.5)$$

where t is a time tag. If let t_c represent the current time, the concept of smoothing ($t < t_c$), filtering ($t = t_c$) and prediction ($t > t_c$) are induced related to modern filtering estimation theory

1.2.2 Estimation Principle

We have noticed two major issues from Table 1.1:

- a) The linear system may not be consistent, namely $\text{rank}(\mathbf{A}) \neq \text{rank}([\mathbf{A} \ \mathbf{b}])$. In this case there exists no solution unless the system can be converted into a consistent system.
- b) The linear system may have infinite number of solutions. In this case there exists no unique solution unless constraints are applied.

When with inconsistent and unlimited number of possible solutions, converting the system into a consistent one and finding a unique “best” solution among all possible ones is the major objective of so-called optimal estimation. For that purpose, the estimation principle to define the optimality must be developed with respect to different applications.

There are different estimation principles used in optimal estimation where different “Best” could be defined. Given below are just some examples where x is the unknown parameter, \hat{x} the estimate of x , ℓ the observation, $f(\cdot)$ the probability density function.

- a) $\text{Prob}(|\hat{x} - x| > c) = \min!$

where c is an arbitrary small positive value.

Best: minimum property for $|\hat{x} - x| > c$

$$b) \sum_{i=1}^n |\hat{x} - x_i| = \min!$$

Best: minimum sum of absolute estimate errors

$$c) \sum_{i=1}^n (\hat{x} - x_i)^2 = \min!$$

Best: minimum sum of square estimate errors

Although different estimation principles lead to different estimators, some common “good” properties on an estimator are highly expected regardless of the applications. They are given below:

a) Consistence

$$\lim_{n \rightarrow \infty} \text{Prob}(|\hat{x} - x| < \varepsilon) = 1$$

where ε is a small value and the estimate converges to the true value.

b) Unbiasedness

$$E(\hat{x}) = x$$

where the expected value of the estimate is the same as the true value. For small sample size, it usually uses $E(\hat{x}) = E(x)$ as the condition.

c) Minimum Mean Variance

$$E\{(\hat{x} - E(\hat{x}))^T (\hat{x} - E(\hat{x}))\} = \min!$$

where the estimate error variance (mean squares error) is less or equal to that of any others.

d) Efficiency

Unbiasedness and minimum variance

Different estimation principles would result in different estimation methods and some typical estimation methods are given in the following. A detailed description of each method can be found in literature.

a) Maximum Posteriori Estimation (MP)

$$f(\mathbf{x} \mid \ell) = \max!$$

or

$$\ln f(\mathbf{x} \mid \ell) = \max!$$

b) Maximum Likelihood Estimation (ML)

$$f(\ell \mid \mathbf{x}) = \max!$$

or

$$\ln f(\ell \mid \mathbf{x}) = \max!$$

c) Minimum Variance Estimation Method (MV)

$$\text{Trace } E\{(\mathbf{x} - \hat{\mathbf{x}}(\ell))(\mathbf{x} - \hat{\mathbf{x}}(\ell))^T\} = \min!$$

d) Least Squares Estimation (LS)

$$[\ell - \ell(\hat{\mathbf{x}})]^T P [\ell - \ell(\hat{\mathbf{x}})] = \min!$$

Note that the least squares method does not need probability distribution information in principle. But it must have such information if we want to construct statistical properties about the estimate.

All above estimators are unbiased and efficient when they are linear systems and distributed according to Gaussian distribution.

There are many other estimation methods available such as robust estimators including ridge estimation, minimum sum of absolute residual and M-estimator.

1.2.3 Analysis of Estimation Results

Analysis of estimation results is necessary in applications for the assurance of the quality or correctness of the estimate. For that purpose, the mathematical models including the functional and stochastic ones and the assumptions associated with them must be verified or tested for any unexpected significant discrepancy before they can be actually used. This is important because big discrepancies such as blunders in observations could distort the estimate significantly.

Subsequently, the use of wrong estimation solutions is dangerous to applications especially for life-critical ones.

There are many different types of analysis that could be carried out to assess the mathematical models and the obtained estimation results. Major analysis includes

- a) Mathematical model analysis
- b) Bias identification and analysis
- c) Distribution analysis
- d) Blunder detection and identification
- e) Variance-covariance analysis
- f) Residual analysis

2

LEAST SQUARES ESTIMATION

In this chapter, least squares equations are derived using a general scheme from which all the cases can be deduced. The methodologies used in the derivation are useful for the derivation of other estimation methods such as Kalman filtering (Chapter 3) and collocation (Chapter 5).

2.1 The Least Squares Problem

Least squares estimation is the standard method to obtain a unique set of values for a set of unknown parameters (\mathbf{x}) from a redundant set of observables (\mathbf{l}) through a known **mathematical model** ($\mathbf{f}(\mathbf{x}, \mathbf{l})$).

Before we treat the general situation, let us describe the **least squares problem** for the linear-explicit case, that is

$$\mathbf{l} = \mathbf{f}(\mathbf{x}) , \quad (2.1)$$

$$\begin{matrix} \mathbf{l} & = & \mathbf{A} & \mathbf{x} , \\ n \times 1 & & n \times u & u \times 1 \end{matrix} \quad (2.2)$$

where n observables are related to u unknown parameters through a **design matrix** \mathbf{A} . We know the observables have some unknown correction (residuals). Denoting these by \mathbf{r} , and the observed value of the observables by \mathbf{l} , equation 2.2 becomes

$$\begin{matrix} \mathbf{l} & + & \mathbf{r} & = & \mathbf{A} & \mathbf{x} . \\ n \times 1 & & n \times 1 & & n \times u & u \times 1 \end{matrix} \quad (2.3)$$

The least squares estimate for \mathbf{x} is obtained under the condition that the quadratic form

* from Chapter 2 of the 629 Lecture Notes "The method of least squares: a synthesis of advances" by EJ Krakiwsky.

$$\hat{\mathbf{r}}^T \mathbf{P} \hat{\mathbf{r}} = \text{minimum} \quad (2.4)$$

where the weight matrix

$$\mathbf{P} = \sigma_o^2 \mathbf{C}_l^{-1} \quad (2.5)$$

is related to the a priori variance factor (σ_o^2) and variance-covariance matrix of the observations (\mathbf{C}_l). It is through the help of this condition that two equations, in addition to equation 2.3, are obtained, thereby yielding least squares estimates for \mathbf{x} and \mathbf{r} - denoted by $\hat{\mathbf{x}}$ and $\hat{\mathbf{r}}$, respectively.

We now consider the general situation where our mathematical model of m equations is implicit and non-linear, that is

$$\mathbf{f}(\hat{\mathbf{x}}, \hat{\mathbf{l}}) = 0, \quad (2.6)$$

that the accuracy estimates exist for the observations \mathbf{C}_l , and the parameters are treated as quasi-observables \mathbf{C}_x . The corresponding weight matrices are

$$\mathbf{P}_l = \sigma_o^2 \mathbf{C}_l^{-1} \quad (2.7)$$

and

$$\mathbf{P}_x = \sigma_o^2 \mathbf{C}_x^{-1} \quad (2.8)$$

The least squares estimates $\hat{\mathbf{x}}$ and $\hat{\mathbf{r}}$ for this general situation are obtained under the condition that

$$[\hat{\mathbf{r}}^T \mathbf{C}_l^{-1} \hat{\mathbf{r}} + \hat{\boldsymbol{\delta}}^T \mathbf{C}_x^{-1} \hat{\boldsymbol{\delta}}] = \text{minimum}, \quad (2.9)$$

where $\hat{\boldsymbol{\delta}}$ are corrections to the parameters as explained immediately below.

We chose to work with linear sets of equations. Thus we approximate our mathematical model (equation 2.6) by a linear Taylor series as follows:

$$\begin{aligned} \mathbf{f}(\hat{\mathbf{x}}, \hat{\mathbf{l}}) &= \mathbf{f}(\mathbf{x}, \mathbf{l}) + \frac{\partial \mathbf{f}}{\partial \mathbf{x}} \Big|_{\mathbf{x}, \mathbf{l}} (\hat{\mathbf{x}} - \mathbf{x}) + \frac{\partial \mathbf{f}}{\partial \mathbf{l}} \Big|_{\mathbf{x}, \mathbf{l}} (\hat{\mathbf{l}} - \mathbf{l}) \\ &= \mathbf{f}(\mathbf{x}, \mathbf{l}) + \frac{\partial \mathbf{f}}{\partial \mathbf{x}} \Big|_{\mathbf{x}, \mathbf{l}} \hat{\boldsymbol{\delta}} + \frac{\partial \mathbf{f}}{\partial \mathbf{l}} \Big|_{\mathbf{x}, \mathbf{l}} \hat{\mathbf{r}} = 0. \end{aligned} \quad (2.10)$$

This **misclosure vector**

$$\underset{nx1}{\mathbf{w}} = \underset{nx1}{\mathbf{f}}(\underset{mx1}{\mathbf{x}}, \underset{ux1}{\mathbf{l}}) \quad (2.11)$$

is the mathematical model evaluated with the quasi-observed values of the parameters (\mathbf{x}) and the observed values of the observables (\mathbf{l}). When \mathbf{f} is evaluated with some approximate values of the parameters (\mathbf{x}^0), we denote the misclosure vector as

$$\mathbf{w}^0 = \mathbf{f}(\mathbf{x}^0, \mathbf{l}) . \quad (2.11a)$$

The **first design matrix** is

$$\mathbf{A} = \left. \frac{\partial \mathbf{f}}{\partial \mathbf{x}} \right|_{\mathbf{x}, \mathbf{l}} . \quad (2.12)$$

and the **second design matrix** is

$$\mathbf{B} = \left. \frac{\partial \mathbf{f}}{\partial \mathbf{l}} \right|_{\mathbf{x}, \mathbf{l}} . \quad (2.13)$$

The final value (adjusted) of the vector of parameters is

$$\hat{\mathbf{x}} = \mathbf{x} + \hat{\boldsymbol{\delta}} . \quad (2.14)$$

The final value (adjusted) of the vector of observables is

$$\hat{\mathbf{l}} = \mathbf{l} + \hat{\mathbf{r}} . \quad (2.15)$$

The linearized mathematical model in symbolic form is

$$\underset{mxu}{\mathbf{A}} \underset{ux1}{\hat{\boldsymbol{\delta}}} + \underset{mxn}{\mathbf{B}} \underset{nx1}{\hat{\mathbf{r}}} + \underset{mx1}{\mathbf{w}} = \mathbf{0} , \quad (2.16)$$

where $\hat{\boldsymbol{\delta}}$ and $\hat{\mathbf{r}}$ are least squares estimates making $\hat{\mathbf{x}}$ and $\hat{\mathbf{l}}$ also least squares estimates.

2.2 The Least Squares Normal Equations

The least squares normal equations relating the unknown quantities $\hat{\delta}$ and \hat{r} to the known quantities A, B, w, C_l^{-1} , and C_x^{-1} is obtained from the variation function

$$\phi = \hat{r}^T C_l^{-1} \hat{r} + \hat{\delta}^T C_x^{-1} \hat{\delta} + 2 \hat{k}^T (A \hat{\delta} + B \hat{r} + w) , \quad (2.17)$$

where the newly introduced unknown quantity \hat{k} is the vector of m Lagrange correlates, where m = number of Equations. To find the minimum of the two quadratic forms subject to the constraint function (linearized math model) is known as the **extremal problem** with constraints. The Lagrange method is the standard method of solving this problem.

First the derivatives of the variation function with respect to the variates \hat{r} and $\hat{\delta}$ are taken and set equal to zero to determine the extremum, minimum in the case, namely

$$\frac{1}{2} \frac{\partial \phi}{\partial \hat{r}} = \hat{r}^T C_l^{-1} + \hat{k}^T B = 0 \quad (2.18)$$

$$\frac{1}{2} \frac{\partial \phi}{\partial \hat{\delta}} = \hat{\delta}^T C_x^{-1} + \hat{k}^T A = 0 \quad (2.19)$$

The transpose of the above two equations and the linearized mathematical model constitute the three equations of the **least squares normal equation system**:

$$\begin{array}{l} C_l^{-1} \hat{r} + B^T \hat{k} = 0 , \\ \text{nxn} \quad \text{nx1} \quad \text{nxm} \quad \text{mx1} \\ \\ C_x^{-1} \hat{\delta} + A^T \hat{k} = 0 , \\ \text{uxu} \quad \text{ux1} \quad \text{uxm} \quad \text{mx1} \\ \\ A \hat{\delta} + B \hat{r} + w = 0 . \\ \text{mxu} \quad \text{ux1} \quad \text{mxn} \quad \text{nx1} \quad \text{mx1} \end{array} \quad (2.20)$$

The most expanded form of the least squares normal equation system in block matrix form is

$$\begin{pmatrix} \mathbf{C}_1^{-1} & \mathbf{B}^T & 0 \\ \mathbf{B} & 0 & \mathbf{A} \\ 0 & \mathbf{A}^T & \mathbf{C}_x^{-1} \end{pmatrix} \begin{pmatrix} \hat{\mathbf{r}} \\ \hat{\mathbf{k}} \\ \hat{\boldsymbol{\delta}} \end{pmatrix} + \begin{pmatrix} 0 \\ \mathbf{w} \\ 0 \end{pmatrix} = 0, \quad (2.21)$$

with a coefficient matrix of dimensions $n + m + u$. A solution for the vector comprising $\hat{\mathbf{r}}$, $\hat{\mathbf{k}}$, and $\hat{\boldsymbol{\delta}}$ is possible by directly inverting the coefficient matrix. This is not efficient, thus a normal equation system is derived where the inversions are smaller.

We use a special elimination technique, e.g. Thomson (1969) and Wells and Krakiwsky (1971). Given a matrix equation system

$$\begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{pmatrix} \begin{pmatrix} \mathbf{x} \\ \mathbf{y} \end{pmatrix} + \begin{pmatrix} \mathbf{u} \\ \mathbf{v} \end{pmatrix} = 0 \quad (2.22)$$

where $\mathbf{A}, \mathbf{B}, \mathbf{C}$ and \mathbf{D} constitute the known coefficient matrix, \mathbf{x} and \mathbf{y} the unknown vector, and \mathbf{u} and \mathbf{v} the known vector, \mathbf{x} is eliminated by forming a modified coefficient matrix and known vector as follows:

$$[\mathbf{D} - \mathbf{C}\mathbf{A}^{-1}\mathbf{B}] \mathbf{y} + [\mathbf{v} - \mathbf{C}\mathbf{A}^{-1}\mathbf{u}] = 0 \quad (2.23)$$

(\mathbf{A} must be non-singular). The proof is simple and left to the reader as an exercise.

We return to the problem at hand and first eliminate $\hat{\mathbf{r}}$ from equation 2.21, where \mathbf{D} becomes the lower two by two hyper-matrix of equation 2.21.

$$\left(\begin{pmatrix} 0 & \mathbf{A} \\ \mathbf{A}^T & \mathbf{C}_x^{-1} \end{pmatrix} - \begin{pmatrix} \mathbf{B} \\ 0 \end{pmatrix} \mathbf{C}_1 [\mathbf{B}^T \ 0] \right) \begin{pmatrix} \hat{\mathbf{k}} \\ \hat{\boldsymbol{\delta}} \end{pmatrix} + \left(\begin{pmatrix} \mathbf{w} \\ 0 \end{pmatrix} - \begin{pmatrix} \mathbf{B} \\ 0 \end{pmatrix} \mathbf{C}_1 [0] \right) = 0, \quad (2.24)$$

$$\begin{pmatrix} -\mathbf{B}\mathbf{C}_1\mathbf{B}^T & \mathbf{A} \\ \mathbf{A}^T & \mathbf{C}_x^{-1} \end{pmatrix} \begin{pmatrix} \hat{\mathbf{k}} \\ \hat{\boldsymbol{\delta}} \end{pmatrix} + \begin{pmatrix} \mathbf{w} \\ 0 \end{pmatrix} = 0. \quad (2.25)$$

Then eliminate $\hat{\mathbf{k}}$ using the same technique:

$$[\mathbf{C}_x^{-1} - \mathbf{A}^T (-\mathbf{B}\mathbf{C}_1\mathbf{B}^T)^{-1} \mathbf{A}] \hat{\boldsymbol{\delta}} + 0 - \mathbf{A}^T (-\mathbf{B}\mathbf{C}_1\mathbf{B}^T)^{-1} \mathbf{w} = 0, \quad (2.26)$$

$$[\mathbf{A}^T (\mathbf{B}\mathbf{C}_1\mathbf{B}^T)^{-1} \mathbf{A} + \mathbf{C}_x^{-1}] \hat{\boldsymbol{\delta}} + \mathbf{A}^T (\mathbf{B}\mathbf{C}_1\mathbf{B}^T)^{-1} \mathbf{w} = 0, \quad (2.27)$$

and

$$\hat{\delta} = -[A^T (BC_1 B^T)^{-1} A + C_x^{-1}]^{-1} A^T (BC_1 B^T)^{-1} w \quad (2.28)$$

The solution for $\hat{\delta}$ is made during the first expression from equation system 2.25, namely

$$\begin{aligned} -(BC_1 B^T) \hat{k} + A \hat{\delta} + w &= 0, \\ \hat{k} &= (BC_1 B^T)^{-1} (A \hat{\delta} + w). \end{aligned} \quad (2.29)$$

The solution for \hat{k} is made using the first expression from equation system 2.21, namely,

$$\begin{aligned} C_1^{-1} \hat{r} + B^T \hat{k} &= 0, \\ \hat{r} &= -C_1 B^T \hat{k}. \end{aligned} \quad (2.30)$$

Equations 2.28, 2.29 and 2.30 represent the alternative solution to the least squares normal equation 2.21.

2.3 Derivation of Variance-Covariance Matrices

In this section we derive the variance-covariance matrices for the residual vector \hat{r} , the parameter vector $\hat{\delta}$, the final value of the parameter vector \hat{x} , and the final value of the observable vector \hat{l} .

We make extensive use of the **covariance law** which states that, given a functional relationship

$$y = f(z) \quad (2.31)$$

between two random vectors y and z along with the variance-covariance matrix of z , (C_z) , the variance-covariance matrix of y is given by

$$C_y = \left(\frac{\partial f}{\partial z} \right) C_z \left(\frac{\partial f}{\partial z} \right)^T. \quad (2.32)$$

Note the following relationship between the variance-covariance C and weight P :

$$\mathbf{P} = \sigma_0^2 \mathbf{C}^{-1}.$$

Variance-Covariance Matrix for $\hat{\mathbf{x}}$

We follow Krakiwsky (1968), Kouba (1970) and Wells and Krakiwsky (1971) in deriving the variance-covariance matrix for the final value of the parameters $\hat{\mathbf{x}}$.

According to equations 2.14 and 2.28

$$\hat{\mathbf{x}} = \mathbf{x} + \hat{\boldsymbol{\delta}}, \quad (2.33)$$

$$\hat{\mathbf{x}} = \mathbf{x} - [\mathbf{A}^T \mathbf{M}^{-1} \mathbf{A} + \mathbf{C}_x^{-1}]^{-1} \mathbf{A}^T \mathbf{M}^{-1} \mathbf{w}, \quad (2.34)$$

where we have let

$$\mathbf{M} = \mathbf{B} \mathbf{C}_l \mathbf{B}^T. \quad (2.35)$$

Recalling equation 2.11,

$$\mathbf{w} = \mathbf{f}(\mathbf{x}, \mathbf{l}) \quad (2.36)$$

we see that $\hat{\mathbf{x}}$ is a function of two independent random variables - the a priori estimate of the parameters (*quasi-observable) (\mathbf{x}), and the observed value of the observables (\mathbf{l}).

Applying the covariance law to equation 2.33 yields (assuming henceforth that \mathbf{x} and \mathbf{l} are statistically independent)

$$\mathbf{C}_{\hat{\mathbf{x}}} = \left(\frac{\partial \hat{\mathbf{x}}}{\partial \mathbf{x}} \right) \mathbf{C}_x \left(\frac{\partial \hat{\mathbf{x}}}{\partial \mathbf{x}} \right)^T + \left(\frac{\partial \hat{\mathbf{x}}}{\partial \mathbf{l}} \right) \mathbf{C}_l \left(\frac{\partial \hat{\mathbf{x}}}{\partial \mathbf{l}} \right)^T. \quad (2.37)$$

From 2.34,

$$\left(\frac{\partial \hat{\mathbf{x}}}{\partial \mathbf{x}} \right) = \mathbf{I} - [\mathbf{A}^T \mathbf{M}^{-1} \mathbf{A} + \mathbf{C}_x^{-1}]^{-1} \mathbf{A}^T \mathbf{M}^{-1} \mathbf{A} \quad (2.38)$$

since from equation 2.12

$$\left(\frac{\partial \mathbf{w}}{\partial \mathbf{x}} \right) = \frac{\partial \mathbf{f}(\mathbf{x}, \mathbf{l})}{\partial \mathbf{x}} = \mathbf{A}. \quad (2.39)$$

Also from 2.34,

$$\left(\frac{\partial \hat{\mathbf{x}}}{\partial \mathbf{l}} \right) = - [\mathbf{A}^T \mathbf{M}^{-1} \mathbf{A} + \mathbf{C}_x^{-1}]^{-1} \mathbf{A}^T \mathbf{M}^{-1} \mathbf{B} \quad (2.40)$$

since from equation 2.13

$$\left(\frac{\partial \mathbf{w}}{\partial \mathbf{l}} \right) = \frac{\partial \mathbf{f}(\mathbf{x}, \mathbf{l})}{\partial \mathbf{l}} = \mathbf{B} . \quad (2.41)$$

Before proceeding further with the derivations, it will prove useful to derive the covariance matrix for $\hat{\delta}$ (equation 2.28), and to do this, we first need the weight coefficient matrix of \mathbf{w} .

Applying the covariance law to equation 2.36 yields

$$\mathbf{C}_w = \left(\frac{\partial \mathbf{f}}{\partial \mathbf{x}} \right) \mathbf{C}_x \left(\frac{\partial \mathbf{f}}{\partial \mathbf{x}} \right)^T + \left(\frac{\partial \mathbf{f}}{\partial \mathbf{l}} \right) \mathbf{C}_l \left(\frac{\partial \mathbf{f}}{\partial \mathbf{l}} \right)^T . \quad (2.42)$$

and hence

$$\mathbf{C}_w = \mathbf{A} \mathbf{C}_x \mathbf{A}^T + \mathbf{B} \mathbf{C}_l \mathbf{B}^T . \quad (2.43)$$

From equation 2.34

$$\hat{\delta} = - [\mathbf{A}^T \mathbf{M}^{-1} \mathbf{A} + \mathbf{C}_x^{-1}]^{-1} \mathbf{A}^T \mathbf{M}^{-1} \mathbf{w} . \quad (2.44)$$

Applying the covariance law to the above equation and taking into account equation 2.43

$$\mathbf{C}_{\hat{\delta}} = \left(\frac{\partial \hat{\delta}}{\partial \mathbf{w}} \right) \mathbf{C}_w \left(\frac{\partial \hat{\delta}}{\partial \mathbf{w}} \right)^T \quad (2.45)$$

$$= (- [\mathbf{A}^T \mathbf{M}^{-1} \mathbf{A} + \mathbf{C}_x^{-1}]^{-1} \mathbf{A}^T \mathbf{M}^{-1}) (\mathbf{A} \mathbf{C}_x \mathbf{A}^T + \mathbf{B} \mathbf{C}_l \mathbf{B}^T) \\ (- [\mathbf{A}^T \mathbf{M}^{-1} \mathbf{A} + \mathbf{C}_x^{-1}]^{-1} \mathbf{A}^T \mathbf{M}^{-1})^T \quad (2.46)$$

$$\mathbf{C}_{\hat{\delta}} = [\mathbf{A}^T \mathbf{M}^{-1} \mathbf{A} + \mathbf{C}_x^{-1}]^{-1} \mathbf{A}^T \mathbf{M}^{-1} (\mathbf{A} \mathbf{C}_x \mathbf{A}^T + \mathbf{M}) \mathbf{M}^{-1} \mathbf{A} \\ [\mathbf{A}^T \mathbf{M}^{-1} \mathbf{A} + \mathbf{C}_x^{-1}]^{-1} \quad (2.47)$$

since $\mathbf{B} \mathbf{C}_l \mathbf{B}^T = \mathbf{M}$ and \mathbf{M} is symmetric, and letting

$$\mathbf{N} = \mathbf{A}^T \mathbf{M}^{-1} \mathbf{A} , \quad (2.48)$$

align horizontally

$$\mathbf{C}_{\hat{\delta}} = [\mathbf{N} + \mathbf{C}_x^{-1}]^{-1} \mathbf{A}^T \mathbf{M}^{-1} (\mathbf{A} \mathbf{C}_x \mathbf{A}^T + \mathbf{M}) \mathbf{M}^{-1} \mathbf{A} [\mathbf{N} + \mathbf{C}_x^{-1}]^{-1} \quad (2.49)$$

and expanding terms

$$\mathbf{C}_{\hat{\delta}} = [\mathbf{N} + \mathbf{C}_x^{-1}]^{-1} \mathbf{N} \mathbf{C}_x \mathbf{N} [\mathbf{N} + \mathbf{C}_x^{-1}]^{-1} + [\mathbf{N} + \mathbf{C}_x^{-1}]^{-1} \mathbf{N} [\mathbf{N} + \mathbf{C}_x^{-1}]^{-1} . \quad (2.50)$$

We now turn to the task of determining $\mathbf{C}_{\hat{x}}$. Substituting equations 2.38 and 2.40 into 2.37 and using 2.48, we find

$$\begin{aligned} \mathbf{C}_{\hat{x}} &= (\mathbf{I} - [\mathbf{N} + \mathbf{C}_x^{-1}]^{-1} \mathbf{N}) \mathbf{C}_x (\mathbf{I} - [\mathbf{N} + \mathbf{C}_x^{-1}]^{-1} \mathbf{N})^T \\ &\quad + (-[\mathbf{N} + \mathbf{C}_x^{-1}]^{-1} \mathbf{A}^T \mathbf{M}^{-1} \mathbf{B}) \mathbf{C}_1 (-[\mathbf{N} + \mathbf{C}_x^{-1}]^{-1} \mathbf{A}^T \mathbf{M}^{-1} \mathbf{B})^T \\ &= (\mathbf{I} - [\mathbf{N} + \mathbf{C}_x^{-1}]^{-1} \mathbf{N}) \mathbf{C}_x (\mathbf{I} - \mathbf{N} [\mathbf{N} + \mathbf{C}_x^{-1}]^{-1}) \\ &\quad + [\mathbf{N} + \mathbf{C}_x^{-1}]^{-1} \mathbf{A}^T \mathbf{M}^{-1} \mathbf{B} \mathbf{C}_1 \mathbf{B}^T \mathbf{M}^{-1} \mathbf{A} [\mathbf{N} + \mathbf{C}_x^{-1}]^{-1} \\ &= \mathbf{C}_x - \mathbf{C}_x \mathbf{N} [\mathbf{N} + \mathbf{C}_x^{-1}]^{-1} - [\mathbf{N} + \mathbf{C}_x^{-1}]^{-1} \mathbf{N} \mathbf{C}_x \\ &\quad + [\mathbf{N} + \mathbf{C}_x^{-1}]^{-1} \mathbf{N} \mathbf{C}_x \mathbf{N} [\mathbf{N} + \mathbf{C}_x^{-1}]^{-1} + [\mathbf{N} + \mathbf{C}_x^{-1}]^{-1} \mathbf{N} [\mathbf{N} + \mathbf{C}_x^{-1}]^{-1} . \end{aligned} \quad (2.51)$$

Noting that the last two terms of the above equations are identical to $\mathbf{C}_{\hat{\delta}}$ (equation 2.50), we can write

$$\mathbf{C}_{\hat{x}} = \mathbf{C}_x + \mathbf{C}_{\hat{\delta}} - \mathbf{C}_x \mathbf{N} [\mathbf{N} + \mathbf{C}_x^{-1}]^{-1} - [\mathbf{N} + \mathbf{C}_x^{-1}]^{-1} \mathbf{N} \mathbf{C}_x . \quad (2.52)$$

The above expression can be shown to be equivalent to the inverse of the coefficient matrix of the normal equation system 2.28 for $\hat{\delta}$, that is

$$\begin{aligned} \mathbf{C}_{\hat{x}} &= [\mathbf{A}^T (\mathbf{B} \mathbf{C}_1 \mathbf{B}^T)^{-1} \mathbf{A} + \mathbf{C}_x^{-1}]^{-1} , \\ &= [\mathbf{A}^T \mathbf{M}^{-1} \mathbf{A} + \mathbf{C}_x^{-1}]^{-1} , \end{aligned} \quad (2.53)$$

$$\mathbf{C}_{\hat{x}} = [\mathbf{N} + \mathbf{C}_x^{-1}]^{-1} . \quad (2.54)$$

To prove this we begin by factoring-out the term

$$[\mathbf{N} + \mathbf{C}_x^{-1}]^{-1}$$

from equation 2.51, then multiply terms, and finally cancel like terms. The details are the following:

$$\begin{aligned}
 \mathbf{C}_{\hat{x}} &= [\mathbf{N} + \mathbf{C}_x^{-1}]^{-1} ([\mathbf{N} + \mathbf{C}_x^{-1}] \mathbf{C}_x - [\mathbf{N} + \mathbf{C}_x^{-1}] \mathbf{C}_x \mathbf{N} [\mathbf{N} + \mathbf{C}_x^{-1}]^{-1} - \mathbf{N} \mathbf{C}_x \\
 &\quad + \mathbf{N} \mathbf{C}_x \mathbf{N} [\mathbf{N} + \mathbf{C}_x^{-1}]^{-1} + \mathbf{N} [\mathbf{N} + \mathbf{C}_x^{-1}]^{-1}) \\
 &= [\mathbf{N} + \mathbf{C}_x^{-1}]^{-1} [\mathbf{N} \mathbf{C}_x + \mathbf{I} - \mathbf{N} \mathbf{C}_x \mathbf{N} [\mathbf{N} + \mathbf{C}_x^{-1}]^{-1} - \mathbf{N} [\mathbf{N} + \mathbf{C}_x^{-1}]^{-1} \\
 &\quad - \mathbf{N} \mathbf{C}_x + \mathbf{N} \mathbf{C}_x \mathbf{N} [\mathbf{N} + \mathbf{C}_x^{-1}]^{-1} + \mathbf{N} [\mathbf{N} + \mathbf{C}_x^{-1}]^{-1}] \\
 &= [\mathbf{N} + \mathbf{C}_x^{-1}]^{-1} \mathbf{I} = [\mathbf{N} + \mathbf{C}_x^{-1}]^{-1} .
 \end{aligned}$$

The estimated variance factor is

$$\hat{\sigma}_o^2 = \frac{\hat{\mathbf{r}}^T \mathbf{C}_1^{-1} \hat{\mathbf{r}} + \hat{\boldsymbol{\delta}}^T \mathbf{C}_x^{-1} \hat{\boldsymbol{\delta}}}{v} . \quad (2.55)$$

The degrees of freedom*

$$v = m - u + m_x , \quad (2.56)$$

where m is the number of equations in \mathbf{f} , u the number of parameters to be estimated, and m_x the number of parameters weighted. The proof of equation 2.55 is beyond the scope of this work, see for example, Hamilton (1964), Wells and Krakiwsky (1971).

Variance-Covariance Matrix for $\hat{\mathbf{l}}$

We begin from the definition of the final (adjusted) observables (equation 2.15),

$$\hat{\mathbf{l}} = \mathbf{l} + \hat{\mathbf{r}} . \quad (2.57)$$

Using equations 2.29 and 2.30

$$\hat{\mathbf{l}} = \mathbf{l} - \mathbf{C}_1 \mathbf{B}^T \hat{\mathbf{k}}$$

* Equation 2.56 is only an approximation, see Bossler (1972) for a complete and rigorous treatment.

$$\begin{aligned}
&= \mathbf{I} - \mathbf{C}_l \mathbf{B}^T \mathbf{M}^{-1} (\mathbf{A} \hat{\boldsymbol{\delta}} + \mathbf{w}) \\
&= \mathbf{I} - \mathbf{C}_l \mathbf{B}^T \mathbf{M}^{-1} \mathbf{A} \hat{\boldsymbol{\delta}} - \mathbf{C}_l \mathbf{B}^T \mathbf{M}^{-1} \mathbf{w}, \quad (2.58)
\end{aligned}$$

and after using equation 2.44

$$\hat{\mathbf{I}} = \mathbf{I} + \mathbf{C}_l \mathbf{B}^T \mathbf{M}^{-1} \mathbf{A} [\mathbf{A}^T \mathbf{M}^{-1} \mathbf{A} + \mathbf{C}_x^{-1}]^{-1} \mathbf{A}^T \mathbf{M}^{-1} \mathbf{w} - \mathbf{C}_l \mathbf{B}^T \mathbf{M}^{-1} \mathbf{w}. \quad (2.59)$$

Applying the covariance law to the above equation yields

$$\mathbf{C}_{\hat{\mathbf{I}}} = \left(\frac{\partial \hat{\mathbf{I}}}{\partial \mathbf{l}} \right) \mathbf{C}_l \left(\frac{\partial \hat{\mathbf{I}}}{\partial \mathbf{l}} \right)^T + \left(\frac{\partial \hat{\mathbf{I}}}{\partial \mathbf{x}} \right) \mathbf{C}_x \left(\frac{\partial \hat{\mathbf{I}}}{\partial \mathbf{x}} \right)^T. \quad (2.60)$$

where

$$\left(\frac{\partial \hat{\mathbf{I}}}{\partial \mathbf{l}} \right) = \mathbf{I} + \mathbf{C}_l \mathbf{B}^T \mathbf{M}^{-1} \mathbf{A} [\mathbf{A}^T \mathbf{M}^{-1} \mathbf{A} + \mathbf{C}_x^{-1}]^{-1} \mathbf{A}^T \mathbf{M}^{-1} \mathbf{B} - \mathbf{C}_l \mathbf{B}^T \mathbf{M}^{-1} \mathbf{B}, \quad (2.61)$$

$$\left(\frac{\partial \hat{\mathbf{I}}}{\partial \mathbf{x}} \right) = \mathbf{C}_l \mathbf{B}^T \mathbf{M}^{-1} \mathbf{A} [\mathbf{A}^T \mathbf{M}^{-1} \mathbf{A} + \mathbf{C}_x^{-1}]^{-1} \mathbf{A}^T \mathbf{M}^{-1} \mathbf{A} - \mathbf{C}_l \mathbf{B}^T \mathbf{M}^{-1} \mathbf{A}. \quad (2.62)$$

Substituting equations 2.61 and 2.62 into 2.60, noting equation 2.50, and collecting terms we get

$$\begin{aligned}
\mathbf{C}_{\hat{\mathbf{I}}} &= \mathbf{C}_l + \mathbf{C}_l \mathbf{B}^T \mathbf{M}^{-1} \mathbf{A} \mathbf{C}_{\hat{\boldsymbol{\delta}}} \mathbf{A}^T \mathbf{M}^{-1} \mathbf{B} \mathbf{C}_l \\
&\quad + \mathbf{C}_l \mathbf{B}^T \mathbf{M}^{-1} \mathbf{A} \mathbf{C}_x \mathbf{A}^T \mathbf{M}^{-1} \mathbf{B} \mathbf{C}_l \\
&\quad - \mathbf{C}_l \mathbf{B}^T \mathbf{M}^{-1} \mathbf{A} (\mathbf{A}^T \mathbf{M}^{-1} \mathbf{A} + \mathbf{C}_x^{-1})^{-1} \mathbf{A}^T \mathbf{M}^{-1} \mathbf{A} \mathbf{C}_x \mathbf{A}^T \mathbf{M}^{-1} \mathbf{B} \mathbf{C}_l \\
&\quad - \mathbf{C}_l \mathbf{B}^T \mathbf{M}^{-1} \mathbf{A} \mathbf{C}_x \mathbf{A}^T \mathbf{M}^{-1} \mathbf{A} (\mathbf{A}^T \mathbf{M}^{-1} \mathbf{A} + \mathbf{C}_x^{-1})^{-1} \mathbf{A}^T \mathbf{M}^{-1} \mathbf{B} \mathbf{C}_l \\
&\quad - \mathbf{C}_l \mathbf{B}^T \mathbf{M}^{-1} \mathbf{B} \mathbf{C}_l. \quad (2.63)
\end{aligned}$$

Using equation 2.52, the above equation can be written in terms of $\mathbf{C}_{\hat{\mathbf{x}}}$, namely

$$\mathbf{C}_{\hat{\mathbf{I}}} = \mathbf{C}_l + \mathbf{C}_l \mathbf{B}^T \mathbf{M}^{-1} \mathbf{A} \mathbf{C}_{\hat{\mathbf{x}}} \mathbf{A}^T \mathbf{M}^{-1} \mathbf{B} \mathbf{C}_l - \mathbf{C}_l \mathbf{B}^T \mathbf{M}^{-1} \mathbf{B} \mathbf{C}_l. \quad (2.64)$$

Weight and Coefficient Matrix for $\hat{\mathbf{r}}$

Expressing $\hat{\mathbf{r}}$ as a function of \mathbf{w} (analogous to equation 2.34), and applying the covariance law, we get after considering equation 2.64 that

$$\begin{aligned} \mathbf{C}_{\hat{\mathbf{r}}} &= \mathbf{C}_l \mathbf{B}^T \mathbf{M}^{-1} \mathbf{B} \mathbf{C}_l - \mathbf{C}_l \mathbf{B}^T \mathbf{M}^{-1} \mathbf{A} \mathbf{C}_{\hat{\mathbf{x}}} \mathbf{A}^T \mathbf{M}^{-1} \mathbf{B} \mathbf{C}_l \\ &= \mathbf{C}_l - \mathbf{C}_{\hat{\mathbf{l}}} . \end{aligned} \quad (2.65a)$$

The above equation corresponds to common sense, namely the variances of the observables after the adjustment are smaller than the variances of the observables (observations) before the adjustment since

$$\mathbf{C}_{\hat{\mathbf{l}}} = \mathbf{C}_l - \mathbf{C}_{\hat{\mathbf{r}}} . \quad (2.65b)$$

See equations 2.105 and 2.106 as examples.

2.4 Generation of the Standard Cases

Combined Case with Weighted Parameters (\mathbf{A} , \mathbf{B} , \mathbf{C}_l^{-1} , \mathbf{C}_x^{-1})

The general case of a non-linear implicit model with weighted parameters is known as the combined case with weighted parameters. It has a solution given by the following equations (2.28, 2.14, 2.29, 2.30, 2.15, 2.47, 2.52, 2.53, 2.64, 2.65, 2.43, 2.55, 2.56):

$$\begin{aligned} \hat{\boldsymbol{\delta}} &= -[\mathbf{A}^T (\mathbf{B} \mathbf{C}_l \mathbf{B}^T)^{-1} \mathbf{A} + \mathbf{C}_x^{-1}]^{-1} \mathbf{A}^T (\mathbf{B} \mathbf{C}_l \mathbf{B}^T)^{-1} \mathbf{w} \\ &= -[\mathbf{N} + \mathbf{C}_x^{-1}]^{-1} \mathbf{A}^T (\mathbf{B} \mathbf{C}_l \mathbf{B}^T)^{-1} \mathbf{w} \end{aligned} \quad (2.66)$$

$$= -[\mathbf{N} + \mathbf{C}_x^{-1}]^{-1} \mathbf{u} ,$$

$$\hat{\mathbf{x}} = \mathbf{x} + \hat{\boldsymbol{\delta}} \quad (2.67)$$

$$\begin{aligned} \hat{\mathbf{k}} &= (\mathbf{B} \mathbf{C}_l \mathbf{B}^T)^{-1} (\mathbf{A} \hat{\boldsymbol{\delta}} + \mathbf{w}) \\ &= \mathbf{M}^{-1} (\mathbf{A} \hat{\boldsymbol{\delta}} + \mathbf{w}) , \end{aligned} \quad (2.68)$$

$$\hat{\mathbf{r}} = -\mathbf{C}_l \mathbf{B}^T \hat{\mathbf{k}} \quad (2.69)$$

$$\hat{\mathbf{l}} = \mathbf{l} + \hat{\mathbf{r}} \quad (2.70)$$

$$\begin{aligned} \mathbf{C}_{\hat{\delta}} = & [\mathbf{A}^T (\mathbf{B}\mathbf{C}_l\mathbf{B}^T)^{-1} \mathbf{A} + \mathbf{C}_x^{-1}]^{-1} \mathbf{A}^T (\mathbf{B}\mathbf{C}_l\mathbf{B}^T)^{-1} \mathbf{A} \mathbf{C}_x \mathbf{A}^T (\mathbf{B}\mathbf{C}_l\mathbf{B}^T)^{-1} \mathbf{A} \\ & [\mathbf{A}^T (\mathbf{B}\mathbf{C}_l\mathbf{B}^T)^{-1} \mathbf{A} + \mathbf{C}_x^{-1}]^{-1} \end{aligned} \quad (2.71)$$

$$\begin{aligned} & + [\mathbf{A}^T (\mathbf{B}\mathbf{C}_l\mathbf{B}^T)^{-1} \mathbf{A} + \mathbf{C}_x^{-1}]^{-1} \mathbf{N} [\mathbf{A}^T (\mathbf{B}\mathbf{C}_l\mathbf{B}^T)^{-1} \mathbf{A} + \mathbf{C}_x^{-1}]^{-1} \\ = & [\mathbf{N} + \mathbf{C}_x^{-1}]^{-1} \mathbf{N} \mathbf{C}_x \mathbf{N} [\mathbf{N} + \mathbf{C}_x^{-1}]^{-1} + [\mathbf{N} + \mathbf{C}_x^{-1}]^{-1} \mathbf{N} [\mathbf{N} + \mathbf{C}_x^{-1}]^{-1}, \end{aligned}$$

$$\begin{aligned} \mathbf{C}_{\hat{x}} = & \mathbf{C}_x + \mathbf{C}_{\hat{\delta}} - \mathbf{C}_x \mathbf{A}^T (\mathbf{B}\mathbf{C}_l\mathbf{B}^T)^{-1} \mathbf{A} [\mathbf{A}^T (\mathbf{B}\mathbf{C}_l\mathbf{B}^T)^{-1} \mathbf{A} + \mathbf{C}_x^{-1}]^{-1} \\ & - [\mathbf{A}^T (\mathbf{B}\mathbf{C}_l\mathbf{B}^T)^{-1} \mathbf{A} + \mathbf{C}_x^{-1}]^{-1} \mathbf{A}^T (\mathbf{B}\mathbf{C}_l\mathbf{B}^T)^{-1} \mathbf{A} \mathbf{C}_x \\ = & \mathbf{C}_x + \mathbf{C}_{\hat{\delta}} - \mathbf{C}_x \mathbf{N} [\mathbf{N} + \mathbf{C}_x^{-1}]^{-1} - [\mathbf{N} + \mathbf{C}_x^{-1}]^{-1} \mathbf{N} \mathbf{C}_x \\ = & [\mathbf{A}^T (\mathbf{B}\mathbf{C}_l\mathbf{B}^T)^{-1} \mathbf{A} + \mathbf{C}_x^{-1}]^{-1} = [\mathbf{N} + \mathbf{C}_x^{-1}]^{-1} \end{aligned} \quad (2.72)$$

$$\mathbf{C}_{\hat{l}} = \mathbf{C}_l + \mathbf{C}_l \mathbf{B}^T \mathbf{M}^{-1} \mathbf{A} \mathbf{C}_{\hat{x}} \mathbf{A}^T \mathbf{M}^{-1} \mathbf{B} \mathbf{C}_l - \mathbf{C}_l \mathbf{B}^T \mathbf{M}^{-1} \mathbf{B} \mathbf{C}_l \quad (2.73)$$

$$\mathbf{C}_{\hat{r}} = \mathbf{C}_l \mathbf{B}^T \mathbf{M}^{-1} \mathbf{B} \mathbf{C}_l - \mathbf{C}_l \mathbf{B}^T \mathbf{M}^{-1} \mathbf{A} \mathbf{C}_{\hat{x}} \mathbf{A}^T \mathbf{M}^{-1} \mathbf{B} \mathbf{C}_l \quad (2.74)$$

$$\mathbf{C}_w = \mathbf{A} \mathbf{C}_x \mathbf{A}^T + \mathbf{B} \mathbf{C}_l \mathbf{B}^T \quad (2.75)$$

$$\hat{\sigma}_o^2 = \frac{\hat{\mathbf{r}}^T \mathbf{C}_l^{-1} \hat{\mathbf{r}} + \hat{\delta}^T \mathbf{C}_x^{-1} \hat{\delta}}{v} \quad (2.76)$$

$$v = m - u + m_x, \quad (2.77)$$

If the reference variance σ_o^2 is unknown, then $\hat{\sigma}_o^2$ can be used to scale the above as follows:

$$\hat{\mathbf{C}}_{\hat{\delta}} = \hat{\sigma}_o^2 \mathbf{C}_{\hat{\delta}} \quad (2.78)$$

$$\hat{\mathbf{C}}_{\hat{x}} = \hat{\sigma}_o^2 \mathbf{C}_{\hat{x}} \quad (2.79)$$

$$\hat{\mathbf{C}}_{\hat{l}} = \hat{\sigma}_o^2 \mathbf{C}_{\hat{l}} \quad (2.80)$$

$$\hat{\mathbf{C}}_{\hat{r}} = \hat{\sigma}_o^2 \mathbf{C}_{\hat{r}} \quad (2.81)$$

$$\hat{\mathbf{C}}_w = \hat{\sigma}_o^2 \mathbf{C}_w \quad (2.82)$$

In the above expressions

$$\mathbf{M} = \mathbf{B} \mathbf{C}_1 \mathbf{B}^T \quad (2.83)$$

$$\mathbf{N} = \mathbf{A}^T (\mathbf{B} \mathbf{C}_1 \mathbf{B}^T)^{-1} \mathbf{A} = \mathbf{A}^T \mathbf{M}^{-1} \mathbf{A} \quad (2.84)$$

$$\mathbf{u} = \mathbf{A}^T (\mathbf{B} \mathbf{C}_1 \mathbf{B}^T)^{-1} \mathbf{w} = \mathbf{A}^T \mathbf{M}^{-1} \mathbf{w} \quad (2.85)$$

An intriguing mathematical fact concerning the variance-covariance matrices for the parameters is that the inverse of the coefficient matrix of the normal equations is the covariance matrix of the final (adjusted) parameters $\hat{\mathbf{x}}$ and not of the solution vector $\hat{\boldsymbol{\delta}}$ [Kouba 1970].

If all the parameters are weighted, note that the number of degrees of freedom becomes equal to the number of equations. This is analogous to the condition case below in which all observations are weighted. This is not surprising for, in the present case, all quantities are also weighted. Schmid and Schmid [1965] call this *generalized least squares*.

One should also note that the variance-covariance matrix of \mathbf{w} is defined with the a priori variance factor σ_o^2 which allows statistical testing **before** the adjustment takes place, if σ_o^2 is indeed known. In the case that σ_o^2 is not known, then an estimate may be obtained from the adjustment itself. Hamilton [1964] shows that in the latter case (σ_o^2 not known), the confidence region for the adjusted parameters $\hat{\mathbf{x}}$ are given in terms of the Fisher distribution, while if σ_o^2 is known, the confidence region is described through the multivariate Chi-squared distribution.

Combined Case ($\mathbf{A}, \mathbf{B}, \mathbf{C}_1^{-1}, \mathbf{C}_x^{-1} = 0$)

The combined case is characterized by a non-linear implicit mathematical model with no weights on the parameters. We deduce the corresponding set of expressions from the general case by considering that if there are no weights then \mathbf{C}_x^{-1} is equal to zero. This implies that \mathbf{x} is a constant vector (now denoted by \mathbf{x}^0), and its variance covariance matrix \mathbf{C}_{x0} does not exist. As a consequence, both \mathbf{C}_x and \mathbf{C}_x^{-1} do not exist. Also note that the partial derivatives of $\hat{\mathbf{x}}$ with respect to \mathbf{x}^0 will also be a null matrix. Upon substitution of the three null matrices into equations 2.9 through 2.66, we get the desired results:

$$\hat{\boldsymbol{\delta}} = - [\mathbf{A}^T (\mathbf{B} \mathbf{C}_1 \mathbf{B}^T)^{-1} \mathbf{A}]^{-1} \mathbf{A}^T (\mathbf{B} \mathbf{C}_1 \mathbf{B}^T)^{-1} \mathbf{w}^0$$

$$= -\mathbf{N}^{-1}\mathbf{u} , \quad (2.86)^*$$

$$\hat{\mathbf{x}} = \mathbf{x}^0 + \hat{\boldsymbol{\delta}} , \quad (2.87)$$

$$\begin{aligned} \hat{\mathbf{k}} &= (\mathbf{B}\mathbf{C}_1\mathbf{B}^T)^{-1} (\mathbf{A} \hat{\boldsymbol{\delta}} + \mathbf{w}^0) \\ &= \mathbf{M}^{-1} (\mathbf{A} \hat{\boldsymbol{\delta}} + \mathbf{w}^0), \end{aligned} \quad (2.88)$$

$$\hat{\mathbf{r}} = -\mathbf{C}_1\mathbf{B}^T \hat{\mathbf{k}} , \quad (2.89)$$

$$\hat{\mathbf{l}} = \mathbf{l} + \hat{\mathbf{r}} \quad (2.90)$$

$$\mathbf{C} \hat{\boldsymbol{\delta}} = \mathbf{N}^{-1} = [\mathbf{A}^T (\mathbf{B}\mathbf{C}_1\mathbf{B}^T)^{-1} \mathbf{A}]^{-1} = \mathbf{C}_{\hat{\mathbf{x}}} , \quad (2.91)$$

$$\mathbf{C} \hat{\mathbf{l}} = \mathbf{C}_1 + \mathbf{C}_1\mathbf{B}^T\mathbf{M}^{-1}\mathbf{A} \mathbf{N}^{-1}\mathbf{A}^T\mathbf{M}^{-1}\mathbf{B}\mathbf{C}_1 - \mathbf{C}_1 \mathbf{B}^T\mathbf{M}^{-1} \mathbf{B}\mathbf{C}_1 , \quad (2.92)$$

$$\mathbf{C} \hat{\mathbf{r}} = \mathbf{C}_1\mathbf{B}^T\mathbf{M}^{-1}\mathbf{B} \mathbf{C}_1 - \mathbf{C}_1\mathbf{B}^T\mathbf{M}^{-1} \mathbf{A} \mathbf{N}^{-1}\mathbf{A}^T\mathbf{M}^{-1}\mathbf{B}\mathbf{C}_1 , \quad (2.93)$$

$$\mathbf{C}_w = \mathbf{B}\mathbf{C}_1\mathbf{B}^T , \quad (2.94)$$

$$\hat{\sigma}_o^2 = \frac{\hat{\mathbf{r}}^T \mathbf{C}_1^{-1} \hat{\mathbf{r}}}{v} \quad (2.95)$$

$$v = m - u . \quad (2.96)$$

We witness that the weight coefficient matrix of the correction vector $\hat{\boldsymbol{\delta}}$ and adjusted vector $\hat{\mathbf{x}}$ are identical and equal to the inverse of the coefficient matrix of the normal equations. The degrees of freedom is calculated as the difference between the number of equations and the number of unknown parameters.

Parametric Case ($\mathbf{A}, \mathbf{B} = -\mathbf{I}, \mathbf{C}_1^{-1}, \mathbf{C}_x^{-1} = 0$)

The parametric case is characterized by a non-linear explicit model. This means that the observables can be explicitly expressed as some non-linear function of the parameters thus the reason for the second design matrix is to be equal to a minus identity matrix. Setting \mathbf{B} equal to $-\mathbf{I}$ in the combined case with no weights on the parameters we get the following expressions:

* $\mathbf{w}^0 = f(\mathbf{x}^0, \mathbf{l})$

$$\hat{\delta} = -[A^T C_l^{-1} A]^{-1} A^T C_l^{-1} w^0, \quad (2.97)$$

$$\hat{x} = x^0 + \hat{\delta} \quad (2.98)$$

$$\hat{k} = C_l^{-1} (A \hat{\delta} + w^0), \quad (2.99)$$

$$\hat{r} = C_l \hat{k}, \quad (2.100)$$

$$\hat{l} = 1 + \hat{r}, \quad (2.101)$$

$$\hat{\sigma}_o^2 = \frac{\hat{r}^T C_l^{-1} \hat{r}}{v}, \quad (2.102)$$

$$v = m - u = n - u, \quad (2.103)$$

$$C_{\hat{\delta}} = C_{\hat{x}} = [A^T C_l^{-1} A]^{-1}, \quad (2.104)$$

$$C_{\hat{l}} = A [A^T C_l^{-1} A]^{-1} A^T, \quad (2.105)$$

$$C_{\hat{r}} = C_l - A [A^T C_l^{-1} A]^{-1} A^T = C_l - C_{\hat{l}}, \quad (2.106)$$

$$C_w = C_l. \quad (2.107)$$

Note that the number of equations equals the number of observations; this is not true for the combined case. We see that the weight coefficient matrix of the adjusted observables has the form of a propagation of errors (covariance law) from the adjusted parameters into these quantities.

Condition Case ($A = 0, B, C_l^{-1}, C_x^{-1} = 0$)

The condition case is characterized by a non-linear model consisting of only observables, thus the first design matrix A in the combined case with no weights on the parameters vanishes, yielding:

$$\begin{aligned} \hat{k} &= [B C_l B^T]^{-1} w \\ &= M^{-1} w, \end{aligned} \quad (2.108)$$

$$\hat{r} = -C_l B^T \hat{k}, \quad (2.109)$$

$$\hat{l} = 1 + \hat{r}, \quad (2.110)$$

$$\hat{\sigma}_o^2 = \frac{\hat{\mathbf{r}}^T \mathbf{C}_1^{-1} \hat{\mathbf{r}}}{v}, \quad (2.111)$$

$$v = m, \quad (2.112)$$

$$\mathbf{C}_{\hat{\mathbf{l}}} = \mathbf{C}_l - \mathbf{C}_l \mathbf{B}^T (\mathbf{B} \mathbf{C}_l \mathbf{B}^T)^{-1} \mathbf{B} \mathbf{C}_l, \quad (2.113)$$

$$\mathbf{C}_{\hat{\mathbf{r}}} = \mathbf{C}_l \mathbf{B}^T \mathbf{M}^{-1} \mathbf{B} \mathbf{C}_l, \quad (2.114)$$

The above case can be used to solve the combined case with weighted parameters as follows, given

$$\mathbf{B}^* \mathbf{r}^* + \mathbf{w} = 0, \quad \mathbf{C}_l^{*-1} \quad (2.115)$$

where

$$\mathbf{B}^* = [\mathbf{A} \quad \mathbf{B}], \quad (2.116)$$

$$\mathbf{r}^* = \begin{pmatrix} \hat{\boldsymbol{\delta}} \\ \hat{\mathbf{r}} \end{pmatrix} \quad (2.117)$$

$$\mathbf{C}_l^* = \begin{pmatrix} \mathbf{C}_x & 0 \\ 0 & \mathbf{C}_l \end{pmatrix}, \quad (2.118)$$

then substitute these definitions of \mathbf{B}^* and \mathbf{C}_l^* into the equations for the condition case immediately above. Note that the design matrix \mathbf{A} and weight matrix \mathbf{C}_x^{-1} both pertain to the weighted parameters.

REFERENCE

BOSSLER, J.D. *Bayesian inference in geodesy*. Ph.D. Dissertation, Dept. of Geodetic Science, The Ohio State University, Columbus, 1972.

HAMILTON, W.C. *Statistics in physical science*. Ronald, New York, 1964.

KOUBA, J. *Generalized sequential least squares expressions and matlan programming*. M.Sc. Thesis, Dept. of Surveying Engineering, University of New Brunswick, Fredericton, 1970.

KRAKIWSKY, E.J. A synthesis of Recent Advances: Method of Least Squares. ENGO629 Lecture Notes, Department of Geomatics Engineering, The University of Calgary, Calgary, Alberta, Canada.

KRAKIWSKY, E.J. Sequential least squares adjustment of satellite triangulation and trilateration in combination with terrestrial data. *Reports of the Department of Geodetic Science*, No. 114, The Ohio State University, Columbus, 1968.

WELLS, D.E. and KRAKIWSKY, E.J. The method of least squares. *Department of Surveying Engineering Lecture Notes* No. 18, University of New Brunswick, Fredericton, 1971.

SCHMID, H.H. and SCHMID, E. A generalized least squares solution for a hybrid measuring system. *The Canadian Surveyor*, XIX, No. 1, Ottawa, 1965.

THOMSON, E.H. *An introduction to the algebra of matrices with some applications*. The University of Toronto Press, 1969.

3

KALMAN FILTERING

In this section, the concept of random process and kinematic modeling are described. The derivation of Kalman filter equations and their implementation aspects are provided.

3.1 Random Process and Properties

Random or stochastic process is a random variable (collection or ensemble) evolving with time, usually denoted as $\{x(t)\}$. It is thus an extension of the concept of conventional random variable in statistics and it is fundamental for real-time signal processing such as Kalman filtering.

Taking each $x(t_i)$, $i=1,2, \dots$ as a random variable at many points in time (for instance at equally spaced intervals) or as a random process:

$$x(t_1), x(t_2), \dots \quad (3.1)$$

We can analyze the stochastic characteristics such as correlation of the signal by observing values of it at different time using the methods developed before. Another possible way to obtain the stochastic properties of a random process is to have a large collection or ensemble by measuring the values of each signal at one point in time. For example, in analyzing ocean waves we may record them at different geographical locations at the same time. In such as case, we would obtain a set, or ensemble, of record:

$${}_1x(t_1), {}_2x(t_1), \dots \quad (3.2)$$

from which we could also obtain a mean value, mean squares value and so on.

Now if based on information from equation 3.2, we can determine the mean and signal auto-correlation by

$$\mu_x(k) = \lim_{T \rightarrow \infty} \frac{1}{T} \int_0^T x(t)_k dt \quad (3.3)$$

$$R_{xx}(t, t + \tau, k) = \lim_{T \rightarrow \infty} \frac{1}{T} \int_0^T x(t)_k x(t + \tau) dt \quad (3.4)$$

where k is considered fixed once data has been chosen.

If we have another random process denoted as

$$y(t_1), y(t_2), \dots \quad (3.5)$$

we can define the following cross-correlation

$$R_{xy}(t, t + \tau, k) = \lim_{T \rightarrow \infty} \frac{1}{T} \int_0^T x(t)_k y(t + \tau) dt \quad (3.6)$$

Let us first consider discrete data set

$$x_1, x_2, \dots \text{ and } y_1, y_2, \dots \quad (3.7)$$

which are distributed with uniform probability density function, the cross correlation/covariance can be determined by

$$R_{xy}(\tau) \approx \frac{1}{N+1} [x_0 y_\tau + x_1 y_{1+\tau} + \dots + x_N y_{N+\tau}] \quad (3.8)$$

using $N+1$ samples.

On the other hand, if based on information from equation 3.2, the mean and auto-correlation value can be evaluated by

$$\mu_x(t_1) = \lim_{N \rightarrow \infty} \frac{1}{N} \sum_{k=1}^N x(t_1)_k \quad (3.9)$$

$$R_{xx}(t_1, t_1 + \tau, k) = \lim_{N \rightarrow \infty} \frac{1}{N} \sum_{k=1}^N x(t_1)_k x(t_1 + \tau)_k \quad (3.10)$$

where time t_1 is considered fixed once data has been chosen. If considering another random process of $y(t)$, we can similarly define the cross-correlation over the ensemble space

$$R_{xy}(t_1, t_1 + \tau, k) = \lim_{N \rightarrow \infty} \frac{1}{N} \sum_{k=1}^N x(t_1)_k y(t_1 + \tau) \quad (3.11)$$

We now found that the two ways have provided with different expressions and also possibly with different results. Since both approaches could be employed in applications, we are interested in if the results are the same statistically derived based on equation 3.3 -3.4 and equation 3.9 - 3.10 under certain conditions. If yes, it will greatly simplify the underlying data acquisition procedures and data analysis and at the same time also improve the flexibility of data acquisition.

In the following we discuss the concept of stationarity and ergodic for a random process.

- Stationary Random Data

- a) Strongly stationary

When the moments and joint moments of the random process $\{x(t)\}$ are time-invariant, $\{x(t)\}$ is called strongly stationary. It describes the property of the random process over time. For many practical applications, an assumption for strong stationarity is justified if the following weak stationarity is verified.

- b) Weakly stationary

A (weakly) stationary time series is a collection of random variables $\{x_t\}$ defined for all real t or all integers t , as the case may be, with the following properties:

$$1) \mu_x(t_1) = \lim_{N \rightarrow \infty} \frac{1}{N} \sum_{k=1}^N x(t_1)_k = \mu_x = \text{constant, in other words } E(x_t) \text{ is constant or independent of time.}$$

$$2) R_{xx}(t_1, t_1 + \tau, k) = \lim_{N \rightarrow \infty} \frac{1}{N} \sum_{k=1}^N x(t_1)_k x(t_1 + \tau) = R_{xx}(\tau), \text{ in other words, } E(x_{t_1} x_{t_2}) \text{ depends on only } t_2 - t_1 \text{ or time lag.}$$

Weak stationarity means that as time t_1 varies the mean value and the autocorrelation do not change.

Since $E(x)$ is constant, it may be estimated and subtracted from the series with little effect on what follows. Henceforth we shall assume that $E(x) = 0$, and state where necessary the modifications to be made if the series is subtracted from the

data. Properties 2) implies that $\text{Var}(x)$ is constant and that $\text{cov}(x_t, x_{t+\tau}) = R_{xx}(\tau)$ does not depend on t , which is the theoretical autocovariance at lag τ of $\{x(t)\}$ and may be estimated.

- Ergodic Random Data

If $\{x(t)\}$ is stationary, the process is called ergodic if the time-averaged moments and joint-moments are equal to the corresponding ensemble moments and joint-moments, that is, if k denotes the k^{th} sample function, the following properties hold:

$$\mu_x(k) = \lim_{T \rightarrow \infty} \frac{1}{T} \int_0^T x(t) dt = \mu_x = \text{constant} \quad (3.11)$$

$$R_{xx}(t, t + \tau, k) = \lim_{T \rightarrow \infty} \frac{1}{T} \int_0^T x(t)_k x(t + \tau) dt = R_{xx}(\tau) \quad (3.12)$$

It describes the relationship of the stochastic properties of a random process in the ensemble and time domain. Note that only stationary processes can be ergodic.

With the help of the previous discussions, we are able to define auto and cross correlation functions in random processes. If $\{x(t)\}$ and $\{y(t)\}$ are two arbitrary but stationary random processes, then they can be characterized by constant mean values

$$\mu_x = E\{X\} = \int_{-\infty}^{\infty} x p(x) dx; \quad (3.13)$$

$$\mu_y = E\{Y\} = \int_{-\infty}^{\infty} y p(y) dy; \quad (3.14)$$

and auto and cross correlation functions:

$$\left. \begin{aligned} R_{xx}(\tau) &= E\{x_1 x_2\} = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x_1 x_2 p(x_1, x_2) dx_1 dx_2 \\ R_{yy}(\tau) &= E\{y_1 y_2\} = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} y_1 y_2 p(y_1, y_2) dy_1 dy_2 \\ R_{xy}(\tau) &= E\{x_1 y_2\} = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x_1 y_2 p(x_1, y_2) dx_1 dy_2 \end{aligned} \right\} \quad (3.15)$$

where $x_1 = x(t)$, $x_2 = x(t + \bullet)$, $y_1 = y(t)$, $y_2 = y(t + \bullet)$.

Note that, if the means are subtracted, the auto and cross covariance functions are computed in place of the R-functions, and

$$\left. \begin{aligned} C_{xx}(\tau) &= R_{xx}(\tau) - \mu_x^2 \\ C_{yy}(\tau) &= R_{yy}(\tau) - \mu_y^2 \\ C_{xy}(\tau) &= R_{xy}(\tau) - \mu_x \mu_y \end{aligned} \right\} \quad (3.16)$$

Since the processes are stationary, the joint pdf's are independent of t . Two processes will be uncorrelated when $C_{xy}(\tau) = 0$ for all τ .

When the processes are also ergodic, the time averages can be used to estimate mean values and auto/cross-correlation functions instead of the ensemble averages. Given in the following are some important properties related to correlation functions:

a) $R_{xx}(0) = E\{x^2\} = \psi_x^2$ (mean square value)

$$C_{xx}(0) = R_{xx}(0) - \mu_x^2 = \psi_x^2 - \mu_x^2 = \sigma_x^2 \text{ (variance)}$$

b) $R_{xx}(-\tau) = R_{xx}(\tau)$ (even function)

$$R_{xy}(-\tau) = R_{yx}(\tau) \text{ (neither even, nor odd)}$$

c) $|R_{xx}(\tau)| \leq R_{xx}(0)$ for all τ

$$|R_{xy}(\tau)|^2 \leq R_{xx}(0)R_{yy}(0)$$

d) If $\{x(t)\}$ contains a periodic component, R_{xx} will also contain a periodic component with the same period.

e) If $x(t)$ does not contain any constant or periodic component,

$$\lim_{\tau \rightarrow \infty} R_{xx}(\tau) = 0$$

f) $|C_{xy}(\tau)|^2 \leq C_{xx}(0)C_{yy}(0) = \sigma_x^2 \sigma_y^2$

g) Normalized cross-covariance function or correlation coefficient function satisfies $\rho_{xy}(\tau) = \frac{C_{xy}(\tau)}{\sigma_x \sigma_y}$, $-1 \leq \rho_{xy}(\tau) \leq 1$

Random process properties can be described in the frequency domain where the cross- (or auto-, when $x = y$) spectral density function is defined as

$$S_{xy}(f) = F\{R_{xy}(\tau) = \int_{-\infty}^{\infty} R_{xy}(\tau) e^{-j2\pi f\tau} d\tau\} \quad (3.17)$$

and the coherence function (corresponding to the correlation coefficient function):

$$\gamma_{xy}^2(f) = \frac{|S_{xy}(f)|^2}{S_{xx}(f)S_{yy}(f)} = \frac{|G_{xy}(f)|^2}{G_{xx}(f)G_{yy}(f)}$$

$$0 \leq \gamma_{xy}^2(f) \leq 1$$

Given in the following are important properties associated with cross-spectral density functions:

$$a) \quad S_{xy}(-f) = S_{xy}^*(f) = S_{yx}(f)$$

$$b) \quad S_{xx}(-f) = S_{xx}^*(f) = S_{xx}(f)$$

$$c) \quad R_{xy}(\tau) = F^{-1}\{S_{xy}(f)\} = \int_{-\infty}^{\infty} S_{xy}(f) e^{j2\pi f\tau} df$$

$$d) \quad R_{xx}(0) = E\{X^2(t)\} = \psi_x^2 = \int_{-\infty}^{\infty} S_{xx}(f) df$$

$$e) \quad |S_{xy}(f)|^2 \leq S_{xx}(f)S_{yy}(f)$$

3.2 Gauss-Markov Process

It is well known that the ordinary differential equations play an important role in the analysis of deterministic systems. The importance stems in large part from certain nice mathematical properties that they enjoy. The following differential equation

$$\frac{dx(t)}{dt} = \dot{x}(t) = f(x(t)) \quad (3.18)$$

indicates that the rate of change of x at time t depends only on x at t (now), and not on x in the past, i.e., $x(t')$, $t' < t$. As a result of this, for any $t_1 < t_2$,

$$x(t_2) = g(t_2; x(t_1), t_1) \quad (3.19)$$

That is, the solution at t_2 is a function of $x(t_1)$ and does not depend on $x(t')$, $t' < t_1$.

This property of ordinary differential equations has lead to establish its stochastic analog in a class of process called Markov process. It describes the dynamic system's behaviors over time.

A random process $\{x(t)\}$ is called first order Markov Process if for $\{t_i : t_i < t_{i+1}\}$, the conditional density function satisfies

$$p\{x(t_n) | x(t_1), x(t_2), \dots, x(t_{n-1})\} = p\{x(t_n) | x(t_{n-1})\}. \quad (3.20)$$

The equation says that the probabilistic properties of the process in the future, once is in a given state, does not depend on how the process arrived at the given state. This property is sometimes referred to as the generalized causality principle: the future can be predicted from knowledge of the present. In fact, Markov property is a basic assumption that is made in the study of stochastic dynamic systems.

Consider a random sequence $\{x_i, i = 1, 2, \dots, n\}$, where the x_i are mutually independent. The process is a Markov sequence since

$$p\{x_n | x_1, x_2, \dots, x_{n-1}\} = p\{x_n | x_{n-1}\} = p(x_n). \quad (3.21)$$

Markov process concept can be extended to include situations that the value of the random variable at a given epoch could depend on the behavior of the values of the random variables in the past but within a limited time interval/memory. This results in different orders for a Markov process.

- 1) Zero-order Markov: the value of the variable at any epoch is independent of the value at any other epochs.

In zero-order, the variable value cannot be predicted even if the variable is estimated at earlier epochs. Such process is the "white sequence" or "white noise".

- 2) First order Markov: the variable at one epoch is only related to its value at the last epoch. The variable at the last epoch is replaced by the first derivative of the random function in a continuous case. The process is the basis of the state space model used in Kalman filtering.

- 3) Second order Markov: the variable at one epoch is related to its value at the last two epochs. The variable at the two last epochs is replaced by the first and second derivatives of the random function in a continuous case.

Markov processes higher than the second order are seldom used in applications.

A Markov process with Gaussian probability density functions describing the process is called Gauss-Markov process. Gauss-Markov process is a process that has been widely used in positioning and navigation applications.

3.3 Kinematic Modeling and Transition Matrix

A kinematic model is the mathematical expression for the predictability of the motion of a dynamic system. This predictability means that the parameters of interests for the system under consideration are not entirely random, but have values that are, within certain bounds, related to their values at an earlier epoch.

Consider a vehicle traveling along a straight line. After a couple of positions are determined, future positions can be predicted by extrapolation. This can be realized through a function model such as constant velocity model within a small time interval. However, the uncertainty associated with the prediction grows with time if no new measurements are made.

Similar to other modeling problems, a kinematic model consists of two components:

- 1) Functional model: parameter prediction based on previous results
- 2) Stochastic model: describe the uncertainty or precision associated with the predicted parameters. This depends on the imperfection of the functional model used to describe the kinematics of the underlying system.

Since the function model is used to describe the deterministic part of the dynamic system and differential equations are often used for this purpose. The general procedure in practical applications for kinematic modeling include

- 1) Describe the multi-variable dynamic system process to be estimated as an n-th order system of linear differential equations driven by white noise.
- 2) Arrange differential equations in convenient vector (state-space) form:

$$\dot{\mathbf{x}}(t) = \mathbf{F}(t)\mathbf{x}(t) + \mathbf{G}(t)\mathbf{w}(t) \quad \mathbf{Q}(t) \quad (3.22)$$

where $Q(t)$ is the spectral density matrix.

In fact, higher order of differential equations can be expressed by first order linear differential equations under the concept of state-space model. For example, given a higher order differential equation

$$\ddot{x}(t) + a_2 \dot{x}(t) + a_1 \dot{x}(t) + a_0 x(t) - w(t) = 0 \quad (3.23)$$

which can be expressed by the following vector expression

$$\begin{bmatrix} \dot{x}(t) \\ \ddot{x}(t) \\ \ddot{x}(t) \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ -a_0 & -a_1 & -a_2 \end{bmatrix} \begin{bmatrix} x(t) \\ \dot{x}(t) \\ \ddot{x}(t) \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ w(t) \end{bmatrix} \quad (3.24)$$

Let $x_1(t) = x(t)$, $x_2(t) = \dot{x}(t)$, $x_3(t) = \ddot{x}(t)$, then we have

$$\begin{bmatrix} \dot{x}_1(t) \\ \dot{x}_2(t) \\ \dot{x}_3(t) \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ -a_0 & -a_1 & -a_2 \end{bmatrix} \begin{bmatrix} x_1(t) \\ x_2(t) \\ x_3(t) \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ w(t) \end{bmatrix} \quad (3.25)$$

$$\Leftrightarrow \dot{x}(t) = Fx(t) + w(t) \quad (3.26)$$

We see that the first differential equation has been described by a first order vector differential equation model which has been often called a state space model. Solving the vector differential equation for step size leads to obtain the following discrete form:

$$x_{k+1} = \Phi_{k+1,k} x_k + w_k \quad Q_k \quad (3.27)$$

where $\Phi_{k+1,k}$, called transition matrix, can be obtained by linear system theory while the covariance matrix Q_k associated with w_k can be obtained from random process theory.

In geomatics applications, the kinematic model is used to describe the motion and the dynamic noise of a vehicle whose instantaneous position and velocity are sought, which is the key difference from static applications. A general form of the kinematic model for a system can be represented by the state space model, in which a set of first order linear differential equations express deviations from a reference trajectory, i.e.

$$\dot{x}(t) = Fx(t) + Gw(t), \quad Q(t) \quad (3.28)$$

where

x is the state vector ($m \times 1$),

\dot{x} is the time derivative of the state vector ($m \times 1$),

F is the system dynamic matrix ($m \times m$),

G is the coefficient matrix of the random forcing function ($m \times m$),

w is the system noise vector ($m \times 1$) which is usually assumed to be white noise,

t is time, and

m is the number of states in the state vector.

$Q(t)$ is the spectral density matrix.

The state vector contains all system parameters such as for instance the position unknowns in GPS positioning as well as the system biases such as GPS receiver clock offset. The solution of the first order homogeneous differential equation, $\dot{x}(t) = Fx(t)$, can be written as

$$x(t) = \Phi(t, t_0)x(t_0) \quad (3.29)$$

where Φ is called transition matrix and satisfies the equations

$$\Phi(t_0, t_0) = I \quad (3.30)$$

$$\dot{\Phi}(t, t_0) = F\Phi(t, t_0). \quad (3.31)$$

The particular solution of equation 3.28 with random forcing function $w(t)$ can be written as

$$x(t) = \Phi(t, t_0)x(t_0) + \int_{t_0}^t \Phi(t, \tau)G(\tau)w(\tau)d\tau \quad (3.32)$$

which is often called the matrix superposition integral (Gelb, 1974). The corresponding variance-covariance matrix of the state is given by

$$\begin{aligned} C_x(t) &= E[x(t)x(t)^T] \\ &= \Phi(t, t_0)C_x(t_0)\Phi(t, t_0)^T + \int_{t_0}^t \int_{t_0}^t \Phi(t, \tau)G(\tau)E[w(\tau)w(s)^T]G(s)^T\Phi(t, s)^T d\tau ds \end{aligned} \quad (3.33)$$

where it is assumed that $x(t_0)$ and $w(t)$ are uncorrelated and $C_x(t_0)$ is the variance covariance matrix of the initial state $x(t_0)$ at t_0 . When $w(t)$ is white noise, equation 3.43 further reduces to

$$C_x(t) = \Phi(t, t_0)C_x(t_0)\Phi(t, t_0)^T + \int_{t_0}^t \Phi(t, \tau)G(\tau)Q(\tau)G(\tau)^T\Phi(t, \tau)^T d\tau \quad (3.34)$$

For a time invariant system (i.e., F is a constant matrix), the transition matrix $\Phi(t, t_0)$ is only a function of the time difference $(t-t_0)$. In this case, the transition matrix can thus be expressed as the matrix exponential (Gelb, 1974).

$$\Phi(t, t_0) = e^{F\Delta t} \quad (3.35)$$

where $\Delta t = t - t_0$. The above equation can be expanded into a Taylor series

$$\Phi(t, t_0) = I + \Delta t F + \frac{1}{2!} \Delta t^2 F F + \frac{1}{3!} \Delta t^3 F F F + \dots \quad (3.36)$$

where I is an identity matrix. For a small time interval Δt , it may be sufficiently approximated by

$$\Phi(t, t_0) \approx I + \Delta t F \quad (3.37)$$

Transforming equations 3.32 and 3.34 into discrete forms, we have

$$x_{k+1} = \Phi_{k+1,k} x_k + w_k \quad (3.38)$$

$$C_{xk+1} = \Phi_{k+1,k} C_{xk} \Phi_{k+1,k}^T + C_{wk}, \quad (3.39)$$

where

$$w_k = \int_0^{\Delta t} \Phi(t_k, \tau) G(\tau) w(\tau) d\tau \quad (3.40)$$

$$C_{wk} = \int_0^{\Delta t} \Phi(t_k, \tau) G(\tau) Q(\tau) G(\tau)^T \Phi(t_k, \tau)^T d\tau \quad (3.41)$$

where $\Delta t = t_{k+1} - t_k$.

3.4 A Geomatics Example of Kinematic Modelling

The state model of a random process presented in section 3.3 is general in form and will accommodate a wide variety of situations. All that is required is that the process under consideration be related to white noise via a linear differential equation. The noise sequence in the problem at hand can be modeled as Gauss-Markov process. In GPS navigation and kinematic positioning, the first order Gauss-Markov process is the most popular one and is discussed in some detail below.

A kinematic model having a first order Gauss-Markov process can be written as

$$\dot{x}(t) = -\beta x(t) + w(t) \quad (3.42)$$

where β is the inverse of the correlation length (we here assume for simplicity that all states have the same β). A larger value of β , i.e., a short correlation length, allows a large change in the state vector from one epoch to the next; a small value of β , i.e., a longer correlation length, describes a strong correlation between subsequent epochs and will allow little variation in the state vector. $\beta=0$ results in another popular process called random walk process. $w(t)$ is a white noise sequence which satisfies the equations

$$E(w(t)) = 0, \quad \text{and} \quad (3.43)$$

$$E(w(t)w(t+\tau)) = C_w(t, t+\tau) = Q\delta(\tau), \quad (3.44)$$

where Q is the spectral density matrix of $w(t)$ and $\delta(\tau)$ is the Dirac delta function. The relationship between the variance-covariance matrix and spectral density matrix can be found in Gelb (1974). Using equations 3.35 and 3.41, we have.

$$\Phi_{k+1,k} = e^{-\beta\Delta t} \mathbf{I} \quad (3.45)$$

$$C_{wk} = \frac{1}{2\beta} (1 - e^{-\beta\Delta t}) Q \quad (3.46)$$

The states in the system model should be selected so as to sufficiently describe the system behaviour in terms of the actual dynamics and data rate. For instance, the motion of a moving vehicle can be represented in three-dimensional space by

$$r(t) = \{\varphi(t), \lambda(t), h(t)\}^T, \quad (3.47)$$

$$\text{or} \quad \dot{r}(t) = \{v_N(t), v_E(t), v_h(t)\}^T, \quad (3.48)$$

$$\text{or } \ddot{\mathbf{r}}(t) = \{a_N(t), a_E(t), a_h(t)\}^T, \quad (3.49)$$

where

$\mathbf{r}(t)$ is the position vector of the vehicle at time t ,

ϕ, λ, h are latitude, longitude and height, respectively,

v_N, v_E, v_h are the velocities of the vehicle in the directions of north, east and height, respectively, and

a_N, a_E, a_h are the accelerations of the vehicle in the directions of north, east and height, respectively.

These three representations are in principle equivalent if the appropriate initial values are known and the time history is continuous. In practice, however, the representation closest to the measurement is chosen. For GPS positioning, either a constant velocity model or a constant acceleration model is often applied, where the choice depends on the dynamic environment of the moving vehicle and the observation data rate. In general, the inclusion of acceleration vector will make no significant improvement in positioning accuracy in the case of moderate vehicle dynamics. For the situation of high dynamics when the acceleration is desired, the acceleration states should be included and in this case raw acceleration data should be made in addition to GPS observables, because GPS does not provide raw acceleration data.

In the following, a constant velocity dynamic model is described with respect to GPS navigation and kinematic positioning (Gao, 1992).

For GPS navigation, an eight parameter state vector is chosen to describe the vehicle dynamics, which are defined as follows:

$$\mathbf{x} = \{\delta\phi, \delta\lambda, \delta h, \delta v_N, \delta v_E, \delta v_h, \delta T, \delta \dot{T}\}^T \quad (3.50)$$

where δ represents a correction to the parameter. The receiver clock error, along with the position and velocity parameters, are also modeled. The dynamic matrix, F , for this model is given as

$$F = \begin{pmatrix} 0 & 0 & 0 & \frac{1}{R} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \frac{1}{R \cos \varphi} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & -\beta_{v_N} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -\beta_{v_E} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & -\beta_{v_h} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & -\beta_{d\dot{T}} \end{pmatrix} \quad (3.51)$$

where R is the radius of the earth. The corresponding transition matrix, Φ , can be obtained based on equation 3.45, i.e.,

$$\Phi_k = \begin{pmatrix} 1 & 0 & 0 & \frac{1 - e^{-\beta_{v_N} \Delta t}}{\beta_{v_N} R} & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & \frac{1 - e^{-\beta_{v_E} \Delta t}}{\beta_{v_E} R \cos \varphi} & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & \frac{1 - e^{-\beta_{v_h} \Delta t}}{\beta_{v_h}} & 0 & 0 \\ 0 & 0 & 0 & e^{-\beta_{v_N} \Delta t} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & e^{-\beta_{v_E} \Delta t} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & e^{-\beta_{v_h} \Delta t} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & \frac{1 - e^{-\beta_{v_{d\dot{T}}} \Delta t}}{\beta_{v_{d\dot{T}}}} \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & e^{-\beta_{v_{d\dot{T}}} \Delta t} \end{pmatrix} \quad (3.52)$$

The corresponding variance-covariance matrix of the system process noise related to the state vector can be calculated by equation 3.41, i.e.,

$$C_{wk} = \int_0^{\Delta t} \Phi(\tau) Q(\tau) \Phi(\tau)^T d\tau = (C_{ij})_{8 \times 8} \quad (3.53)$$

with the spectral density of the system noise given by

$$Q = \text{diag}(0, 0, 0, q_{v_N}, q_{v_E}, q_{v_h}, 0, q_{d\dot{T}}) \quad (3.54)$$

The explicit expression of the non-zero elements in C_{wk} are given below:

$$\begin{aligned}
C_{11} &= \frac{q_{v_N}}{\beta_{v_N}^2} \left\{ \Delta \Delta t \frac{2}{\beta_{v_N}} (1 - e^{-\beta_{v_N} \Delta t}) + \frac{1}{2\beta_{v_N}} (1 - e^{-2\beta_{v_N} \Delta t}) \right\} \\
C_{22} &= \frac{q_{v_E}}{\beta_{v_E}^2} \left\{ \Delta \Delta t \frac{2}{\beta_{v_E}} (1 - e^{-\beta_{v_E} \Delta t}) + \frac{1}{2\beta_{v_E}} (1 - e^{-2\beta_{v_E} \Delta t}) \right\} \\
C_{33} &= \frac{q_{v_h}}{\beta_{v_h}^2} \left\{ \Delta \Delta t \frac{2}{\beta_{v_h}} (1 - e^{-\beta_{v_h} \Delta t}) + \frac{1}{2\beta_{v_h}} (1 - e^{-2\beta_{v_h} \Delta t}) \right\} \\
C_{44} &= \frac{q_{v_N}}{2\beta_{v_E}} (1 - e^{-2\beta_{v_E} \Delta t}) \\
C_{55} &= \frac{q_{v_E}}{2\beta_{v_E}} (1 - e^{-2\beta_{v_E} \Delta t}) \\
C_{66} &= \frac{q_{v_h}}{2\beta_{v_h}} (1 - e^{-2\beta_{v_h} \Delta t}) \\
C_{77} &= \frac{q_{d\dot{T}}}{\beta_{d\dot{T}}^2} \left\{ \Delta \Delta t \frac{2}{\beta_{d\dot{T}}} (1 - e^{-\beta_{d\dot{T}} \Delta t}) + \frac{1}{2\beta_{d\dot{T}}} (1 - e^{-2\beta_{d\dot{T}} \Delta t}) \right\} \\
C_{88} &= \frac{q_{d\dot{T}}}{2\beta_{d\dot{T}}} (1 - e^{-2\beta_{d\dot{T}} \Delta t}) \\
C_{14} = C_{41} &= \frac{q_{v_N}}{\beta_{v_N}} \left\{ \frac{1}{\beta_{v_N}} (1 - e^{-\beta_{v_N} \Delta t}) - \frac{1}{2\beta_{v_N}} (1 - e^{-2\beta_{v_N} \Delta t}) \right\} \\
C_{25} = C_{52} &= \frac{q_{v_E}}{\beta_{v_E}} \left\{ \frac{1}{\beta_{v_E}} (1 - e^{-\beta_{v_E} \Delta t}) - \frac{1}{2\beta_{v_E}} (1 - e^{-2\beta_{v_E} \Delta t}) \right\} \\
C_{36} = C_{63} &= \frac{q_{v_h}}{\beta_{v_h}} \left\{ \frac{1}{\beta_{v_h}} (1 - e^{-\beta_{v_h} \Delta t}) - \frac{1}{2\beta_{v_h}} (1 - e^{-2\beta_{v_h} \Delta t}) \right\} \\
C_{78} = C_{87} &= \frac{q_{d\dot{T}}}{\beta_{d\dot{T}}} \left\{ \frac{1}{\beta_{d\dot{T}}} (1 - e^{-\beta_{d\dot{T}} \Delta t}) - \frac{1}{2\beta_{d\dot{T}}} (1 - e^{-2\beta_{d\dot{T}} \Delta t}) \right\}
\end{aligned} \tag{3.55}$$

For a first-order Gauss-Markov process, the value q can be determined from the following equation:

$$q = 2\sigma^2\beta \tag{3.56}$$

where σ is the standard deviation of the system noise. We see that the spectral density is determined by the actual system noise level and the noise correlation behavior in the sequence.

For kinematic positioning with double differenced pseudorange, carrier phase and phase rate observables between receivers and satellites where the receiver clock offsets are cancelled out, a six state parameter state vector can be defined as follows:

$$\mathbf{x} = \{\delta\phi, \delta\lambda, \delta h, \delta v_N, \delta v_E, \delta v_h\}^T \quad (3.57)$$

A similar expression of the transition matrix Φ_k and dynamic variance matrix C_{wk} as the stand-alone GPS navigation case can be obtained, except that the elements for the clock errors are not presented in the matrix. Also, the dynamic matrix F and spectral density matrix Q become

$$F = \begin{pmatrix} 0 & 0 & 0 & \frac{1}{R} & 0 & 0 \\ 0 & 0 & 0 & 0 & \frac{1}{R\cos\phi} & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & -\beta_{v_N} & 0 & 0 \\ 0 & 0 & 0 & 0 & -\beta_{v_E} & 0 \\ 0 & 0 & 0 & 0 & 0 & -\beta_{v_h} \end{pmatrix} \quad (3.58)$$

and

$$Q = \text{diag}(0, 0, 0, q_{v_E}, q_{v_E}, q_{v_h}) \quad (3.59)$$

3.5 Kalman Filtering

Filtering is a data processing/estimation method for applications that the parameters of interest of the dominant system errors, or both, are time-varying. Furthermore, the time variation is more or less predictable. A time-varying system can be described by the following mathematical model:

1) In continuous case

$$\text{Dynamic model:} \quad \dot{\mathbf{x}}(t) = F\mathbf{x}(t) + G\mathbf{w}(t) \quad Q(t) \quad (3.60a)$$

$$\text{Measurement model} \quad y(t) = A(t)\mathbf{x}(t) + \varepsilon(t) \quad R(t) \quad (3.60b)$$

2) In discrete case

$$\text{Dynamic model} \quad x_{k+1} = \Phi_{k+1,k} x_k(t) + w_k \quad Q_k \quad (3.61a)$$

$$\text{Measurement model} \quad y_k = A_k x_k + \varepsilon_k \quad R_k \quad (3.61b)$$

We have already given details how to develop the dynamic model of equation 3.60a for continuous case and of equation 3.61a for discrete case. As to the measurement models given in equations 3.60b and 3.71b, the development is similar to the static cases that we are already familiar with.

Filtering has its origin in electrical signal processing, and consequently the literature mainly reflects a preoccupation with signal processing and communication engineering applications. There are various approaches toward filtering with respect to the problem in equation 3.60 and 3.61. Kalman filter is perhaps the best known of the techniques that have gained wide acceptance across a wide spectrum of physical and engineering science, and more recent years in mathematics. In Geomatics, it was with the advent of satellite navigation that had pushed the geodetists to use Kalman filter as a major data processing tool.

A Kalman filter is simply an optimal recursive data processing algorithm. There are many way of defining optimal often meaning that errors are minimized in some respect, the Kalman filter is optimal with respect to virtually any criterion that makes sense. One aspect of this optimality is that the Kalman filter incorporates all information that can be provided to it. It processes all available measurements regardless of their precision, to estimate the current value of the variables of interest, with the use of

- a) knowledge of the system and measurement device dynamics
- b) the statistical description of the system noises, measurement errors, and uncertainty in the dynamic models, and
- c) any available information about initial conditions of the variables of interest.

The word “recursive” in the previous description means that, unlike certain data processing concepts, the Kalman filter does not require all previous data to be kept in storage and reprocessed every time a new measurement is taken. This is of vital importance to the practicality of filter implementation.

The Kalman filtering method was developed by R.E. Kalman in 1960 which provides a recursive solution of the discrete-data filtering problem. The advance digital computer technology made it possible to consider implementing recursive solution in a number of real-time applications and quickly found wide

applications in engineering, and nowadays to numerous areas that beyond most people's imagination.

The derivations provided below are from Brown (1983) which is found quite intuitive and easy to follow. The Kalman filter equations can also be derived based on Least Squares method as shown in Krakiwsky (1992).

Assume the dynamic model of a random process has the form of

$$x_{k+1} = \Phi_k x_k + w_k \quad (3.62)$$

where

$x_k =$ (n x 1) process state vector at time t_k .

$\Phi_k =$ (n x n) matrix relating x_k to x_{k+1} in the absence of a forcing (If x_k is a sample of continuous process, Φ_k is the usual state transition matrix)

$w_k =$ (n x 1) vector- assumed to be a white (uncorrelated) sequence known covariance structure.

and the observation (measurement) model of the process at a specific time epoch t_k has a linearized form of

$$z_k = H_k x_k + v_k \quad (3.63)$$

where

$z_k =$ (m x 1) vector measurement at time t_k .

$H_k =$ (m x n) matrix giving the ideal (noiseless) connection between measurement and the state vector at time t_k .

$v_k =$ (m x 1) measurement error- assumed to be a white sequence with known covariance structure and uncorrelated with the w_k sequence.

For the stochastic models of the dynamic and measurement models, the mean values for the w_k and v_k vectors are assumed to equal to zero and their covariance matrices are given by

$$E[w_k w_i^T] = \begin{cases} Q_k & i = k \\ 0, & i \neq k \end{cases} \quad (3.64)$$

$$E[v_k v_i^T] = \begin{cases} R_k, & i = k \\ 0, & i \neq k \end{cases} \quad (3.65)$$

$$E[w_k v_i^T] = 0, \quad \text{for all } k \neq i \quad (2.66)$$

Assume at this point that we have an initial estimate of the process at some point in time t_k , and that this estimate is based on all of our knowledge about the process prior to t_k . This prior (or a priori) estimate will be denoted as \hat{x}_k^- where the "hat" denotes estimate, and the "super minus" is a reminder that this is our best estimate prior to assimilating the measurement at t_k . Also assume that the error covariance matrix associated with \hat{x}_k^- is known, i.e. we define the estimation error to be

$$e_k^- = x_k - \hat{x}_k^- \quad (3.67)$$

and the covariance matrix is

$$P_k^- = E[e_k^- e_k^{-T}] = E[(x_k - \hat{x}_k^-)(x_k - \hat{x}_k^-)^T] \quad (3.68)$$

In many cases, we begin the estimation problem with no prior measurements. Thus, in this case, if the process mean is zero, the initial estimate is zero, and the associated error covariance matrix is just the covariance matrix of x itself.

With the assumption of a prior estimate \hat{x}_k^- , the next step is to use the measurement z_k to improve the prior estimate in accordance with the equation

$$\hat{x}_k = \hat{x}_k^- + K_k(z_k - H_k \hat{x}_k^-) \quad (3.69)$$

where

\hat{x}_k = updated estimate

K_k = blending factor (yet to be determined)

The problem now is to find the particular blending factor K_k that yields an updated estimate that is optimal in some sense. We use minimum mean-square error as the performance criterion. Toward this end we first form the expression for the error covariance matrix associated with the updated (posteriori) estimate.

$$P_k = E[e_k e_k^T] = E[(x_k - \hat{x}_k)(x_k - \hat{x}_k)^T] \quad (3.70)$$

Next, we substitute equation 3.63 into equation 3.69 and then substitute the resulting expression for \hat{x}_k into equation 3.70. The result is

$$P_k = E\{[(x_k - \hat{x}_k^-) - K_k(H_k x_k + v_k - H_k \hat{x}_k^-)][(x_k - \hat{x}_k^-) - K_k(H_k x_k + v_k - H_k \hat{x}_k^-)]^T\} \quad (3.71)$$

Now, performing the indicated expectation and noting the $(x_k - \hat{x}_k^-)$ is the a priori estimation error that is uncorrelated with the measurement error v_k , we have

$$P_k(I - K_k H_k)P_k^- (I - K_k H_k)^T + K_k R_k K_k^T \quad (3.72)$$

Notice here that equation 3.72 is a perfectly general expression for the updated error covariance matrix, and it applies for any gain K_k , suboptimal or otherwise.

Returning to the optimization problem, we wish to find the particular K_k that minimizes the individual terms along the major diagonal of P_k , because these terms represent the estimation error variances for the elements of the state vector being estimated. The optimization can be done a number of ways. Our derivation here will follow the completing-the-square approach, because it does not involve the use of special matrix differentiation formulas which are required in the differential calculus approach. We now temporarily drop the subscripts in our equations in order to avoid unnecessary clutter in derivation. Equation 3.72 is then rewritten without subscripts as

$$P = (I - KH)P^- (I - KH)^T + KRK^T \quad (3.73)$$

This expression for P may now be expanded and terms regrouped to yield

$$P = P^- - \underbrace{KHP^- - P^- H^T K^T}_{\text{Linear in } K} + \underbrace{K(HP^- H^T + R)K^T}_{\text{Quadratic in } K} \quad (3.74)$$

P can be seen to quadratic in K , and we now wish to do the matrix equivalent of "completing the square." First we assume $(HP^- H^T + R)$ to be symmetric and positive definite. It can be written in factored form as SS^T , that is,

$$SS^T = HP^- H^T + R \quad (3.75)$$

Using equation 3.85, the expression for P may now be rewritten as

$$P = P^- - KHP^- - P^- H^T K^T + KSS^T K^T \quad (3.76)$$

Now we complete the square and write P in the form

$$P = P^- + (KS - A)(KS - A)^T - AA^T \quad (3.77)$$

where A does not involve K . If equation 3.77 is expanded and compared term by term with equation 3.76, we see that the following equality must be true:

$$KSA^T + AS^TK^T = KHP^- + P^-H^TK^T \quad (3.78)$$

It is easily verified that if we let A be

$$A = P^-H^T(S^T)^{-1} \quad (3.79)$$

then equation 3.78 is satisfied, and equation 3.77 is equivalent to equation 3.76.

We now note that the first and third terms in equation 3.77 do not involve K . Only the middle term involves K , and it is the product of a matrix and its transpose, which ensures that all terms along the major diagonal will be nonnegative. We wish to minimize the diagonal terms of P . Clearly, then the best we can possibly hope to do is to adjust K to make the middle term of equation 3.77 zero. Thus, we choose K to be such that

$$KS = A \quad (3.80)$$

Or, using equation 3.79, we have

$$\begin{aligned} K &= AS^{-1} \\ &= P^-H^T(S^T)^{-1}S^{-1} \\ &= P^-H^T(SS^T)^{-1} \end{aligned} \quad (3.81)$$

But SS^T is just the factored form of $(HPH^T + R)$, so we have for our final expression for the optimum K (with the subscripts reinserted):

$$K_k = P_k^-H_k^T(H_kP_k^-H_k^T + R_k)^{-1} \quad (3.82)$$

This particular K_k , namely the one that minimizes the mean square estimation error, is called the Kalman gain.

The covariance matrix associated with the optimal estimate may now be computed. Referring to equations 3.73 and 3.74, and reinserting the subscripts, we have

$$P_k = (I - K_kH_k)P_k^-(I - K_kH_k)^T + K_kR_kK_k^T \quad (3.83)$$

$$= P_k^- - K_kH_kP_k^- - P_k^-H_k^TK_k^T + K_k(H_kP_k^-H_k^T + R_k)K_k^T \quad (3.84)$$

Routine substitution of the optimal gain expression, equation 3.82 into equation 3.84 leads to

$$P_k = P_k^- - P_k^- H_k^T (H_k P_k^- H_k^T + R_k)^{-1} H_k P_k^- \quad (3.85)$$

or

$$P_k = P_k^- - K_k (H_k P_k^- H_k^T + R_k) K_k^T \quad (3.86)$$

or

$$P_k = (I - K_k H_k) P_k^- \quad (3.87)$$

Of the three expressions for P_k , the latter one given by equation 3.87 is the simplest, so it is used more frequently than the other. Again, note that equation 3.83 is valid for any gain, suboptimal or other wise, whereas equations 3.85, 3.86, and 3.87 are valid only for the Kalman (optimal) gain.

We now have a means of assimilating the measurement at t_k by the use of equation 3.69 with K_k set equal to the Kalman gain as given by equation 3.82. Note we need \hat{x}_k^- and P_k^- to accomplish this, and we can anticipate a similar need at the next step in order to make optimal use of the measurement z_{k+1} . The updated estimate \hat{x}_k^+ is easily projected ahead via the transition matrix. We are justified in ignoring the contribution of w_k in equation 3.62 because it has zero mean and is uncorrelated with the previous w 's. Thus we have

$$\hat{x}_{k+1}^- = \Phi_k \hat{x}_k \quad (3.88)$$

The error covariance matrix associated with \hat{x}_{k+1}^- is obtained by first forming the expression for the a priori error

$$\begin{aligned} e_{k+1}^- &= x_{k+1} - \hat{x}_{k+1}^- \\ &= (\Phi_k x_k + w_k) - \Phi_k \hat{x}_k \\ &= \Phi_k e_k + w_k \end{aligned} \quad (3.89)$$

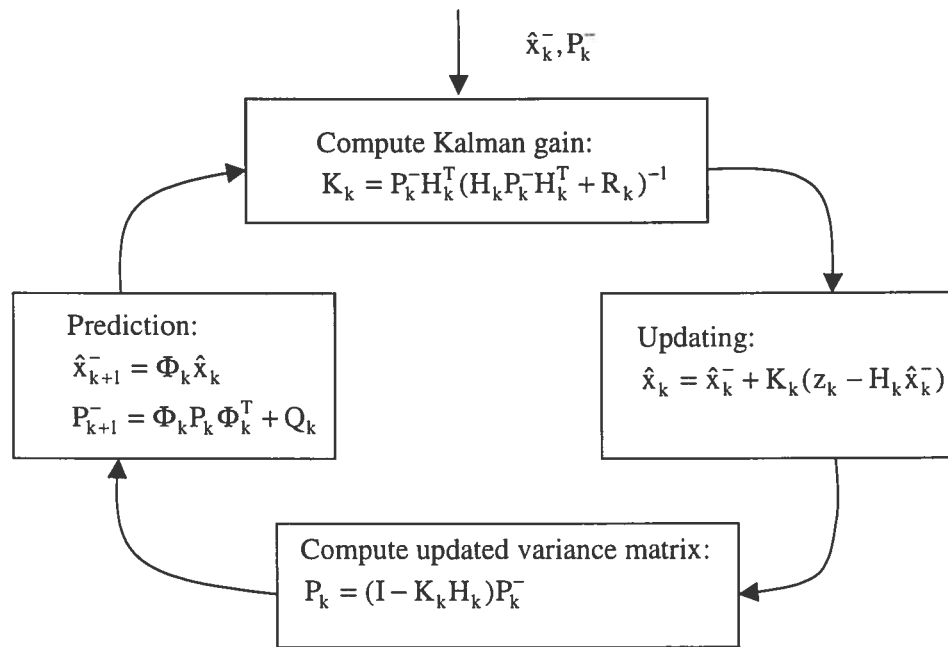
We now note that w_k and e_k are uncorrelated, and thus we can write the expression for P_{k+1}^- as

$$\begin{aligned} P_{k+1}^- &= E[e_{k+1}^- e_{k+1}^{-T}] = E[(\Phi_k e_k + w_k)(\Phi_k e_k + w_k)^T] \\ &= \Phi_k P_k \Phi_k^T + Q_k \end{aligned} \quad (3.90)$$

We now have the needed quantities at time t_{k+1} , and the measurement z_{k+1} can be assimilated just as in the previous step.

Equations 3.69, 3.82, 3.87, 3.88, and 3.90 comprise the Kalman filter recursive equations. It should be clear that once the loop is entered, it could be continued as infinitum.

The pertinent equations and the sequence of computational steps shown in the following figure which summarizes what is now known as the Kalman filter.



3.6 Implementation Aspects of Kalman Filtering

3.6.1 Model Non-Linearity

In general both the kinematic model and the measurement model can be non-linear with the form of

$$\text{Dynamic model:} \quad \dot{x}(t) = f(x(t)) + Gw(t) \quad Q(t) \quad (3.91)$$

$$\text{Measurement model:} \quad y(t) = g(x(t)) + \varepsilon(t) \quad R(t) \quad (3.92)$$

Note the dynamic and measurement noise vectors are still assumed to be a linear additive fashion, which does not generally hold true in practice. A method to deal with the non-linearity in the model is to use a first-order approximation where the model can be expanded in a Taylor series neglecting the higher order terms as follows:

$$\text{Dynamic model:} \quad \dot{x}(t) = f(x^0(t)) + F[x(t) - x^0(t)] + Gw(t) \quad Q(t) \quad (3.93)$$

$$\text{Measurement model:} \quad y(t) = g(x^0(t)) + A[x(t) - x^0(t)] + \varepsilon(t) \quad R(t) \quad (3.94)$$

where

$$F = \left. \frac{\partial f(x(t))}{\partial x(t)} \right|_{x(t)=x^0(t)} \quad (3.96a)$$

$$A = \left. \frac{\partial g(x(t))}{\partial x(t)} \right|_{x(t)=x^0(t)} \quad (3.96b)$$

and equations 3.93 and 3.94 are linear systems.

If a nominal trajectory is available and linearization is done about points of this trajectory, the filter is often called a linearized Kalman filter (LKF). If the linearization is done about the predicted state estimate the filter is called an extended Kalman filter. If the solution is iterated until a certain stopping criterion is met to get the most recent state estimate, the filter is called an iterated extended Kalman filter (IEKF) which has been the most frequently used filter in geomatics applications to deal with non-linearity in the filter model (Salzmann, 1988). If only consider the non-linearity in the measurement model, the iterative measurement update equation has the form of

$$\hat{x}_{k(i+1)} = \hat{x}_k^- + K_{k(i)}(z_k - g(\hat{x}_{k(i)}) - H_{k(i)}(\hat{x}_k^- - \hat{x}_{k(i)})) \quad (3.97)$$

3.6.2 Colored Noise in System Model

As you have seen, the Kalman filter equations are derived based on the assumption that the system noise and measurement noise sequences are white and un-correlated. When system noise and measurement noise sequences are correlated over time, that is,

$$E[w_k w_i^T] = Q_{ki} \neq 0 \quad (3.98)$$

$$E[v_k v_i^T] = R_{ki} \neq 0 \quad (3.99)$$

$$E[w_k v_i^T] \neq 0 \quad (3.100)$$

for all k and i . Such noise sequences are called colored noise sequence. The standard Kalman filter cannot provide optimal solutions without modification of its filtering equations.

State augmentation is the most generally applied technique in this case. Typical example is the method of shaping filter. But it increases the computational load considerably compared to the standard Kalman filter. A set of Kalman filtering equations have been developed by Gao et al. (1992) which is able to consider consecutive correlation between the measurement sequences without the augmentation of the state vector. Consideration of existing correlation in system models can improve the estimation precision and provide more realistic results. Further readings on the topic refer to Maybeck (1979) and Anderson and Moor (1979).

3.6.3 Computational Efficiency

Recall the Kalman filter equations, we should notice that the major computational load is the computations related to several major matrix operations especially the inverse operation which can account for up to 90% of the total computational load in Kalman filtering. There are different methods available with efforts to improve the computational efficiency since the Kalman filter equations given in 3.5 can be algebraically manipulated into a variety of form. An alternative form, called Bayes filtering sometimes, which has merits in certain conditions compared to the standard Kalman filter equations in terms of computational efficiency, will be presented in the following. The derivations are based on Brown (1983).

With the expression for updating the error covariance, equation 3.87,

$$P = (I - KH) P^- \quad (3.101)$$

where the subscripts are omitted to save writing.

As the Kalman gain is given by equation 3.148,

$$K = P^- H^T (H P^- H^T + R)^{-1} \quad (3.102)$$

substituting equation 3.102 into equation 3.101 yields

$$P = P^- - P^-H^T(HP^-H^T + R)^{-1}HP^- \quad (3.103)$$

We now wish to show that if the inverses of P , P^- , and R exist, P^- can be written as

$$P^{-1} = (P^-)^{-1} + H^TR^{-1}H \quad (3.104)$$

which is an alternative expression for the calculation of the updated state variance matrix.

The proof of equation 3.104 can simply form the product of the right sides of equations 3.103 and 3.104 is equal to the identity matrix as shown in the following:

$$\begin{aligned} & P^- - P^-H^T(HP^-H^T + R)^{-1}HP^-][(P^-)^{-1} + H^TR^{-1}H] \\ &= I - R^{-1}H^T[(HP^-H^T + R)^{-1} - R^{-1} + (HP^-H^T + R)^{-1}HP^-H^TR^{-1}]H \\ &= I - P^-H^T[(HP^-H^T + R)^{-1}(I + HP^-H^TR^{-1}) - R^{-1}]H \\ &= I - P^-H^T[R^{-1} - R^{-1}]H \\ &= I \end{aligned}$$

Since $K = R^{-1}H^T(HP^-H^T + R)^{-1}$, insertion of PP^{-1} and $R^{-1}R$ will not alter the gain. Thus, K can be written as

$$\begin{aligned} K &= PP^{-1}P^-H^TR^{-1}(HP^-H^T + R)^{-1} \\ &= PP^{-1}P^-H^TR^{-1}(HP^-H^TR^{-1} + I)^{-1} \end{aligned}$$

Application of equation 3.104 for P^{-1} yields the following equation

$$\begin{aligned} K &= P[(P^-)^{-1} + H^TR^{-1}H]P^-H^TR^{-1}(HP^-H^TR^{-1} + I)^{-1} \\ &= P(I + H^TR^{-1}HP^-)H^TR^{-1}(HP^-H^TR^{-1} + I)^{-1} \\ &= PH^TR^{-1}(I + HP^-H^TR^{-1})(HP^-H^TR^{-1} + I)^{-1} \\ &= PH^TR^{-1} \end{aligned} \quad (3.105)$$

which is an alternative expression for the gain matrix.

Now the equations 3.104 and 3.105 can be rewritten with the subscripts reinserted:

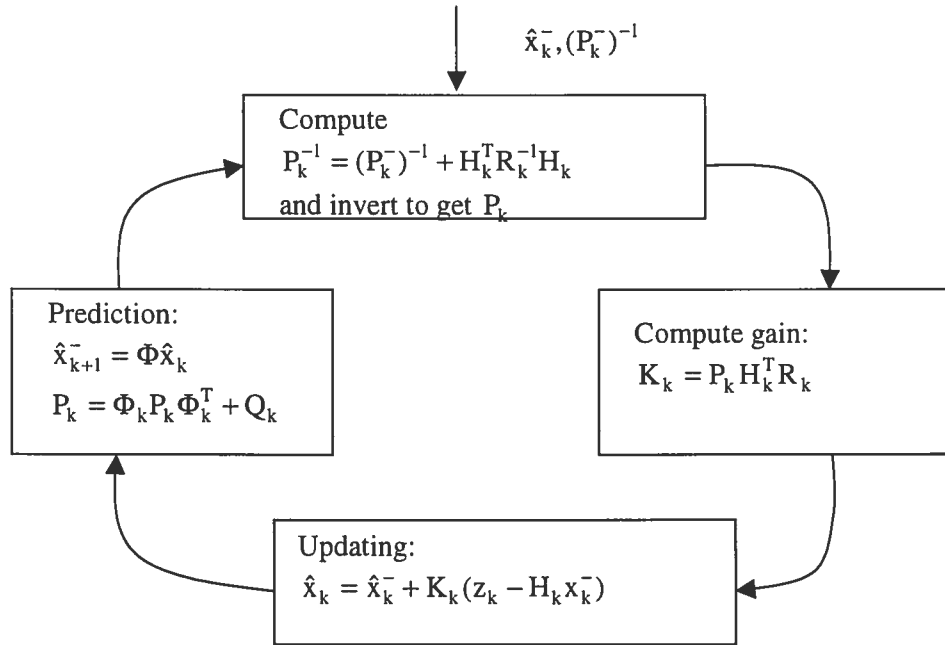
$$P_k^{-1} = (P_k^-)^{-1} + H_k^TR_k^{-1}H_k \quad (3.106)$$

$$K_k = P_k H_k^T R_k^{-1} \quad (3.107)$$

Based on the above alternative equations to first calculate the updated state variance matrix and then the gain matrix, an alternative Kalman filtering implementation has been obtained.

As seen in 3.106, the updated error covariance can now be computed without first finding the gain and the determination of the gain matrix is done after and requires P_k , that is, if equation 3.107 is to be used, K_k must be computed after the P_k computation. In other words, the order in which the P_k and K_k computations based on the alternative expressions is reversed from that presented in 3.5.

The alternative Kalman filter algorithm is summarized in the following figure. Note from the figure that two $(n \times n)$ matrix inversions are required for each recursive loop. If the order of the state vector is large, this leads to obvious computational problems. Nevertheless, the alternative algorithm has some useful applications.



3.6.4 Numerical Stability

As to computational stability problem, equation 3.97 can involve small differences of large numbers, particularly if there exist extremely accurate measurements compared to the rest of other measurements. This could cause

numerical instability problems. Ways to improve the numerical stability of the kalman filter can simple and straightforward through revised matrix operation procedures. Recall the Kalman filter equations, we notice that the major computational load is the computations related to matrix operations especially the inverse operation which can account for up to 90% of the total computational load in Kalman filtering. There are different methods available with efforts to improve the computational efficiency and at the same time to maintain a high computational stability to avoid problem such as divergence, nonnegative positive covariance matrix etc. An equivalent form of equation 3.79 is

$$P_k = (I - A_k K_k) P_k^- (I - A_k K_k)^T + K_k R_k K_k^T \quad (3.110)$$

which can improve the symmetry and positive definiteness of the filtered covariance matrix of the states and it is also insensitive to small errors in the computed filter gain matrix (Maybeck, 1979). The computation load of equation 3.110, however, is considerably greater than that using equation 3.79.

There are other more complicated treatments for the improvement of filter stability. Square root filter and U-D factorization filter are two popular approaches that have found wide applications. Square root filter is based on the decomposition of a nonnegative definiteness covariance matrix into the form of LL^T , where L is normally square, but not necessarily nonnegative definite, and it not unique. The matrix L is also often called a Cholesky factor. The U-D filter decomposes a symmetric covariance matrix into the form of UDU^T , where U is a unitary upper triangular (i.e. with ones along the diagonal) and D is a diagonal matrix. For a detailed description of the methods the reader is referred to Anderson and Moore (1979), Maybeck (1979) and Bierman (1977).

REFERENCES

- Anderson, B.D.O. and J.B. Moore (1979). *Optimal Filtering*. Prentice-Hall, Englewood-Cliffs, NJ.
- Bierman (1977). *Factorization Methods for Discrete Sequential Estimation*. Academic Press, New York.
- Brown, R.G. (1983). *Introduction to Random Signal Analysis and Kalman Filtering*, Jone Wiley & Sons, New York.
- Gao, Y., E.J. Krakiwsky and Z.W. Liu (1992). A New Algorithm for Filtering a Correlated Measurement Sequence, *Manuscripta Geodetica*, Vol. 17, No. 2.

-
- Gao, Y. (1992). *A Robust Quality Control System for GPS Navigation and Kinematic Positioning*, PhD Thesis, Department of Geomatics Engineering, The University of Calgary, Calgary, Alberta, Canada.
- Gelb, A. (editor) (1974). *Applied Optimal Estimation*, MIT Press, Cambridge, MA.
- Krakiwsky, E.J. (1992). *The Method of Least Squares: a Synthesis of Advances*, ENGO629 lecture Notes, Department of Geomatics Engineering, The University of Calgary, Calgary, Alberta, Canada.
- Maybeck (1979). *Stochastic Models, Estimation, and Control*. Vol. 1, Academic Press, New York.
- Salzmann, (1988). *Some Aspects of Kalman Filtering*. Technical Report, Department of Surveying Engineering, University of New Brunswick, Fredericton, N.B., Canada.

4

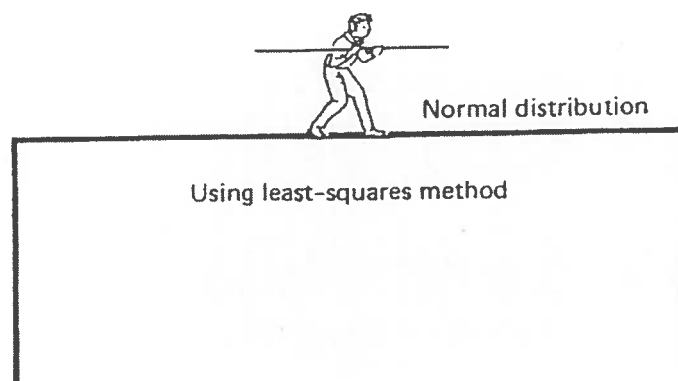
ROBUST ESTIMATION

Outlined in this chapter are the fundamental aspects of robust statistics. The concept behind robust statistics is first introduced. The influence function is then treated along with a discussion of its robust characteristics. Finally, Huber's min-max M-estimator is discussed since it is the most important estimator in robust statistics.

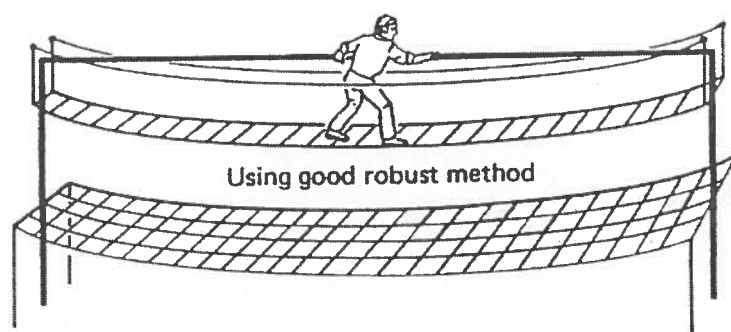
4.1 Concept of Robust Statistics

Robust statistics is concerned with the fact that many assumptions commonly made in classical statistics, such as normality, are at most approximations to reality and that deviations from the assumptions due to blunders, for instance, can falsify results.

The theory has obtained real attention only in the last two decades after the work of Tukey (1960), Huber (1964) and Hampel (1968), although the concept of robust statistics may date back to the beginning of statistics. Tukey (1960) seems to be one of the first statisticians who actually felt that something was wrong with the conventional theory in cases where the normal distribution requirement was not fulfilled. A decisive step in development of the theory for robust statistics, however, was made by Huber (1964). He considered a generalized maximum likelihood estimator and approached the robustness problem by looking for a min-max solution over an approximately normal class of distributions. His paper resulted in a series of papers (Hampel, 1968, 1971; Andrews et al., 1972; Collins, 1976).



(a) Classical statistics



(b) Robust statistics

Figure 4.1: Difference between classical and robust statistics (after Hampel et al., 1986)

Robustness, in general, means insensitivity to deviations from certain assumptions about the parametric model, and thus a robust estimator should be able to provide a good solution even with deviations like blunders. Robust statistics is defined in Hampel et al. (1986) as the statistics of approximate parametric models. Figure 4.1 is a vivid depiction of the difference between classic and robust statistics.

To date, a great variety of approaches have been developed (Huber, 1981; Hampel et al., 1986). The M-estimator is one of the most sophisticated, and more applications can be found with this method in practice (e.g., Masreliez, 1972; Zarembka, 1974; Kubik, 1982; Gao, 1991). In the remainder of this chapter,

important concepts behind the M-estimator are discussed. For a detailed discussion, see Huber (1981) and Hampel et al. (1986).

4.2 Influence Function

Suppose $y_1, y_2, \dots, y_n \in Y$ (sample or observation space) are observations of the unknown parameter $x \in X$ (parameter space) and they are independently, identically distributed with distribution F . Also suppose that the parameter model consists of a family of probability distributions F_x (Hampel et al., 1986). In the following discussion, T denotes a class of estimates and F a class of distributions.

Definition 4.1: The influence function (IF) of a functional or estimator $T \in T$ with a distribution $F \in F$ is given by

$$IF(y; T, F) = \lim_{t \rightarrow 0} \frac{T[(1-t)F + t\Delta_y] - T(F)}{t} \quad (4.1)$$

if the limit exists, where Δ_y is the probability measure which puts mass 1 at the point y .

The above quantity, considered as a function of y , has been introduced by Hampel (1968, 1974) and is perhaps the most useful concept in robust statistics. The influence function describes the (approximate and standardized) effect of observations on the estimator T , given a (large) sample with distribution F . Roughly speaking, the influence function $IF(y; T, f)$ is the first derivative of T for an underlying distribution F . For robustness, the importance of the influence function lies in its heuristic interpretation, namely, it describes the effect of an infinitesimal contamination in the observations on the estimate, standardized by the mass of the contamination. One could say it gives a picture of the infinitesimal behaviour of the asymptotic value, so it measures the asymptotic bias caused by contamination in the observations. In classical statistics, one then may assume that the observations y are distributed exactly as one of the F_x , which is the distribution assumed by the parametric model and undertaken to estimate x based on the data at hand.

In large samples, an important property of an estimator is its asymptotic variance. The asymptotic variance associated with the estimator T can be determined based on the influence function (Hampel et al., 1986)

$$V(T, F) = \int IF(y; T, F)^2 dF(y) \quad (4.2)$$

This integral equation indicates the important relation between the influence function and the asymptotic variance. Therefore, the influence function allows an immediate, and simple, heuristic assessment of the asymptotic properties of the estimate, since it allows us to guess at an explicit formula of equation 4.2 for the asymptotic variance. A Cramer-Rao inequality equation can be established, similar to that of conventional statistics, if the corresponding T is Fisher consistent (Hampel et al. 1986), i.e.,

$$T(Fx) = x, \quad (4.3)$$

which means that in the given model the estimator T asymptotically measures the true value, i.e., the estimator is unbiased. The probability density function of the distribution Fx can be denoted by f_x , and let $F^* = Fx^*$ where x^* is some fixed element of $x \in X$. The Cramer-Rao inequality equation can then be written as

$$\int IF(y; T, F^*)^2 dF^*(y) \geq \frac{1}{J(F^*)} \quad (4.4)$$

where

$$J(F^*) = \int \left\{ \frac{\partial}{\partial x} [\ln f_x(y)]_{x^*} \right\}^2 dF^* \quad (4.5)$$

and $J(F^*)$ is called the Fisher information matrix. This inequality relation provides a lower bound for the asymptotic variance of the estimator T and thus it can be used to evaluate the efficiency of an estimator based on the following equation (Hampel et al., 1986):

$$\text{eff} = \frac{1/J(F^*)}{V(T, F^*)} \quad (4.6)$$

The equality holds if and only if

$$IF(y; T, F^*) \text{ is proportional to } \frac{\partial}{\partial x} [\ln f_x(y)]_{x^*} \quad (4.7)$$

This means that the estimator T is asymptotically efficient if and only if

$$IF(y; T, F^*) = J(F^*)^{-1} \frac{\partial}{\partial x} [\ln f_x(y)]_{x^*} \quad (4.8)$$

In this case, the asymptotic variance of an estimator approaches its lower bound defined by $\frac{1}{J(F^*)}$ with a maximal efficiency of $\text{eff} = 1$.

Consider the well-known arithmetic mean, $T_n = \sum_{i=1}^n y_i$, as an example under the assumption of Gaussian distribution Φ with a probability density function of $\phi(y) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}y^2)$. The corresponding influence function (also see Figure 4.2) becomes (Hampel et al., 1986):

$$IF(y; T, \Phi) = y. \quad (4.9)$$

The associated asymptotic variance and Fisher information matrix with this influence function are both equal to one and thus the efficiency of the estimator is equal to 1 (Hampel et al., 1986). This proves that the arithmetic mean is an optimal estimator in terms of efficiency when the underlying distribution is Gaussian.

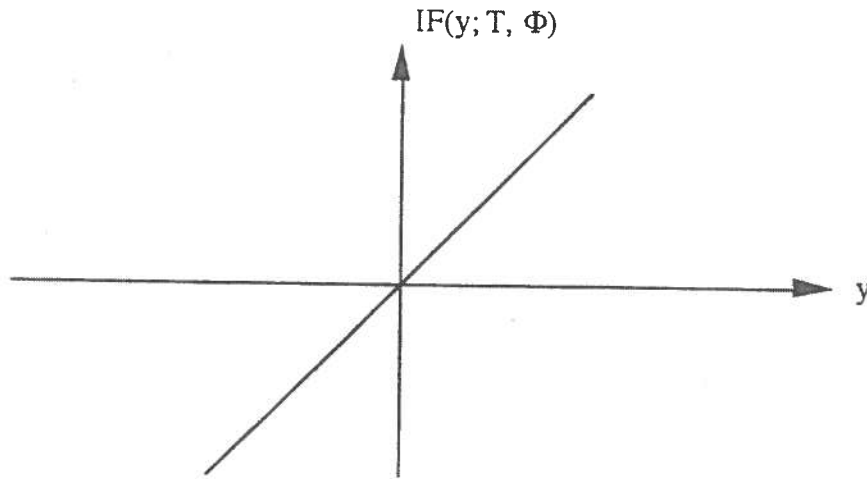


Figure 4.2: Influence Function of the Arithmetic Mean

4.3 Robust Criteria Based on the Influence Function

The influence function describes the effect that an additional observation has on the estimator and thus enables us to construct a variety of robust estimators that satisfy pre-determined conditions related to the estimate. It becomes the basis for quantifying the robustness of an estimator. Robustness measures of an estimator are described below based on the influence function, where the robustness means that a small change in the parametric model should result in only a small change in the estimate (Hampel et al., 1986). For robustness,

we wish to determine the influence function so that the resulting estimator will be able to resist unexpected deviations, e.g., blunders. On the other hand, we desire that the resulting estimate is reasonably good or has only small efficiency degradation in the case that the data actually enjoys the normal assumptions. The efficiency that we sacrifice for robustness in this case is sometimes referred to as the "premium" that we pay to gain protection in very non-normal cases. In the following, three most important robustness measures are discussed which were introduced by Hampel (1968. 1971).

Definition 4.2: An estimator T is said to be robust at F if the y^* defined below is bounded

$$\gamma^*(T, F) = \sup_y |IF(y; T, F)|, \quad (4.10)$$

where the supremum is taken over all y where $IF(y; T, f)$ exists.

$\gamma^*(T, F)$ is often called gross error sensitivity of T at F which measures the worst (approximate) influence that a small amount of contamination of fixed size can cause on the value of the estimator. Therefore, it may be regarded as an upper bound on the (standardized) asymptotic bias of the estimator. This criterion is important in order to quantify the robustness of an estimator since an estimator has robustness of efficiency only if y^* is bounded.

Definition 4.3: The rejection point is given by

$$\rho^* = \inf\{r > 0; IF(y; T, F) = 0 \text{ when } |y| > r\} \quad (4.11)$$

If there exists no such r , then $\rho^* = \infty$ by definition of the infimum.

The rejection point defines a threshold and the influence function vanishes beyond this threshold. In other words, if the influence function is zero in some region, then the contamination in those points does not have any influence at all on the estimates. This quantity is very useful for practical applications in order to eliminate the influence of extremely large errors.

Definition 4.4: Local shift sensitivity is defined by

$$\lambda^* = \sup_{y \neq z} \frac{|IF(z; T, F) - IF(y; T, F)|}{|z - y|} \quad (4.12)$$

where z is some neighbouring point around the observation y and $IF(y; T, F)$, $IF(z; T, F)$, are the corresponding influence functions.

The difference of the influence functions associated with y and z in equation 4.12 are used to describe the effect of shifting an observation slightly from the point y to some neighbouring point z . This quantity A^* then describes the normalized worst effect on an estimator due to the shift in the observation.

The robustness measures defined above are based on the influence function and thus they are entirely local concepts. The concept of the breakdown point defined below, however, is a global robustness measure which can be applied to complement the above local robustness measures in the robustness analysis of an estimator.

Definition 4.5: The breakdown point of an estimator is defined by

$$\varepsilon^* = \sup\{\varepsilon \leq 1; \text{there is a compact set } K_\varepsilon \subset X \text{ such that } \pi(F, G) < \varepsilon$$

$$\text{implies } G(\{T_n \in K_\varepsilon\}) \xrightarrow{n \rightarrow \infty} 1\}$$
(4.13)

where $\pi(F, G)$ denotes the Prohorov distance between the distributions F and G (Hampel et al., 1986).

The breakdown point describes the largest possible fraction of the observations for which there is a bound on the change in the estimate when that fraction of the sample is altered without bound. In other words, this quantity describes how much contaminated data an estimator can stand before it becomes useless.

Considering again the example of the arithmetic mean, we now turn to examine its robustness properties using the notions introduced above. From equation 4.9 (also see Figure 4.2), it is seen that the gross-sensitivity y^* in equation 4.10 is equal to infinity. In addition, its rejection point p^* defined by equation 4.11 is also infinite. The local-shift sensitivity A^* in equation 4.12 is finite and equal to 1, but it may be the lowest value of A^* possible for Fisher-consistent estimators (Hampel et al., 1986). Thus, the arithmetic mean is not a robust estimator which does not have protection against possible deviations such as blunders in the observation. Furthermore, the breakdown point of the arithmetic mean E^* is equal to zero and the estimator is also not robust in a global sense.

From the discussion above, we realize that a robust estimator should have an influence function that is bounded (finite gross error sensitivity), is moderately continuous (finite local shift sensitivity), and has a finite rejection point, and the breakdown point of the estimator should be as large as possible.

These criteria can then be used to formulate the optimal influence function and subsequently to define the robustness of an estimator.

4.4 Min-Max M-Estimator

Let the parametric model under consideration be described by the following linear system in matrix form, namely:

$$y = Ax + \varepsilon, \quad (4.14)$$

or for individual observations

$$y_i = a_i x + \varepsilon_i \quad i = 1, 2, \dots, n. \quad (4.15)$$

where y is an n -dimensional vector of observations that are assumed to be independently and identically distributed with distribution $F \in \mathcal{F}$, x is an m -dimensional vector of unknown parameters, ε is the error vector with dimension of n , and a_i is the i -th row vector of the design matrix $A = (a_{ij})$, i.e.,

$$a_i = (a_{i1}, a_{i2}, \dots, a_{im}), \quad i = 1, 2, \dots, n. \quad (4.16)$$

The distribution F_x of the observations for the linear model is usually assumed to be a normal distribution. The well-known maximum likelihood estimator (MLE) is thus defined as the value $T_n = T(y_1, y_2, \dots, y_n)$ that maximizes $\prod_{i=1}^n f_{T_n}(y_i - a_i x)$, or equivalently by

$$\sum_{i=1}^n -\ln f_{T_n}(y_i - a_i T_n) = \text{Minimum}_{T_n}, \quad (4.17)$$

where \ln denotes the natural logarithm. This estimator has many well-defined and valuable statistical properties under the assumption that F_x is a normal distribution, and has been widely applied. In robust statistics, Huber (1964) proposed to generalize this to

$$\sum_{i=1}^n \rho(y_i, T_n) = \text{Minimum}_{T_n}, \quad (4.18)$$

where ρ is some given function, usually a non-negative symmetric function. We see that the regular maximum likelihood estimator given in equation 4.17 becomes a special case of the estimator defined by equation 4.18. Suppose that P

has a derivative $\psi(y, x) = \frac{\partial}{\partial x} \rho(y, x)$, so that the estimate T_n of equation 4.18 satisfies the following implicit equation

$$\sum_{i=1}^n \psi(y_i, T_n) = 0. \quad (4.19)$$

This equation provides an alternative way to calculate the solution of the extreme problem in equation 4.18 by solving a set of equations.

Definition 4.6: Given the linear model in equation 4.14, any estimator defined by equation 4.18 or equation 4.19 is called an M-estimator.

The name "M-estimator" comes from "Generalized Maximum Likelihood" (Huber, 1964) since it reduces to a usual maximum likelihood estimator when a special p-function is chosen.

Equations 4.18 and 4.19 are not always equivalent, but usually the latter equation is very useful in the search for the solution to the former equation. Let the solution T_n of equation 4.18 correspond to the functional $T(F)$ given by (considering a one-dimensional case for simplicity)

$$\int \psi(y, T(F)) dF(y) = 0, \quad (4.20)$$

for all distributions F for which the integral is defined. The expression of the influence function can then be rewritten as (Hampel et al., 1986)

$$IF(y; T, F) = \frac{\psi(y, T(F))}{-\int \frac{\partial}{\partial x} [\psi(z, x)] T(F) dF(z)}, \quad (4.21)$$

under the assumption that the denominator is not equal to zero. This equation describes the relation between the influence function and the ψ -function. The importance of equation 4.21 is that it tells us that the influence function is proportional to ψ -function by a constant. In other words, given an influence function, we can find a 'V-function' which is a constant multiple of it. Therefore, ψ -function may be used itself as the influence function and can be equivalently used to investigate the robustness properties of an estimator.

The variance matrix of the estimator can be calculated by

$$V(T, F) = \frac{\int [\psi(y, t(F))]^2 dF(y)}{\left\{ \int \frac{\partial}{\partial x} [\psi(z, x)] T(F) dF(z) \right\}^2} . \quad (4.22)$$

The development of the robust M-estimator has centred around efficiency and robustness; both are important properties for an estimator. Different choices of ψ -function will result in estimators with different properties in terms of efficiency and robustness. The optimization of the robust criteria and the efficiency of the estimator, however, cannot be both achieved simultaneously (Hampel et al., 1986). Thus, it is important to maximize the efficiency along with the given robustness criteria which results in Huber's min-max problem. With Huber's idea, quantitative evaluation of robustness must somehow be concerned with the maximum degradation of performance possible for a deviation from the underlying assumptions (Huber, 1964). An optimally robust procedure should then be one which minimizes this maximum degradation and will be a min-max procedure of some kind. If we use asymptotic performance criteria (like asymptotic variances), we then obtain asymptotic min-max estimates, the well-known Huber's min-max M-estimator, which is presented immediately below.

Definition 4.7: The following extreme problem

$$\text{min-max } V(T, F)$$

$$1) T^* \text{ minimizes } V(T, F^*) \text{ over all } T \in T,$$

$$2) F^* \text{ maximizes } V(T^*, F) \text{ overall } F \in F, \quad (4.23)$$

is called Huber's min-max problem, where T^* , F^* denote the solution of the min-max problem.

Huber's min-max problem actually is concerned with minimizing the maximal asymptotic variance over F and finding the M-estimator T^* satisfying

$$\sup_{F \in F} V(T^*, F) = \text{minimum}_T \sup_{F \in F} V(T, F) . \quad (4.24)$$

The corresponding optimal robust influence function ψ^* , chosen in accordance with the min-max principle, should satisfy the saddle point condition

$$V(\psi(F^*)) \geq V(\psi^*, F^*) \geq V(\psi^*, F) . \quad (4.25)$$

The corresponding estimator T^* is often referred to as the min-max robust estimator. In terms of both efficiency and robustness, an efficiency robust estimate should be one whose efficiency is high at a variety of distributions.

Huber has studied the min-max M-estimator T^* in equation 4.23 and equation 4.24 based on the family of E-contaminated distributions, which is defined as follows:

$$F_\varepsilon = \{F : F = (1 - \varepsilon)G + \varepsilon H, \varepsilon \text{ fixed and } 0 \leq \varepsilon \leq 1, H \text{ symmetric}\}, \quad (4.26)$$

where G is some known symmetric distribution (basic mode) and H is an unknown symmetric distribution (exceptional mode). In practice, F is usually unknown but is believed to lie in some appropriate neighbourhood of the known distribution G , e.g., Gaussian distribution which is often implied in classic statistics. The introduction of H is to account for the fact that a fraction of data may consist of blunders. With respect to the E-contaminated distribution, Huber (1964) solved the min-max problem by finding the well-known least favourable distribution F_0 , i.e., the distribution minimizing the Fisher information matrix $J(F)$ over all $F \in F_\varepsilon$. The least favourable distribution represents the worst possible distribution over the neighbourhood of the assumed mode. The resulting estimator is the asymptotically most efficient M-estimator, i.e., a maximum likelihood estimator for the ε -contaminated distributions F_ε . The optimal robust influence function ψ_0 corresponding to the least favourable distribution F_0 , chosen in accordance with the min-max principle, equals

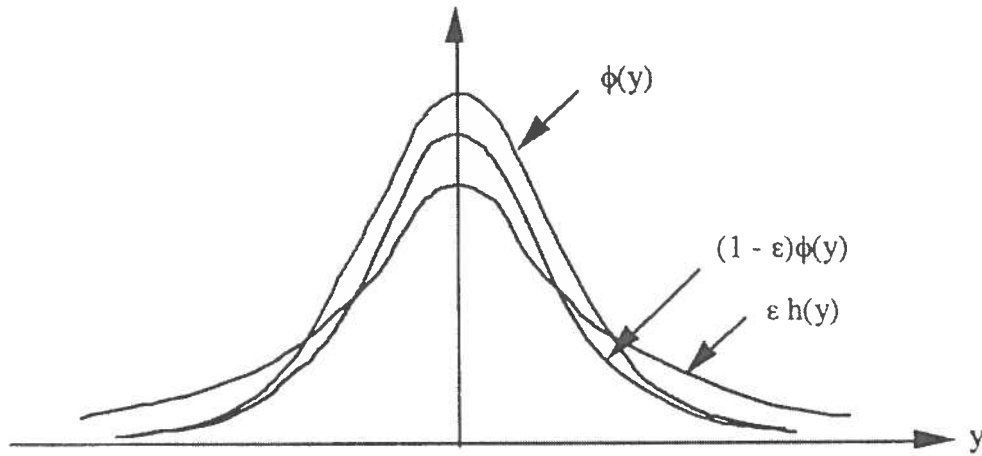
$$\psi_0 = -\frac{F_0''}{F_0'} = -\frac{f_0'}{f_0} \quad (4.27)$$

which satisfies the saddle point condition of equation 4.25. F_0'' and F_0' are the first and second derivatives of F_0 , respectively. f_0 is the probability density function of distribution F_0 and f_0' is the corresponding derivative.

When the ψ_0 is chosen skew-symmetric and strictly monotone, then all M-estimators will be min-max M-estimator and satisfy the following properties (Hampel, 1971; Huber, 1981):

- 1) They are robust and have breakdown point $\varepsilon^*=0.5$, which is the maximal value that a breakdown point can have, if ψ_0 is bounded;
- 2) They are not robust, and the breakdown point $\varepsilon^*=0$, if ψ_0 is unbounded.

When G in equation 4.26 is a Gaussian distribution Φ (with probability density function ϕ), F becomes the ε -contaminated Gaussian distribution (also see Figure 4.3).

Figure 4.3: ε -Contaminated Gaussian Distribution

Further, the well-known normal mixture distribution can be obtained, whose probability density function, denoted as $CN(. | \varepsilon, R_1, R_2)$, equals

$$CN(. | \varepsilon, R_1, R_2) = (1 - \varepsilon)N(. | 0, R_1) + \varepsilon N(. | 0, R_2) , \quad (4.28)$$

where $N(. | 0, R)$ denotes the normal density with zero mean and covariance matrix R . Usually, R_2 is much larger than R_1 and it is used to account for the contamination to the ideal mode.

The least favourable distribution for the family of ε -contaminated Gaussian distributions has been given by Huber (1964), whose probability density function is normal in the middle and exponential in the tails. The corresponding influence function is (also see Figure 4.4(a))

$$\psi_0(v) = \begin{cases} v & |v| \leq k \\ k \operatorname{sign}(v) & |v| > k \end{cases} \quad (4.29)$$

where k is a constant depending on ε . The corresponding estimator is often called Huber's M-estimator that, in terms of efficiency, does the best possible for the family of ε -contaminated distributions and is highly efficient for distributions close to Gaussian. See Huber (1964) for the details of this development. Other influence functions have also been proposed by different authors. Shown in Figures 4.4(b) and (c) are Hampel's and Andrew's influence functions applied to different applications.

The contamination distribution H has been assumed being symmetric in the previous discussion. It is, however, worthwhile to mention that Collins (1976)

has investigated the optimal influence function when the contamination is asymmetric. A detailed discussion is given in his original paper (Collins, 1976).

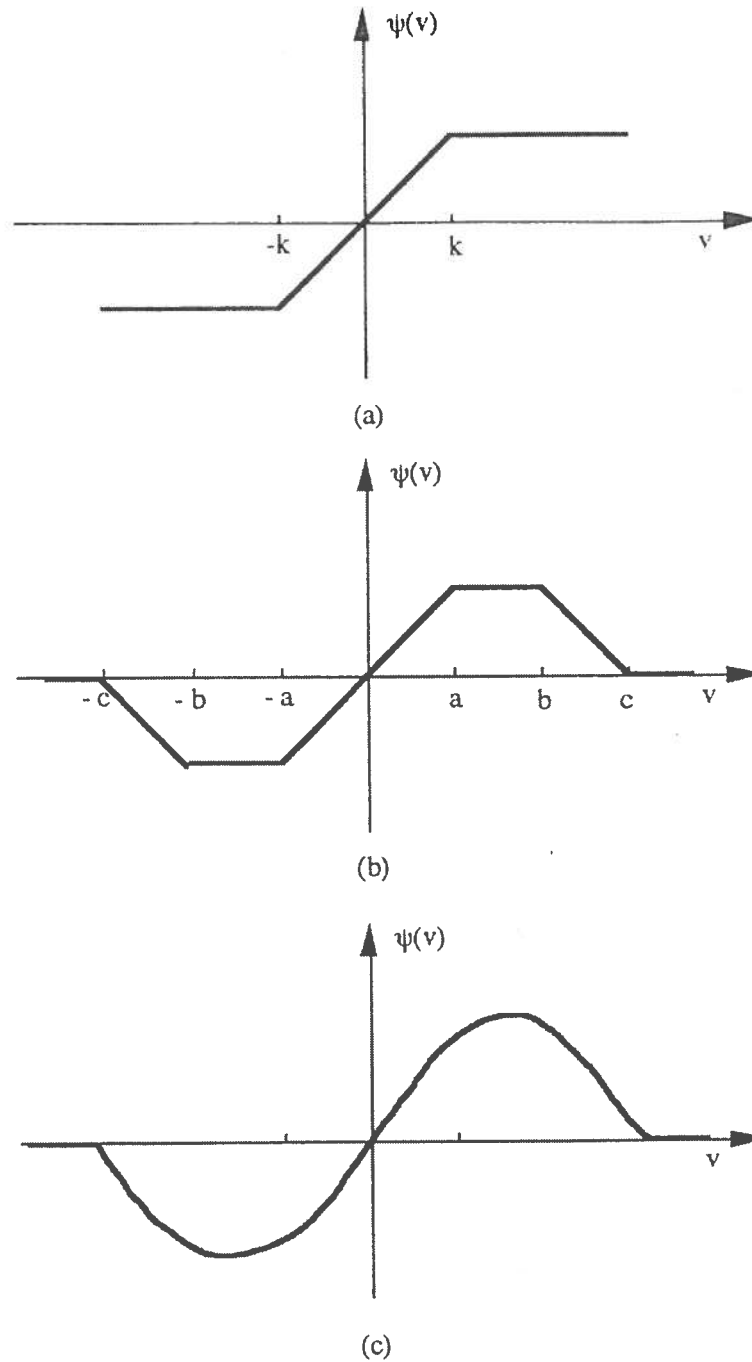


Figure 4.4: Influence Functions: (a) Huber's (b) Hampel's (c) Andrew's

4.5 Robust Least Squares

Conventional least squares estimator implicitly assumes that the observations are normally distributed, which yields a minimum variance unbiased parameter estimate. But numerous studies clearly show that least squares is not a robust estimator which is vulnerable to the larger errors particularly undetected blunders that could significantly distort the obtained solutions. Sometimes even only with a single blunder, the least squares estimate could be greatly away from its true solution.

The standard least squares method can be robustified using the robust estimation concepts introduced in previous sections. In the following, a robustified least squares estimation algorithm is described based on the influence function concept.

Let the parametric model under consideration be described by the following linear system in matrix form, namely:

$$y = Ax + \varepsilon, \quad (4.30)$$

where y is an n -dimensional vector of observations that are assumed to be independently and identically distributed with distribution $F \in \mathcal{F}$, x is an m -dimensional vector of unknown parameters, ε is the error vector with dimension of n , and A is the design matrix. If the observations are correlated and different in accuracy, they can still be converted into the form of equation 4.30.

A way to robustify the least squares estimation process is to introduce a weight function to control the influence of each observation on the estimate. Ideally the observations with blunders should be assigned with a weight as small as possible so that they wouldn't affect the resultant estimate.

The robustified least squares problem is as follows:

$$\left. \begin{array}{l} \sum_{i=1}^n p_i v_i^2 = \min \\ \text{subject to } 1) \ y = Ax + \varepsilon \\ \quad \quad \quad 2) \ p_i \text{ is determined by a robust weight function} \end{array} \right\} \quad (4.31a)$$

and its solution is

$$\hat{x} = (A^T P A)^{-1} A^T P y \quad (4.31b)$$

where $P = \text{diag}(p_1, p_2, \dots, p_n)$.

Since blunders are unknown before estimation, the estimation needs to start from the standard least squares estimation given $P = I$, and a new weight matrix will be determined according to the size of the least squares residuals to repeat the estimation process until the solution converges. There are different robust weight functions available to robustify the least squares estimation and given below are two examples.

$$p_i = \frac{1}{|v_i| + c} \quad \text{where } 0 < c \ll 1 \quad (4.32a)$$

$$p_i = \frac{\hat{\sigma}_0^2}{\hat{\sigma}_i^2} \quad \text{where } \hat{\sigma}_i^2 = \frac{v_i^2}{r_i} \quad \hat{\sigma}_0^2 = \frac{\sum_{i=1}^n v_i^2}{\sum_{i=1}^n r_i} \quad r_i = [I - A(A^T P A)^{-1} A^T P]_{ii} \quad (4.32b)$$

$$p_i = \begin{cases} 1/w_i^2 & w_i \geq 4.13 \\ 1 & w_i < 4.13 \end{cases} \quad \text{where } w_i = \frac{|v_i|}{\hat{\sigma}_0 \sqrt{[P^{-1} - A(A^T P A)^{-1} A]_{ii}}} \quad (4.32c)$$

$$\hat{\sigma}_0^2 = \frac{\sum_{i=1}^n v_i^2}{\sum_{i=1}^n r_i} \quad r_i = [I - A(A^T P A)^{-1} A^T P]_{ii}$$

The influence functions given in Figure 4.4 can also be used to determine robust weight functions via a differentiation process for least squares estimation. For instance, differencing the Huber's influence function of equation 4.29 results in the following weight function:

$$p_i = \begin{cases} 1 & |v_i| \leq k \\ 0 & |v_i| > k \end{cases} \quad (4.33)$$

The success of the above algorithms relies on the assumption that observations with larger residuals more likely contain blunders. In Chapter 6, however, we will find out that such assumption is necessarily holds true in reality since many studies have shown that clean observations may receive larger residuals in least squares estimation and as a result they are assigned with small weights while observations with blunders are given large weights. This would result in very poor solutions. Statistical testing directly based on the standard least squares method may falsify the testing results due to a so-called smoothing effects on the least squares residuals. This would be discussed in Chapter 6.

4.6 Robust Kalman Filtering

Kalman filtering is an efficient algorithm for optimally filtering a Gaussian process. The recursive nature of the algorithm has made it computationally attractive. The non-robustness of the Kalman filtering algorithm, however, was realized in applications where discrepancies from assumed model may lead to significant performance degradation and result in unacceptable solutions. In particular, the usefulness of the filter may be nullified by the phenomenon known as "divergence". Namely, after an extended period of operation of the filter, the solution and variance of the estimates eventually diverge away from the theoretical values governed by the filter equations. Thus, there appears to be considerable motivation for considering filters that are robustified to perform fairly well in non-Gaussian environments.

Considerable effort has been made toward robustification of the Kalman filtering algorithm and most of the contributions in this area have been focused on approximating the pertinent probability densities directly, such as the well-known Gaussian sum method proposed by Alspach et al. (1971). Nevertheless, these methods are not attractive for practical applications because of their computational complexity. For instance, the terms in Gaussian sums approximation increase exponentially as the filter propagates. Therefore, an effort is necessary to reduce the computational load in order to apply robust methods to practical applications.

Masreliez (1972) developed another promising way to tackle this problem. He found that the score function for the innovation sequence plays an important role in obtaining the minimum variance estimator. The theory of robust statistics (Huber, 1964) was applied to formulate a robustified Kalman filter. Compared to above method, the major merit to this approach is the considerable savings in computational time, which is critical for real-time applications. On the other hand, the robustified Kalman filter is very similar in form to the general Kalman filter and thus can be easily implemented into the existing Kalman filtering algorithm with only minor modification. These advantages are quite attractive with respect to the criteria defined for a robust quality control system. In the following, Masreliez's approach is discussed and is then applied to the robust quality control system.

Definition 4.8: An n -dimensional distribution F is said to have properties P1 and P2 if F satisfies the following two properties:

- P1): $f(y_1, y_2, \dots, y_i, \dots, y_n) = f(y_1, y_2, \dots, -y_i, \dots, y_n)$ for each $i = 1, 2, \dots, n$, where $f(\cdot)$ is the probability density function associated with F .

P2): All marginal distributions for F are members of Gaussian mixture distributions $F\epsilon$.

Theorem 4.1: Consider the Kalman filtering models given by equations (3.61a) and (3.61b) and assume that the predicted state vector is Gaussian but the measurement sequences are heavy-tailed non-Gaussian. Suppose that T_k is a linear transformation which insures the distribution of the transformed innovation sequence $v_k^{(-)} = T_k v_k^{(-)} = T_k (z_k - A_k x_k^{(-)})$ to satisfy properties P1 and P2 defined above. Then a min-max Kalman filter can be obtained as follows:

Prediction (time update):

$$x_k^{(-)} = \Phi_{k,k-1} x_{k-1}^{(+)} \quad (4.34)$$

$$C_{xk}^{(-)} = \Phi_{k,k-1} C_{xk-1}^{(+)} \Phi_{k,k-1}^T + Q_k \quad (4.35)$$

Filtering (measurement update):

$$x_k^{(+)} = x_k^{(-)} + C_{xk}^{(-)} A_k^T T_k^T \Psi_k \{v_k^{(-)}\} \quad (4.36)$$

$$C_{xk}^{(+)} \leq C_{xk}^{(-)} - C_{xk}^{(-)} A_k^T T_k^T E_{F_0} \{\Psi_k(v_k^{(-)})\} T_k A_k C_{xk}^{(-)} \quad (4.37)$$

where the vector influence function $\Psi_k \{.\}$ has components $\Psi_{ik} \{v_k^{(-)}\} = \psi \{(v_k^{(-)})_i\}$, and $\Psi' \{.\}$ is the corresponding derivative $\psi \{.\}$ is the odd symmetric scalar influence function corresponding to Huber's min-max estimate for $F\epsilon$. $E_{F_0 \{.\}}$ denotes expectation with respect to the least favourable distribution $F_0 \in F\epsilon$.

Proof: Since the predicted state vector $x_k^{(-)}$ is assumed Gaussian. Equation 4.34 and 4.35 can be obtained directly by means of the variance-covariance propagation law. Let x_k denote the true value of the state vector. The variance matrix of the filtering state $x_k^{(+)}$ can be determined by

$$\begin{aligned} C_{xk}^{(+)} &= E\{(x_k^{(+)} - x_k)(x_k^{(+)} - x_k)^T\} \\ &= E\{[x_k^{(-)} - x_k + C_{xk}^{(-)} A_k^T T_k^T \Psi_k(v_k^{(-)})][x_k^{(-)} - x_k + C_{xk}^{(-)} A_k^T T_k^T \Psi_k(v_k^{(-)})]^T\} \end{aligned} \quad (4.38)$$

Let $x_k' = x_k^{(-)} - x_k$. Then the above equation can be written as

$$C_{xk}^{(+)} = E\{(x_k')(x_k')^T\} + E\{C_{xk}^{(-)} A_k^T T_k^T \Psi_k(v_k^{(-)})(x_k')^T\}$$

$$\begin{aligned}
& + E\{(x_k') \Psi_k(v_k^{(-)})^T T_k A_k C_{xk}^{(-)}\} \\
& + E\{C_{xk}^{(-)} A_k^T T_k^T \Psi_k(v_k^{(-)}) \Psi_k(v_k^{(-)})^T T_k A_k C_{xk}^{(-)}\}
\end{aligned} \tag{4.39}$$

The first term is simply $C_{xk}^{(-)}$ and since $x_k^{(-)}$ is Gaussian.

$$\begin{aligned}
v_k^{(-)} &= T_k(z_k - A_k x_k^{(-)}) = T_k z_k - T_k A_k x_k^{(-)} = T_k z_k - T_k A_k (x_k' + x_k) \\
&= T_k z_k - T_k A_k x_k - T_k A_k x_k' \\
&= T_k(z_k - A_k x_k) - T_k A_k x_k'
\end{aligned} \tag{4.40}$$

Let $v_k^{(-)} = T_k(z_k - A_k x_k^{(-)})$ which is the true transformed innovation sequence and thus is a Gaussian sequence. The second term in equation 4.39 can then be expressed as

$$\begin{aligned}
& E\{C_{xk}^{(-)} A_k^T T_k^T \Psi_k(v_k^{(-)}) (x_k')^T\} \\
&= \int_{R^{m_k}} \int_{R^{n_k}} C_{xk}^{(-)} A_k^T T_k^T \Psi_k(u_k - T_k A_k x_k') (x_k')^T N(x_k' | 0, C_{xk}^{(-)}) dx_k' dF_{u_k}(u_k)
\end{aligned} \tag{4.41}$$

Because

$$x_k' N(x_k' | 0, C_{xk}^{(-)}) = -C_{xk}^{(-)} \frac{d}{dx_k'} N(x_k' | 0, C_{xk}^{(-)}) \tag{4.42}$$

integrating by parts we get

$$\begin{aligned}
& E\{C_{xk}^{(-)} A_k^T T_k^T \Psi_k(v_k^{(-)}) (x_k')^T\} \\
&= - \int_{R^{m_k}} \int_{R^{n_k}} C_{xk}^{(-)} A_k^T T_k^T \Psi_k(u_k - T_k A_k x_k') (x_k')^T \\
&\quad \left\{ \frac{d}{dx_k'} N(x_k' | 0, C_{xk}^{(-)}) \right\}^T C_{xk}^{(-)} dx_k' dF_{u_k}(u_k) \\
&= -C_{xk}^{(-)} A_k^T T_k^T G_k' T_k A_k C_{xk}^{(-)}
\end{aligned} \tag{4.43}$$

where G_k' is a diagonal matrix with elements $E(\Psi_{ik}' \{v_k^{(-)}\}) = E(\psi' \{(v_k^{(-)})_i\})$. Since the transformed innovation sequence satisfies property PI and the influence function ψ is odd, we have

$$C_{xk}^{(+)} = C_{xk}^{(-)} - C_{xk}^{(-)} A_k^T T_k^T (2G_k^{-1} - G_k^{-2}) T_k A_k C_{xk}^{(-)} \quad (4.44)$$

where $G_k^2 = E\{\Psi_k(v_k^{(-)})\Psi(v_k^{(-)})^T\}$ which has elements $E\{\Psi_{ik}^2(v_k^{(-)})\} = E\{\psi^2(v_k^{(-)})_i\}$.

Further, since

$$V(\psi, F) \leq V(\psi, F_0) \quad (4.45)$$

with the defined influence function ψ , we then have

$$2E_F(\psi'(v_k^{(-)})_i) - E_F\{\psi^2\{(v_k^{(-)})_i\}\} \geq E_{F_0}(\psi'\{(v_k^{(-)})_i\}) \quad (4.46)$$

and applying the above inequality into equation 4.44 yields

$$C_{xk}^{(+)} \leq C_{xk}^{(-)} - C_{xk}^{(-)} A_k^T T_k^T E_{F_0}\{\Psi_k(v_k^{(-)})\} T_k A_k C_{xk}^{(-)} \quad (4.47)$$

This robustified Kalman filter has a form similar to the standard Kalman filter and thus preserves the computational advantage of the standard Kalman filtering algorithm. The estimate from this filter is unbiased because the symmetric property of the transformed innovation sequence and the estimate is optimum from the min-max sense. The estimate is also robust in a local sense due to the bounded variance according to equation 4.37. Further, the estimate is globally robust with a breakdown point about 0.5 since the derived filter is a min-max filter with the defined influence function. Masreliez et al. (1977) have also shown that this bound is tight for the family of Gaussian mixture distributions and thus the efficiency of the estimator can be assured.

Shown in Figure 4.2 is the computation flowchart of Masreliez's robust filter. We see from the figure that computationally a linear transformation matrix T_k first needs to be determined with this filter, then is applied to symmetrize and scale the density of the innovation sequence, and finally an influence function is applied to the results to cut off the outliers in the noise distribution. The following Lemma given by Masreliez et al. (1977) assures the existence of this transformation whenever the innovation sequence $v_k^{(-)}$ has a normal mixture distribution.

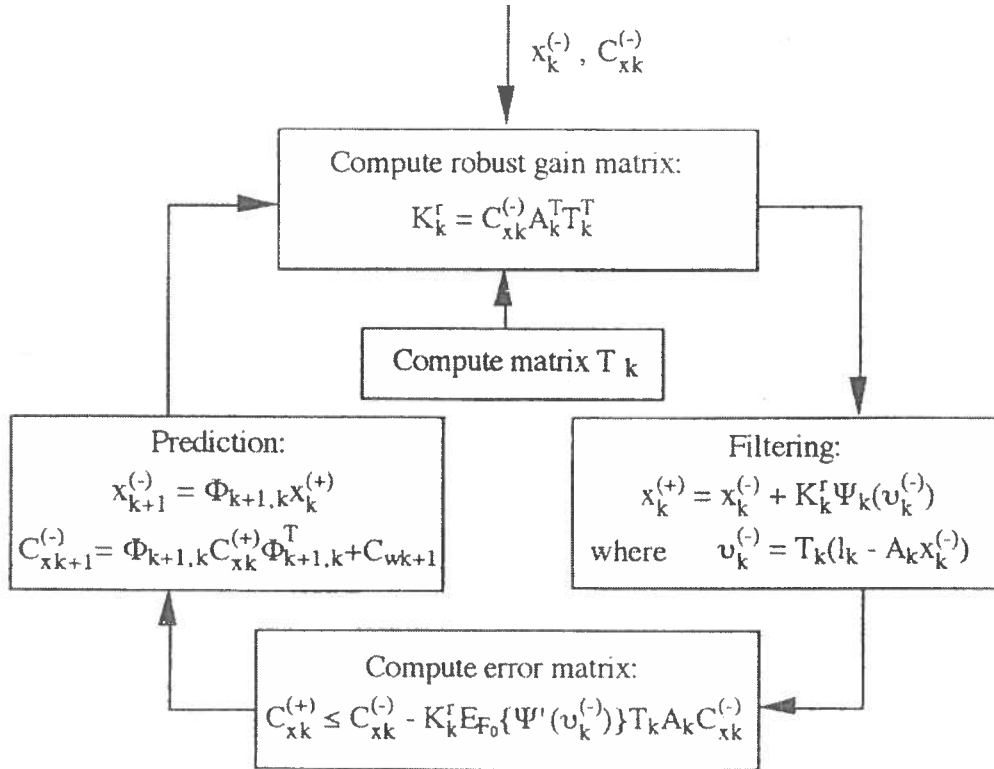


Figure 4.2: Computation flowchart of Masreliez's Robust Filter

Lemma: Let $s = x_1 + x_2$ with x_1, x_2 independent and $x_1 \sim N(0, R_0)$. If $x_2 \sim CN(0, \epsilon, R_1, R_2)$ with $R_1 > 0$, then T exists such that Ts satisfies P1 and P2.

Proof: First suppose that $x_2 \sim CN(0, \epsilon, R_1, R_2)$. Then $s \sim CN(0, \epsilon, R_1', R_2')$ with $R_i' = R_0 + R_i$, $i = 1, 2$. Let T be the linear transformation matrix which simultaneously diagonalizes R_1' and R_1 . T may be written S_1, S_2 with being rotations and a diagonal scaling matrix. S_1, S_2 may be chosen so that $TR_1T^T = I$, which ensures that Ts satisfies P1 and P2 with F_e being a Gaussian mixture distribution.

From the proof, we found that the essence of T with respect to P1 and P2 typically consists of the simultaneous diagonalization of two matrices, i.e., R_1' and R_2' along with a scaling process. Applying the Lemma to the robust Kalman filtering algorithm, let $x_k^{(-)} \sim N(0, C_{xk}^{(-)})$, $l_k \sim CN(0, \epsilon, R_k, R_k')$. Then $v_k^{(-)} \sim CN(0, \epsilon, A_k C_{xk}^{(-)} A_k^T + R_k, A_k C_{xk}^{(-)} A_k^T + R_k')$ and the simultaneous

diagonalization of the matrices of $A_k C_{xk}^{(-)} A_k^T + R_k$ and $A_k C_{xk}^{(-)} A_k^T + R_k'$ can be accomplished once C_{lk}' are specified. The additional scaling of the marginal distributions which will be required, in order that P2 be satisfied, can be determined by the nominal Gaussian covariance matrix R_k .

With the solution from equation (4.36) in Masreliez's robustified Kalman filter, a robust FSG function can be obtained ~ follows:

$$\begin{aligned}\hat{V}_{kr} &= z_k - A_k g(z_k, x_k^{(-)}, R_k, C_{xk}^{(-)}, A_k) \\ &= z_k - A_k x_k^{(-)} - A_k C_{xk}^{(-)} A_k^T T_k^T \psi_k \{v_k^{(-)}\} \\ &= v_k^{(-)} - A_k C_{xk}^{(-)} A_k^T T_k^T \psi_k \{v_k^{(-)}\}\end{aligned}\quad (4.48)$$

where $v_k^{(-)} = T_k(z_k - A_k x_k^{(-)})$.

To use the robust Kalman filter, the additional effort against the standard Kalman filter is that a linear transformation matrix has to be determined such that the transformed innovation sequence satisfies the properties of PI and P2. Below, we will study how the computational effort in the determination of the matrix T_k can be minimized. Proofs of the theorems (Hwang, 1984) given below will be omitted except where they are considered necessary.

Theorem 4.2: Let A be an $n \times n$ symmetric matrix. There exists an orthogonal matrix P such that

$$PAP^T = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_n), \quad (4.49)$$

where λ_i is the i -th eigenvalue of the matrix A .

This theorem tells us that any symmetric matrix has an orthogonal matrix P which diagonalizes the original matrix by equation 4.49.

Theorem 4.3: If A and B are two symmetric matrices, then there exist two orthogonal matrices P_1 and P_2 such that $P_1 A P_1^T = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_n)$ and $P_2 B P_2^T = \text{diag}(\gamma_1, \gamma_2, \dots, \gamma_n)$. In general, $P_1 \neq P_2$. But if $AB=BA$, then $P_1 = P_2 = P$ which simultaneously diagonalizes the two matrices.

This theorem shows that the matrix P , which simultaneously diagonalizes the two matrices, can be determined through an orthogonal decomposition of the original matrix A or B whenever the two matrices are exchangeable.

Theorem 4.4: Let A and B be two symmetric matrices and $B > 0$. Then there exists a non-singular matrix S which simultaneously diagonalizes the two matrices with the following forms:

$$SAS^T = \text{diag}(\mu_1, \mu_2, \dots, \mu_n), \text{ and} \quad (4.50)$$

$$SBS^T = I, \quad (4.51)$$

where μ_i are the roots of $|\lambda B - AI| = 0$.

Proof Since $B > 0$, it can be decom~ into

$$B = LL^T, \quad |L| \neq 0 \quad (4.52)$$

Multiplying A by L^{-1} on the left and $(L^T)^{-1}$ on the right yields

$$C = L^{-1}A(L^T)^{-1}. \quad (4.53)$$

Since matrix C is still symmetric, by applying Theorem 4.2 to it, there exists an orthogonal matrix P such that

$$PCP^T = PL^{-1}A(L^T)^{-1}P^T = \text{diag}(\mu_1, \mu_2, \dots, \mu_n), \quad (4.54)$$

where μ_i are the eigenvalues of the matrix $L^{-1}A(L^T)^{-1}$. Let $S = PL^{-1}$, then

$$SAS^T = \text{diag}(\mu_1, \mu_2, \dots, \mu_n), \quad (4.55)$$

$$SBS^T = PL^{-1}B(L^T)^{-1}P^T = PL^{-1}LL^T(L^T)^{-1}P^T = I \quad (4.56)$$

Therefore, the matrix S simultaneously diagonalizes A and B. Further,

$$|\lambda B - A| = |\lambda SBS^T - SAS^T| = |\lambda I - \text{diag}(\mu_1, \mu_2, \dots, \mu_n)| |P|^2 = 0. \quad (4.57)$$

Thus, $\mu_1, \mu_2, \dots, \mu_n$ are also the roots of $|\lambda B - A| = 0$.

These theorems are useful to minimize the effort in the determination of the matrix T_k , which simultaneously diagonalizes two symmetric matrices required in the implementation of Masreliez's robust Kalman filter. Because the matrix $A_k C_{xk}^{(-)} A_k^T + R_k$ is positive definite in Kalman filtering, the matrix T_k always exists in terms of Theorem 4.4, in which two orthogonal decompositions are involved. The determination of matrix T_k becomes even easier when the matrices

$A_k C_{xk}^{(-)} A_k^T + R_k$ and $A_k C_{xk}^{(-)} A_k^T + R_k'$ are exchangeable, in which only one orthogonal matrix decomposition is required.

REFERENCES

- Alspach, D.L. and Sorenson, H.W. (1971). "Recursive Bayesian estimation using Gaussian sums", *Automatica*, Vol. 6.
- Andrew, D.F., P.J. Bickel, F.R. Hampel, P.J. Hunber, W.H. Rogers and J.W. Tukey (1972). *Robust Estimates of Location: Survey and Advances*. Princeton University Press, Princeton, N.J.
- Collins, J.R. (1976). "Robust estimation of a location parameter in the presence of asymmetry", *Ann. Statist.*, Vol. 4, No. 1.
- Gao, Y. (1991). "A new algorithm of receiver autonomous integrity monitoring (RAIM) for GPS navigation", *Proceedings of ION GPS-91*, Albuquerque, NM, September 11-14, 1991.
- Hwang, L. (1984). *Linear Algebra in System and Control*, Academic Press.
- Hampel, F.R. (1968). *Contributions to the Theory of Robust Estimation*. PhD Thesis, University of California, Berkeley.
- Hampel, F.R. (1971). "A general qualitative definition of robustness". *Ann. Math. Statist.* 42
- Hampel, F.R. (1974). "The influence curve and its role in robust estimation". *J. Am. Statist. Assoc.* 69.
- Hampel, F.R., E.M. Ronchetti, P.J. Rousseeuw and W.A. Stahel (1986). *Robust Statistics: the Approach based on Influence Functions*. John Wiley and Sons, New York.
- Huber, P.J. (1964). "Robust Estimation of a Location Parameter". *Ann. Math. Statist.* 35.
- Huber, P.J. (1981). *Robust Statistics*, John Wiley and Sons, New York.
- Kubik, K.K., P. Frederiksen and W. Wang (1982). "Ah, Robust Estimation". *The Australian Journal of Geodesy, Photogrammetry and Surveying*, No. 42.

Masreliez, C.J. (1972). *Robust Recursive Estimation and Filtering*, PhD Thesis, University of Washington.

Masreliez, C.J. and R.D. Martin (1977). "Robust Bayesian Estimation for Linear Model and Robustifying the Kalman Filter", *IEEE Transaction on Automatic Control*. Vol. AC-22, No. 3.

Tukey, J.W. (1960). "A Survey of Sampling from Contaminated Distribution", *In Contributions to Probability and Statistics*, I. Olkin (ed), Stanford University Press, Stanford, California.

Zarembka, P. (ed) (1974). *Frontiers in Econometrics*. Academic Press, New York.

5

COLLOCATION

Least squares collocation equations, proposed by Moritz [1972], are derived fully based on the equations of the standard adjustment combined case that we have already obtained in Chapter 2.

5.1 Collocation Mathematical Model

We begin with the linear explicit model (equation 2.3)

$$\underset{nx1}{l} = \underset{nxu}{A} \underset{ux1}{x} - \underset{nx1}{r}, \quad (5.1)$$

where l is the vector of n observations, r the vector of n residuals, A the design matrix, and x the vector of u unknown parameters. We now depart slightly from our notation to be consistent with Moritz. The above is rewritten in the form,

$$\underset{nx1}{x} = \underset{nxu}{A} \underset{ux1}{X} - \underset{nx1}{n}, \quad (5.2)$$

where x is called the *measurement* and n its *noise*.

Moritz then extends the above model to

$$\underset{nx1}{x} = \underset{nxu}{A} \underset{ux1}{X} + \underset{nx1}{S'} + \underset{nx1}{n}, \quad (5.3)$$

where the newly introduced quantity S' is called the *signal*. Here we can perhaps interpret S' as the **short-coming in the mathematical model**, that is the *inability* of the model to describe completely (exactly) the actual relationship among the measurements (x) and unknown parameters (X). In other words, one could imagine that there is an *overflow* from the model into some sort of additional correction (S') to the observations. Even though this may seem plausible at first sight, it is the author's belief that is not what is intended in collocation. What is

* from Chapter 4 of the 629 Lecture Notes "The method of least squares: a synthesis of advances" by EJ Krakiwsky.

intended is to contribute the signal directly to the observable, thereby stating that the observable (measurement) has two unknown errors - the signal S' and the noise n . We can consider the noise as a measuring error, or resolution capability, and thus **internal** to the instrument. On the other hand, the signal is thought of as being **external** to the instrument and related to the behaviour of the observable in a particular milieu - like deflections of the vertical in the gravity field, electronically measured distances in the *polluted* atmosphere or in the electron charged ionosphere; or gravity anomalies in the gravity field. An important characteristic of a signal is that it is continuous throughout the domain of a particular *milieu* or *field*. One of the requirements of collocation is that the signal has **known** second moments (covariance matrix), even though the first moments (value of the signal) remain as unknowns to be estimated.

The signal and its variance is not new to geodesists. Since the 1960's we have been calculating the standard deviation of the noise and signal for electronically measured distances from the formula

$$\sigma_d = a + b (\text{distance}).$$

In the above, a is the resolution of the instrument (variance of the noise) and b is a constant in parts per million which when multiplied by the distance is nothing else but the variance of the signal. The latter is a measure of the behaviour of the observed distance in the troposphere. It is rather obvious that there is no covariance in the noise but one suspects that there would be covariance between the signal components of different distances since they are measured in the troposphere. Collocation attempts to account for this correlation through a fully populated variance-covariance matrix for the signal, while in the ordinary least squares treatment, the covariances in the signal are first ignored, then the variance of the signal is combined with the variance of the noise to give one variance (σ_d^2), and finally a solution is made for only one correction (residual) for each measured distance.

In collocation, the condition imposed on the signal is that it be random with zero mean. Thus the measurement x is seen to consist of a systematic part AX , and two random parts, S' and n .

In his development, Moritz introduces the quantity

$$Z = S' + n, \tag{5.4}$$

where S' denotes the signal at the **observation points**. S will be reserved to denote the signal at any point in general without observations. These points (p

in number) are called **computation points** and it is at these points that the signal is said to be *predicted*.

After considering the above equation, the main model becomes

$$\mathbf{x} = \mathbf{AX} + \mathbf{Z} \quad (5.5)$$

and

$$\mathbf{Z} = \mathbf{x} - \mathbf{AX} \quad (5.6)$$

represents the random part of observations after subtracting the systematic part \mathbf{AX} .

This stage constitutes the end of formulating the collocation mathematical model and the beginning of applying the conventional least squares methodology used throughout the previous chapters.

5.2 Least Squares Collocation

We now state the collocation least squares problem. Determine the least squares estimate $\hat{\mathbf{x}}$ in equation 5.3 under the condition that

$$\hat{\mathbf{r}}^T \mathbf{C}_*^{-1} \hat{\mathbf{r}} = \text{minimum}, \quad (5.7)$$

where the residual vector, defined to have the nature of *corrections* as in previous chapters,

$$\hat{\mathbf{r}}^T_{1 \times (p+n)} = \begin{pmatrix} -\hat{\mathbf{S}}^T & -\hat{\mathbf{Z}}^T \\ 1 \times p & 1 \times n \end{pmatrix} \quad (5.8)$$

is made up of two parts - the signal at the computation points, and the random part of the observations. The hyper-covariance matrix

$$\mathbf{C}_*^{-1} = \begin{pmatrix} \mathbf{C}_{ss} & \mathbf{C}_{sx} \\ \mathbf{C}_{xs} & \mathbf{C}_{xx} \end{pmatrix}^{-1}, \quad (5.9)$$

where \mathbf{C}_{ss} is the covariance matrix of the signal, \mathbf{C}_{xx} the variance-covariance matrix of the observable. \mathbf{C}_{sx} and \mathbf{C}_{xs} are the covariance matrices between the signal and the observable.

Remember that in collocation the observable has two random parts - the signal \mathbf{S}' and the noise \mathbf{n} (equation 5.4). Accordingly the covariance matrix

$$\begin{aligned}
 \mathbf{C}_{xx} &= \text{COV}(\mathbf{x}, \mathbf{x}) = \mathbf{M}[\mathbf{Z} \mathbf{Z}^T] = \mathbf{M}[(\mathbf{S}' + \mathbf{n})(\mathbf{S}'^T + \mathbf{n}^T)] \\
 &= \mathbf{M}[\mathbf{S}' \mathbf{S}'^T + \mathbf{n} \mathbf{S}'^T + \mathbf{S}' \mathbf{n}^T + \mathbf{n} \mathbf{n}^T] \\
 &= \mathbf{M}[\mathbf{S}' \mathbf{S}'^T] + \mathbf{M}[\mathbf{n} \mathbf{S}'^T] + \mathbf{M}[\mathbf{S}' \mathbf{n}^T] + \mathbf{M}[\mathbf{n} \mathbf{n}^T] \\
 &= \mathbf{C}_{s's'} + \mathbf{C}_{nn} = \mathbf{C} + \mathbf{D},
 \end{aligned} \tag{5.10}^*$$

after one assumes that the measuring error (\mathbf{n}) has no correlation with the signal (\mathbf{S}') at each observation point. Under the assumption

$$\begin{aligned}
 \mathbf{C}_{sx} &= \mathbf{M}[\mathbf{S} \mathbf{Z}^T] = \mathbf{M}[\mathbf{S}(\mathbf{S}' + \mathbf{n})^T] = \mathbf{M}[\mathbf{S} \mathbf{S}'^T] + \mathbf{M}[\mathbf{S} \mathbf{n}^T] \\
 &= \mathbf{M}[\mathbf{S} \mathbf{S}'^T] = \text{COV}(\mathbf{S}, \mathbf{S}') ,
 \end{aligned} \tag{5.11}$$

and

$$\mathbf{C}_{xs} = \text{COV}(\mathbf{S}, \mathbf{S}') \tag{5.12}$$

are pure signal covariances which describe the correlation between the signal components in the domain of the *observation* and *computation* points.

The above minimum is to be found subject to the constraint function

$$\mathbf{A} \hat{\mathbf{X}} + \mathbf{B}^* \hat{\mathbf{r}}^* + \mathbf{w} = 0, \tag{5.13}$$

where \mathbf{A} is the $n \times u$ design matrix of equation 5.3; \mathbf{X} the u vector of unknown parameters of equation 5.3; \mathbf{r}^* is the $n+p$ vector of equation 5.8; and the newly introduced quantities

$$\mathbf{B}^*_{n \times (n+p)} = \begin{pmatrix} \mathbf{0} & \mathbf{I} & -\mathbf{I} \\ n \times p & & n \times n \end{pmatrix} \tag{5.14}$$

is the second design matrix consisting of a null and minus identity matrix, and

$$\mathbf{w}_{n \times 1} = -\mathbf{x}_{n \times 1} \tag{5.15}$$

* M stands for mathematical expectation - sometimes denoted by E.

is the vector of measurements. In the case that we choose to solve for corrections to some approximate (or observed) parameters, then

$$\mathbf{w}^o = \mathbf{A}\mathbf{X}^o - \mathbf{x}, \quad (5.16)$$

or

$$\mathbf{w} = \mathbf{A}\mathbf{X} - \mathbf{x}. \quad (5.16a)$$

The purpose of introducing the null matrix in the above was to involve the signals \mathbf{S} (to be predicted) in the equations without modifying the original mathematical model given by equation 5.5. To corroborate this, we substitute \mathbf{B}^* , \mathbf{r}^* , and \mathbf{w} into equation 5.13:

$$\mathbf{A}\hat{\mathbf{X}} + \begin{bmatrix} 0 & \mathbf{I} \end{bmatrix} \begin{pmatrix} -\hat{\mathbf{S}} \\ \vdots \\ -\hat{\mathbf{Z}} \end{pmatrix} - \mathbf{x} = 0 \quad (5.17)$$

$$\mathbf{A}\hat{\mathbf{X}} + \hat{\mathbf{Z}} = \mathbf{x} \quad (5.18)$$

which is equation 5.5.

5.3 Least Squares Collocation Equations

The least squares normal equations relating the unknown quantities \mathbf{X} and \mathbf{S} to the known quantities \mathbf{A} , \mathbf{x} , \mathbf{C}_{xx} and \mathbf{C}_{sx} are obtained from the variation function

$$\phi = \hat{\mathbf{r}}^{*T} \mathbf{C}^{*-1} \hat{\mathbf{r}}^* + \hat{\boldsymbol{\delta}}^T \mathbf{C}_x^{-1} \hat{\boldsymbol{\delta}} + 2\hat{\mathbf{k}}^T (\mathbf{A}\hat{\boldsymbol{\delta}} + \mathbf{B}^*\hat{\mathbf{r}}^* + \mathbf{w}), \quad (5.19)$$

where all quantities are defined immediately above and \mathbf{C}_x^{-1} is the weight matrix on the parameters as before. We recognize this to be the variation function of the standard-combined case with weighted parameters. The equations corresponding to the above have been derived in detail in Chapter 2 (equations 2.66 to 2.85). We now specialize these equations to the collocation problem.

Solution for the Parameters

The solution for the parameters is (equation 2.67)

$$\hat{\mathbf{X}} = \mathbf{X} + \hat{\boldsymbol{\delta}}, \quad (5.20)$$

where \mathbf{X} is the vector of weighted parameters, while the correction vector is given by (2.66)

$$\hat{\delta} = - [\mathbf{A}^T (\mathbf{B}^* \mathbf{C}^* \mathbf{B}^{*T})^{-1} \mathbf{A} + \mathbf{C}_x^{-1}]^{-1} \mathbf{A}^T (\mathbf{B}^* \mathbf{C}^* \mathbf{B}^{*T})^{-1} \mathbf{w}. \quad (5.21)$$

In the above

$$\mathbf{B}^* \mathbf{C}^* \mathbf{B}^{*T} = \begin{bmatrix} 0 & \mathbf{I} \end{bmatrix} \begin{pmatrix} \mathbf{C}_{ss} & \mathbf{C}_{sx} \\ \mathbf{C}_{xs} & \mathbf{C}_{xx} \end{pmatrix} \begin{pmatrix} 0 \\ -\mathbf{I} \end{pmatrix} = \mathbf{C}_{xx}. \quad (5.22)$$

Using equation 5.10 and the above,

$$\hat{\delta} = - [\mathbf{A}^T (\mathbf{C} + \mathbf{D})^{-1} \mathbf{A} + \mathbf{C}_x^{-1}]^{-1} \mathbf{A}^T (\mathbf{C} + \mathbf{D})^{-1} \mathbf{w}. \quad (5.23)$$

Note the covariance matrix for the signal at the observation points (\mathbf{C}) and measurement error (\mathbf{D}), and the weight matrix for the parameters (\mathbf{C}_x^{-1}) enter as three separate pieces of information. Also note that the covariance matrix \mathbf{C}_{xs} needed for the prediction of the signal does not affect the solution of the parameters \mathbf{X} since it does not appear in the above equation.

For the case of weighted parameters

$$\mathbf{w} = \mathbf{A}\mathbf{X} - \mathbf{x}, \quad (5.24)$$

where \mathbf{X} is the quasi-observed value of the parameters with weight matrix \mathbf{C}_x^{-1} , and \mathbf{x} are the measurements. For the unweighted case

$$\hat{\delta} = - [\mathbf{A}^T (\mathbf{C} + \mathbf{D})^{-1} \mathbf{A}]^{-1} \mathbf{A}^T (\mathbf{C} + \mathbf{D})^{-1} \mathbf{w}, \quad (5.25)$$

where the misclosure vector \mathbf{w} becomes:

$$\mathbf{w} = -\mathbf{x}, \quad (5.26)$$

or

$$\mathbf{w} = \mathbf{w}^0 = \mathbf{A}\mathbf{X}^0 - \mathbf{x} \quad (5.27)$$

depending upon whether one wishes to solve for the parameters themselves or corrections to some approximate value \mathbf{X}^0 .

Solution for the Signal at the Computation Points

The expression for computing the signal at the computation points follows from the equation for the adjusted observations (equation 2.70) and residual vector (equation 2.69), namely

$$\hat{\mathbf{l}}^* = \mathbf{l}^* + \hat{\mathbf{r}}^*, \quad (5.28)$$

where

$$\hat{\mathbf{r}}^* = -\mathbf{C}^* \mathbf{B}^{*T} \hat{\mathbf{k}} \quad (5.29)$$

and

$$\hat{\mathbf{k}} = (\mathbf{B}^* \mathbf{C}^* \mathbf{B}^{*T})^{-1} (\mathbf{A} \hat{\mathbf{X}} + \mathbf{w}). \quad (5.30)$$

The observed value of the signal can be taken as zero as per the condition imposed upon this quantity - namely it is a random variable with zero mean. Hence the solution for the signal is given by the residual vector. Noting that

$$\hat{\mathbf{l}}^* = \mathbf{0} + \hat{\mathbf{r}}^* = \begin{pmatrix} -\hat{\mathbf{S}} \\ -\hat{\mathbf{Z}} \end{pmatrix} = \begin{pmatrix} \mathbf{C}_{sx} \\ \mathbf{C}_{xx} \end{pmatrix} \mathbf{C}_{xx}^{-1} (\mathbf{A} \hat{\mathbf{X}} + \mathbf{w}) \quad (5.31)$$

the signal is estimated from:

$$\hat{\mathbf{S}} = -\mathbf{C}_{sx} \mathbf{C}_{xx}^{-1} (\mathbf{A} \hat{\mathbf{X}} + \mathbf{w}) = -\mathbf{C}_{sx} (\mathbf{C} + \mathbf{D})^{-1} (\mathbf{A} \hat{\mathbf{X}} + \mathbf{w}). \quad (5.32-5.33)$$

The collocation solution does not contain an equation for \mathbf{Z} , however one can get from equation 5.31 the required expression:

$$\hat{\mathbf{Z}} = -\mathbf{C}_{xx} \mathbf{C}_{xx}^{-1} (\mathbf{A} \hat{\mathbf{X}} + \mathbf{w}) = -(\mathbf{A} \hat{\mathbf{X}} + \mathbf{w}). \quad (5.34)$$

The problem of splitting $\hat{\mathbf{Z}}$ into signal $\hat{\mathbf{S}}$ and noise $\hat{\mathbf{n}}$ estimates is the subject of Section 5.5.

Covariance Matrix for the Parameters

The covariance matrix for the final (adjusted) parameters

$$\hat{\mathbf{X}} = \mathbf{X} + \hat{\boldsymbol{\delta}} \quad (5.35)$$

is, using equations 2.72,

$$\mathbf{C}_{\hat{x}} = [\mathbf{A}^T (\mathbf{B}^* \mathbf{C}^* \mathbf{B}^{*T})^{-1} \mathbf{A} + \mathbf{C}_x^{-1}]^{-1}. \quad (5.36)$$

Considering the collocation definitions of \mathbf{A} , \mathbf{B}^* (equation 5.14), and \mathbf{C}^{*-1} (equations 5.9 and 5.10) we get

$$\mathbf{C}_{\hat{x}} = [\mathbf{A}^T (\mathbf{C} + \mathbf{D})^{-1} \mathbf{A} + \mathbf{C}_x^{-1}]^{-1}. \quad (5.37)$$

In the case that no weights are applied to the parameters

$$\mathbf{C}_{\hat{x}} = [\mathbf{A}^T (\mathbf{C} + \mathbf{D})^{-1} \mathbf{A}]^{-1}, \quad (5.38)$$

which is identical to the collocation equation.

Covariance Matrix of the Signal at the Computation Points

We have seen above that the negative value of the signal at a computation point is like an adjusted observation. The reason for this is that an adjusted observation equals the observation plus the residual. In collocation, the adjusted observations equal zero (for the signal) plus, negative estimated signal (see equations 5.8 and 5.28). Thus the covariance matrix for the signal follows from that of the adjusted observations (equation 2.73), namely

$$\begin{aligned} \mathbf{C}_{\hat{l}^*} = & \mathbf{C}^* + \mathbf{C}^* \mathbf{B}^{*T} (\mathbf{B}^* \mathbf{C}^* \mathbf{B}^{*T})^{-1} \mathbf{A} \mathbf{C}_{\hat{x}} \mathbf{A}^T (\mathbf{B}^* \mathbf{C}^* \mathbf{B}^{*T})^{-1} \mathbf{B}^* \mathbf{C}^* \\ & - \mathbf{C}^* \mathbf{B}^{*T} (\mathbf{B}^* \mathbf{C}^* \mathbf{B}^{*T})^{-1} \mathbf{B}^* \mathbf{C}^* \end{aligned} \quad (5.39)$$

where $\mathbf{C}_{\hat{x}}$ is given above in equation 5.37. Since from adjustments

$$\hat{\mathbf{l}}^* = \mathbf{l}^* + \hat{\mathbf{r}}^*, \quad (5.40)$$

noting \mathbf{l}^* is a null vector in collocation, thus

$$\hat{\mathbf{l}}^* = \mathbf{0} + \hat{\mathbf{r}}^* = \begin{pmatrix} -\hat{\mathbf{S}} \\ -\hat{\mathbf{Z}} \end{pmatrix}, \quad (5.41)$$

then

$$\mathbf{C}_{\hat{l}^*} = \begin{pmatrix} \mathbf{C}_{\hat{s}} & \mathbf{C}_{\hat{s}\hat{z}} \\ \mathbf{C}_{\hat{z}\hat{s}} & \mathbf{C}_{\hat{z}} \end{pmatrix}. \quad (5.42)$$

It is \hat{C}_s^* that we are after. We now specialize equation 5.39 as before by using the definitions of the various terms

$$\begin{aligned}\hat{C}_1^* &= \begin{pmatrix} \hat{C}_s & \hat{C}_{sz} \\ \hat{C}_{zs} & \hat{C}_z \end{pmatrix} \\ &= \begin{pmatrix} C_{ss} & C_{sx} \\ C_{xs} & C_{xx} \end{pmatrix} + \begin{pmatrix} C_{sx} \\ C_{xx} \end{pmatrix} C_{xx}^{-1} A C_x^* A^T C_{xx}^{-1} [C_{xs} \ C_{xx}] \\ &\quad - \begin{pmatrix} C_{sx} \\ C_{xx} \end{pmatrix} C_{xx}^{-1} [C_{xs} \ C_{xx}] \quad .\end{aligned}\tag{5.43}$$

Picking out only the upper left portion from the above hyper-matrix equation, one obtains

$$\begin{aligned}\hat{C}_s &= C_{ss} + C_{sx} C_{xx}^{-1} A C_x^* A^T C_{xx}^{-1} C_{xs} - C_{sx} C_{xx}^{-1} C_{xs} \\ &= C_{ss} + C_{sx} (C + D)^{-1} A C_x^* A^T (C + D)^{-1} C_{xs} - C_{sx} (C + D)^{-1} C_{xs},\end{aligned}\tag{5.44-5.45}$$

which is identical to the collocation equation.

Let us now examine the expressions for the parameters (equation 5.23), signal equation 5.33 and their respective covariance matrices (equations 5.37 and 5.45) to determine the number, form and size of matrices to be inverted. The solution and covariance matrix for the parameters requires two matrix inversions in

$$[A^T (C + D)^{-1} A]^{-1}.\tag{5.46}$$

The first inversion is a fully populated matrix because of the correlation among the signals at the computation points and has dimensions equal to the number of observations. In certain applications this matrix can be reduced to a band matrix and then by using special inversion tactics, like *compacting out the zero parts* the inversion is made faster, e.g. [Krakiwsky and Pope 1967; Isner 1972; Knight and Steeves 1974]. The second inversion is of a fully populated matrix of order equal to the number of parameters. Equations 5.33 and 5.45 (pertaining to the signal) require no additional inverses to that made in the determination of the parameters.

5.4 Stepwise Collocation

We have seen that collocation equations follow directly from the equations of the standard-combined case of adjustments simply by specifying the collocation forms of the design matrices \mathbf{A} , \mathbf{B} , and weight matrix \mathbf{C}^{*-1} . On the other hand, we cannot deduce the sequential collocation equations of Moritz [1973a] from the Kalman (sequential) expressions derived in Chapter 3 and 4. This is because in the latter development the observations of consecutive stages are assumed to be uncorrelated.

This assumption is prohibitive in collocation as the signal part of the observation has a continuous and complete correlation which must be accounted for when grouping the observations in stages. Thus one sees that from this point of view, collocation is a more general method. Below we simply state the sequential collocation equations from Moritz. Recall the collocation form for the solution of the parameters (equations 5.20, 5.23 and 5.25): namely

$$\hat{\mathbf{X}} = [\mathbf{A}^T \mathbf{C}_{xx}^{-1} \mathbf{A}]^{-1} \mathbf{A}^T \mathbf{C}_{xx}^{-1} \mathbf{x}. \quad (5.47)$$

In sequential collocation the matrices and vectors are partitioned in two, such that

$$\mathbf{A} = \begin{pmatrix} \mathbf{A}_1 \\ \mathbf{A}_2 \end{pmatrix}, \quad (5.48)$$

$$\mathbf{x} = \begin{pmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{pmatrix}, \quad (5.49)$$

$$\mathbf{C}_{xx} = \begin{pmatrix} \mathbf{C}_{11} & \mathbf{C}_{12} \\ \mathbf{C}_{21} & \mathbf{C}_{22} \end{pmatrix} \quad (5.50)$$

The solution for the parameters is given by

$$\hat{\mathbf{X}} = \hat{\mathbf{X}}_1 + \mathbf{C}_1 \bar{\mathbf{A}}_2^T \bar{\mathbf{C}}_{22}^{-1} (\mathbf{x}_2 - \mathbf{C}_{21} \mathbf{C}_{11}^{-1} \mathbf{x}_1 - \bar{\mathbf{A}}_2 \hat{\mathbf{X}}_1), \quad (5.51)$$

where

$$\mathbf{C}_1 = [\mathbf{A}_1^T \mathbf{C}_{11}^{-1} \mathbf{A}_1]^{-1}, \quad (5.52)$$

$$\bar{\mathbf{A}}_2 = \mathbf{A}_2 - \mathbf{C}_{21} \mathbf{C}_{11}^{-1} \mathbf{A}_1, \quad (5.53)$$

$$\bar{\mathbf{C}}_{22}^{-1} = [\mathbf{C}_{22} - \mathbf{C}_{21} \mathbf{C}_{11}^{-1} \mathbf{C}_{12} + \bar{\mathbf{A}}_2 \mathbf{C}_1 \bar{\mathbf{A}}_2^T]^{-1}, \quad (5.54)$$

and

$$\hat{\mathbf{X}}_1$$

is the solution using only observations \mathbf{x}_1 , that is

$$\hat{\mathbf{X}}_1 = (\mathbf{A}_1^T \mathbf{C}_{11}^{-1} \mathbf{A}_1)^{-1} \mathbf{A}_1^T \mathbf{C}_{11}^{-1} \mathbf{x}_1. \quad (5.55)$$

The solution for a signal element in two steps, is given by

$$\begin{aligned} \hat{\mathbf{S}}_p = \hat{\mathbf{S}}_1 + (\mathbf{C}_{p2} - \mathbf{C}_{p1} \mathbf{C}_{11}^{-1} \mathbf{C}_{12} - \mathbf{C}_{p1} \mathbf{C}_{11}^{-1} \mathbf{A}_1 \mathbf{C}_1 \bar{\mathbf{A}}_2^T) \bar{\mathbf{C}}_{22}^{-1} \\ (\mathbf{x}_2 - \mathbf{C}_{21} \mathbf{C}_{11}^{-1} \mathbf{x}_1 - \bar{\mathbf{A}}_2 \hat{\mathbf{X}}_1), \end{aligned} \quad (5.56)$$

where the newly introduced terms are defined as

$$\mathbf{C}_p^T = [\mathbf{C}_{p1} \quad \mathbf{C}_{p2}] \quad (5.57)$$

and $\hat{\mathbf{S}}_1$ is the signal computed using only observations \mathbf{x}_1 , that is

$$\hat{\mathbf{S}}_1 = \mathbf{C}_{p1} \mathbf{C}_{11}^{-1} (\mathbf{x}_1 - \mathbf{A}_1 \hat{\mathbf{X}}_1). \quad (5.58)$$

\mathbf{C}_p^T is a row vector of dimensions $1 \times (n_1 + n_2)$ [Moritz 1973a]; here only one signal element ($\hat{\mathbf{S}}_p$) is estimated. However, the same formula could be applied for all p signal elements (vector $\hat{\mathbf{S}}$) by using

$$\mathbf{C}_{sx} = \begin{pmatrix} \mathbf{C}_{s1} & \& \mathbf{C}_{s2} \\ p \times n_1 & & p \times n_2 \end{pmatrix}$$

instead of \mathbf{C}_p .

The covariance matrix for the parameters is

$$\bar{\mathbf{C}}_{\hat{\mathbf{X}}} = \bar{\mathbf{C}}_{\hat{\mathbf{X}}1} - \mathbf{C}_1 \bar{\mathbf{A}}_2^T \bar{\mathbf{C}}_{22}^{-1} \bar{\mathbf{A}}_2 \mathbf{C}_1, \quad (5.59)$$

where

$$\hat{C}_{\hat{x}1} = C_1. \quad (5.60)$$

In the above sequential collocation expressions, one sees that one matrix inversion is necessary, namely that of \bar{C}_{22} which has dimensions equal to the number of observations in the second stage. It is interesting to note how the correlation among all the signal elements is accounted for even though the full matrix

$$C_{xx} = C + D = \bar{C} = \begin{pmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{pmatrix} \quad (5.61)$$

is not explicitly inverted. We stress that it is **implicitly** inverted thereby giving rise to expressions like equations 5.55 and 5.56. Thus we see how the correlation among the signal elements, which is essential for collocation, is accounted for in the sequential expressions.

REFERENCES

- ISNER, J. Determination of surface densities from a combination of gravimetry and satellite altimetry. *Reports of the Department of Geodetic Science*, No. 186, The Ohio State University, Columbus, 1972.
- KNIGHT, W. and STEEVES, R. Partial solution of the variance-covariance matrix for geodetic networks. *The Canadian Surveyor*, 28, No. 5, 1974.
- KRAKIWSKY, E.J. A synthesis of Recent Advances: Method of Least Squares. Lecture Notes #24, Department of Geomatics Engineering, The University of Calgary, Calgary, Alberta, Canada.
- KRAKIWSKY, E.J. and POPE, A.J. Least squares adjustment of satellite observations for simultaneous directions and ranges. *Reports of the Department of Geodetic Science*, No. 86, The Ohio State University, Columbus, 1967.
- MORITZ, H. Advanced least squares methods. *Reports of the Department of Geodetic Science*, No. 75, The Ohio State University, Columbus, 1972.
- MORITZ, H. Determination of the gravity field by collocation. *Bolletino Di Geodesia E Science Affini*, Anno XXXII, 1973.

MORITZ, H. Stepwise and sequential collocation. Reports of the Department of Geodetic Science, No. 203, The Ohio State University, Columbus, 1973a.

STATISTICAL TESTING AND ANALYSIS

Testing statistics and equations for least squares and Kalman filtering are derived in this chapter. Potential problems for the developed test statistics and ways to tackle them are also discussed.

1.0 Statistic Testing in Least Squares

Consider a linear model

$$\mathbf{l} = \mathbf{A}\mathbf{x} + \boldsymbol{\varepsilon} \quad (6.1a)$$

$$\mathbf{C} = \sigma_0^2 \mathbf{P}^{-1} = \sigma_0^2 \mathbf{Q} \quad (6.1b)$$

where \mathbf{A} = the $n \times u$ design matrix; \mathbf{x} = the $u \times 1$ unknown parameter vector; \mathbf{l} = the $n \times 1$ observation vector; $\boldsymbol{\varepsilon}$ = the $n \times 1$ random error vector assumed to be normal distributed; \mathbf{C} = $n \times n$ a priori covariance matrix of the observations; \mathbf{Q} = the $n \times n$ weight coefficient matrix of the observations; \mathbf{P} = the $n \times n$ weight matrix of the observations; and σ_0^2 is the a priori variance factor. By means of the least squares method, we have the following results

$$\hat{\mathbf{x}} = (\mathbf{A}^T \mathbf{C}^{-1} \mathbf{A})^{-1} \mathbf{A}^T \mathbf{C}^{-1} \mathbf{l} \quad (6.2)$$

with the corresponding covariance matrix

$$\mathbf{C}_{\hat{\mathbf{x}}} = (\mathbf{A}^T \mathbf{C}^{-1} \mathbf{A})^{-1} = \sigma_0^2 \mathbf{Q}_{\hat{\mathbf{x}}} \quad (6.3)$$

The estimated residual vector is

$$\hat{\mathbf{v}} = \mathbf{A}\hat{\mathbf{x}} - \mathbf{l} = [\mathbf{A}(\mathbf{A}^T\mathbf{C}^{-1}\mathbf{A})^{-1}\mathbf{A}^T - \mathbf{C}]\mathbf{C}^{-1}\mathbf{l} \quad (6.4)$$

with covariance matrix

$$\mathbf{C}_{\hat{\mathbf{v}}} = \mathbf{C} - \mathbf{A}(\mathbf{A}^T\mathbf{C}^{-1}\mathbf{A})^{-1}\mathbf{A}^T = \sigma_0^2\mathbf{Q}_{\hat{\mathbf{v}}} \quad (6.5)$$

The quadratic sum of the weighted residuals is

$$\mathbf{R} = \hat{\mathbf{v}}^T\mathbf{P}\hat{\mathbf{v}} \quad (6.6)$$

For the case of blunders in the observations, the model equation 6.1a has to be extended to the form

$$\mathbf{E}(\mathbf{l}) = \mathbf{A}\mathbf{x} + \mathbf{H}\nabla \quad \mathbf{C} = \sigma_0^2\mathbf{P}^{-1} \quad (6.7)$$

where $\mathbf{H}\nabla$ represents the suspected blunders in the linear model. The least-square estimate of \mathbf{x} in the extended model becomes

$$\hat{\mathbf{x}}' = (\mathbf{A}^T\mathbf{C}^{-1}\mathbf{A})^{-1}[\mathbf{A}^T\mathbf{C}^{-1}\mathbf{l} - \mathbf{A}^T\mathbf{C}^{-1}\mathbf{H}\nabla] \quad (6.8)$$

and

$$\begin{aligned} \hat{\mathbf{v}}' &= \mathbf{A}\hat{\mathbf{x}}' - \mathbf{l} = [\mathbf{A}(\mathbf{A}^T\mathbf{C}^{-1}\mathbf{A})^{-1}\mathbf{A}^T - \mathbf{C}]\mathbf{C}^{-1}\mathbf{l} \\ &\quad - \mathbf{A}(\mathbf{A}^T\mathbf{C}^{-1}\mathbf{A})^{-1}\mathbf{A}^T\mathbf{C}^{-1}\mathbf{H}\hat{\nabla} = \hat{\mathbf{v}} - \mathbf{A}\mathbf{C}_{\hat{\mathbf{x}}}\mathbf{A}^T\mathbf{C}^{-1}\mathbf{H}\hat{\nabla} \end{aligned} \quad (6.9)$$

Clearly, the existence of blunders will affect the estimation of the unknown parameters. The quadratic sum of the residuals in the extended model now becomes

$$\mathbf{R}' = \hat{\mathbf{v}}'^T\mathbf{P}\hat{\mathbf{v}}' \quad (6.10)$$

and it can be proven that

$$\mathbf{R}' = \mathbf{R} + \Delta\mathbf{R} \quad (6.11)$$

where

$$\Delta\mathbf{R} = \hat{\mathbf{v}}^T\mathbf{P}\mathbf{H}(\mathbf{H}^T\mathbf{P}\mathbf{Q}_{\hat{\mathbf{v}}}\mathbf{P}\mathbf{H})^{-1}\mathbf{H}^T\mathbf{P}\hat{\mathbf{v}} \quad (6.12)$$

ΔR represents the increase of the quadratic form of the residuals due to the presence of blunders and thus it can be used to formulate a statistic to test the significance of possible blunders the observations.

The null hypotheses H_0 , the alternative hypothesis H_a , and test statistic T are given as follows:

$$H_0: E(l) = Ax \quad H_a: E(l) = Ax + H\nabla$$

$$T = \Delta R \sigma_0^{-2} = \hat{v}^T P H (H^T P Q_{\hat{v}} P H)^{-1} H^T P \hat{v} \sigma_0^{-2} \sim \chi^2(p, \lambda^2) \quad (6.13)$$

where H is a $n \times p$ matrix; and ∇ is a $p \times 1$ vector that is the presumed blunder vector for testing. When H_0 is valid, the noncentral parameter λ^2 is equal to zero, otherwise

$$\lambda^2 = \frac{\hat{V}^T H^T P Q_{\hat{v}} P H \hat{V}}{\sigma_0^2} \quad (6.14)$$

Represented in equation 6.13 is a general statistic, with which multiple blunders can be tested. In practice, however, only a one-dimensional test is usually performed because the actual blunders are not known beforehand, i.e., there is difficulty to specify the alternative hypothesis via the H matrix. In this case, the matrix H reduces to a vector of the form

$$H = e_i = [0, 0, \dots, 0, 1, 0, \dots, 0]^T \quad (6.15)$$

which is known as the influence vector in the literature. The test statistic for the i th observation then becomes

$$t_i = \frac{(e_i^T P \hat{v})^2}{\sigma_0^2 (e_i^T P Q_{\hat{v}} P e_i)} \sim \chi^2(\lambda^2) \quad (6.16)$$

or further

$$w_i = \sqrt{t_i} = \frac{e_i^T P \hat{v}}{\sigma_0 \sqrt{e_i^T P Q_{\hat{v}} P e_i}} \sim N(\delta_i, 1) \quad (6.17)$$

where

$$\delta_i = \frac{\nabla l_i \sqrt{e_i^T P Q_{\hat{v}} P e_i}}{\sigma_0} \quad (6.18)$$

in which ∇l_i is the blunder in the i -th observation.

If we assume that the observations are uncorrelated, the statistic can be further simplified as follows:

$$w_i = \frac{|\hat{v}_i|}{\sigma_0 \sqrt{(Q_{\hat{v}})_{ii}}} = \frac{\hat{v}_i}{\sigma_{\hat{v}_i}} \sim N(\delta_i, 1) \quad (6.19)$$

where

$$\delta_i = \nabla l_i \frac{\sqrt{(Q_{\hat{v}} \mathbf{P})_{ii}}}{\sigma_{li}} \quad (6.20)$$

which is the corresponding noncentrality parameter. Equation 6.18 is the test statistic of data snooping developed by Baarda (1968). When an estimated standard deviation is used, the aforementioned statistic has a τ or student's (t) distribution depending on how the standard deviation is computed.

In summary, statistical testing on observations includes two steps:

1) Detection (testing variance)

$$H_0 : E(v) = 0 \quad H_a : E(v) = \nabla$$

$$T = r \frac{\hat{\sigma}^2}{\sigma_0^2} \begin{cases} H_0 \sim \chi^2(r, 0) \\ H_a \sim \chi^2(r, \lambda^2) \end{cases} \quad (6.21)$$

$$\text{where } \lambda^2 = \frac{\nabla^T \mathbf{P}_{\nabla} \nabla}{\sigma_0^2}.$$

2) Identification (testing individual residual)

$$H_0 : E(v_i) = 0 \quad H_a : E(v_i) = \nabla_i$$

$$w_i = \frac{e_i^T \mathbf{P}_l v_i}{\sigma_0 \sqrt{e_i^T \mathbf{P}_l \mathbf{Q}_v \mathbf{P}_l e_i}} \begin{cases} H_0 \sim N(0, 1) \\ H_a \sim N(\delta_i, 1) \end{cases} \quad (6.22)$$

where $\delta_i = \frac{\nabla_i \sqrt{e_i^T P_l Q_v P_l e_i}}{\sigma_0}$ and $e_i = (0 \dots 1 \dots 0)^T$ in which all elements equals zero except the i -th element in the vector.

When the observations are all uncorrelated, then the identification statistic in equation 6.21 reduces to

$$\begin{aligned} w_i &= \frac{v_i}{\sigma_{v_i}} = \frac{v_i}{\sigma_{l_i} \sqrt{r_i}} | H_0 \sim N(0,1) \\ &| H_a \sim N(\delta_i, 1) \end{aligned} \quad (6.23)$$

$$\text{where } \delta_i = \frac{\nabla_i \sqrt{r_i}}{\sigma_{l_i}}$$

2.0 Statistical Testing in Kalman Filtering

In this section, we will discuss innovation-based failure (outlier, blunder, bias etc) detection and identification procedures in Kalman filtering.

Innovation sequence is defined as the difference between the actual system output and the predicted output based on the predicted state, namely:

$$v_k^{(-)} = z_k - H_k \hat{x}_k \quad (6.24)$$

which is obtained directly from equation 3.61a. From equation 6.24, we find that the equation is actually the predicted residual vector of observations. The reason it is called the innovation sequence is that it represents the new information brought in by the latest observation vector. Under the assumed conditions specified in equations 3.64, 3.65 and 3.6, the innovation sequence is a zero mean Gaussian white noise sequence with known covariance matrix, namely:

$$\begin{aligned} E[v_k^{(-)}] &= 0 \\ E[v_k^{(-)} (v_k^{(-)})^T] &= C_{v_k}^{(-)} = R_k + H_k P_k^* H_k^T \end{aligned} \quad (6.25)$$

where $C_{v_k}^{(-)}$ is the covariance matrix of the innovation sequence.

If a failure is present in the system, the above properties are no longer valid and the existing failures will cause the innovation sequence to depart from its

mean and the whiteness properties. Thus, the innovation sequence can be used to detect possible abnormal behaviour of the underlying system. In the following the test statistics for failure detection and identification are derived and their characteristics are explored based on the innovation sequence.

To globally test the existence of any system failures, let H_0 and H_a denote the null and alternative hypothesis, respectively, which are defined below:

$$H_0 : E(v_k^{(-)}) = 0, \quad H_a : E(v_k^{(-)}) \neq 0. \quad (6.26)$$

The test statistics can then be formed directly based on the innovation sequence, i.e.,

$$T_k = v_k^{(-)T} \{C_{v_k}^{(-)}\}^{-1} v_k^{(-)} \quad | H_0 \sim \chi^2(n_k, 0), \quad (6.27a)$$

$$T'_k = v_k^{(-)T} \{C_{v_k}^{(-)}\}^{-1} v_k^{(-)} \quad | H_a \sim \chi^2(n_k, \lambda_k^2), \quad (6.27b)$$

where n_k is the number of observations taken at time t_k and $\lambda_k^2 = \nabla_k^T \{C_{v_k}^{(-)}\}^{-1} \nabla_k$ is the noncentrality parameter in which ∇_k is the failure vector under H_a . When the global test is rejected, i.e., there exist system failure; a more specific alternative hypothesis has to be formulated if one has some idea of the causes leading to the departures from the nominal values (e.g., sensor failure or outliers in the data, etc.).

Below, test statistics for failure detection and identification with the innovation-based approach are derived. The derivation herein, however, is different from that currently used, which is advantageous in exploring the intrinsic characteristics of the test statistics. Considering failure vector ∇_k in the measurement mode, test statistics can be formulated based on the following extended filtering model against the standard mode:

$$E \begin{Bmatrix} x_k^- \\ z_k \end{Bmatrix} = \begin{bmatrix} I & 0 \\ H_k & B_k \end{bmatrix} \begin{Bmatrix} x_k^+ \\ \nabla_k \end{Bmatrix}, \quad \begin{bmatrix} P_k^- & 0 \\ 0 & R_k \end{bmatrix}, \quad (6.28)$$

where ∇_k is the failure vector specified in the test and has dimension of $b_k \times 1$ ($b_k < n_k$). B_k is assumed to be a known matrix with dimensions $n_k \times b_k$ and if full rank. Using the least squares principle, we obtain the following normal equations

$$\begin{aligned}
& \begin{bmatrix} I & H_k^T \\ 0 & B_k^T \end{bmatrix} \begin{bmatrix} [P_k^-]^{-1} & 0 \\ 0 & R_k^{-1} \end{bmatrix} \begin{bmatrix} I & 0 \\ H_k & B_k \end{bmatrix} \begin{bmatrix} x_k^+ \\ \nabla_k \end{bmatrix} \\
&= \begin{bmatrix} I & H_k^T \\ 0 & B_k^T \end{bmatrix} \begin{bmatrix} [P_k^-]^{-1} & 0 \\ 0 & R_k^{-1} \end{bmatrix} \begin{bmatrix} x_k^- \\ z_k \end{bmatrix}.
\end{aligned} \tag{6.29}$$

Eliminating the parameter $x_k^{(+)}$ we have

$$\begin{aligned}
\hat{V}_k &= \{B_k^T R_k^{-1} B_k - B_k^T R_k^{-1} H_k [(P_k^-)^{-1} + H_k^T R_k^{-1} H_k]^{-1} H_k^T R_k^{-1} B_k\}^{-1} \\
&\{B_k^T R_k^{-1} z_k - B_k^T R_k^{-1} H_k [(P_k^-)^{-1} + H_k^T R_k^{-1} H_k]^{-1} [(P_k^-)^{-1} x_k^- + H_k^T R_k^{-1} z_k]\}.
\end{aligned} \tag{6.30}$$

Applying the following well-known relationship to it,

$$(C^{-1} + A^T B A)^{-1} = C - C A^T (B + A C A^T)^{-1} A C,$$

Equation 6.30 can be written as

$$\begin{aligned}
\hat{V}_k &= \{B_k^T R_k^{-1} B_k - B_k^T R_k^{-1} H_k [P_k^- - P_k^- H_k^T (R_k H_k P_k^- H_k^T)^{-1} H_k P_k^-] H_k^T R_k^{-1} B_k\}^{-1} \\
&\{B_k^T R_k^{-1} z_k - B_k^T R_k^{-1} H_k [P_k^- - P_k^- H_k^T (R_k + H_k P_k^- H_k^T)^{-1} H_k P_k^-] [(P_k^-)^{-1} x_k^- + H_k^T R_k^{-1} z_k]\} \\
&= \{B_k^T R_k^{-1} B_k - B_k^T R_k^{-1} H_k P_k^- H_k^T R_k^{-1} B_k + B_k^T R_k^{-1} H_k P_k^- H_k^T (C_{vk}^{(-)})^{-1} H_k P_k^- H_k^T R_k^{-1} B_k\}^{-1} \\
&\{B_k^T R_k^{-1} z_k - B_k^T R_k^{-1} H_k x_k^- + B_k^T R_k^{-1} H_k P_k^- H_k^T (C_{vk}^{(-)})^{-1} H_k x_k^- - \\
&B_k^T R_k^{-1} H_k P_k^- H_k^T R_k^{-1} z_k + B_k^T R_k^{-1} H_k P_k^- H_k^T (C_{vk}^{(-)})^{-1} H_k P_k^- H_k^T R_k^{-1} z_k\}.
\end{aligned} \tag{6.31}$$

Since $H_k P_k^- H_k^T (C_{vk}^{(-)})^{-1} + R_k (C_{vk}^{(-)})^{-1} = I$, equation 6.31 can be further expressed as

$$\begin{aligned}
\hat{V}_k &= \{B_k^T R_k^{-1} B_k - B_k^T (C_{vk}^{(-)})^{-1} H_k P_k^- H_k^T R_k^{-1} B_k\}^{-1} \\
&\{B_k^T R_k^{-1} z_k + B_k^T (C_{vk}^{(-)})^{-1} H_k x_k^- - B_k^T (C_{vk}^{(-)})^{-1} H_k P_k^- H_k^T R_k^{-1} z_k\} \\
&= [B_k^T (C_{vk}^{(-)})^{-1} B_k]^{-1} B_k^T (C_{vk}^{(-)})^{-1} v_k^{(-)},
\end{aligned} \tag{6.32}$$

The corresponding covariance matrix is

$$C_{\nabla k} = [B_k^T (C_{vk}^{(-)})^{-1} B_k]^{-1}. \tag{6.33}$$

The squared norm of $\hat{\nabla}_k$ is a quadratic form and can be formulated as a test statistic to test whether such an error is significant, namely:

$$H_0 : E(\hat{\nabla}_k) = 0, \quad H_a : E(\hat{\nabla}_k) \neq 0, \quad (6.34)$$

$$T_{bk} = \hat{\nabla}_k^T C_{\nabla k}^{-1} \hat{\nabla}_k \quad \left| H_0 \sim \chi^2(b_k, 0), \text{ and} \right. \quad (6.35a)$$

$$T_{bk}' = \hat{\nabla}_k^T C_{\nabla k}^{-1} \hat{\nabla}_k \quad \left| H_a \sim \chi^2(b_k, \lambda_{bk}^2). \right. \quad (6.35b)$$

The noncentrality parameter is

$$\lambda_{bk}^2 = \nabla_k^T [B_k^T (C_{vk}^{(-)})^{-1} B_k]^{-1} \nabla_k. \quad (6.36)$$

Using the relationship in equations 6.32 and 6.33 to equations 6.35a and 6.35b, we obtain

$$T_{bk} = v_k^{(-)T} (C_{vk}^{(-)})^{-1} B_k [B_k^T (C_{vk}^{(-)})^{-1} B_k]^{-1} B_k^T (C_{vk}^{(-)})^{-1} v_k^{(-)} \quad \left| H_0 \sim \chi^2(b_k, 0) \right. \quad (6.37)$$

$$T_{bk}' = v_k^{(-)T} (C_{vk}^{(-)})^{-1} B_k [B_k^T (C_{vk}^{(-)})^{-1} B_k]^{-1} B_k^T (C_{vk}^{(-)})^{-1} v_k^{(-)} \quad \left| H_a \sim \chi^2(b_k, \lambda_{bk}^2). \right. \quad (6.38)$$

It is found that the above statistic has been expressed as a function of the innovation sequence, which is why it is called the innovation-based approach. In an actual application, ∇_k and B_k are usually unknown and difficult to give beforehand. This has been considered as the most difficult task in quality control. In order to implement the procedure in practice, it is thus common to consider only one failure each time, i.e., the vector ∇_k becomes a scalar and the statistic reduces to a one-dimensional test procedure:

$$H_0 : E(\hat{\nabla}_{ik}) = 0, \quad H_a : E(\hat{\nabla}_{ik}) \neq 0, \quad (6.39)$$

$$t_{ik} = \frac{e_i^T (C_{vk}^{(-)})^{-1} v_k^{(-)}}{\sqrt{e_i^T (C_{vk}^{(-)})^{-1} e_i}} \quad \left| H_0 \sim N(0,1), \text{ and} \right. \quad (6.40)$$

$$t_{ik}' = \frac{e_i^T (C_{vk}^{(-)})^{-1} v_k^{(-)}}{\sqrt{e_i^T (C_{vk}^{(-)})^{-1} e_i}} \quad \left| H_a \sim N(\delta_{ik}, 1), \right. \quad (6.41)$$

where $e_i = (0, 0, \dots, 1, \dots, 0, 0)^T$ and $\delta_{ik} = \nabla_{ik} \sqrt{e_i^T (C_{vk}^{(-)})^{-1} e_i}$ is the noncentrality parameter.

Implementation of the innovation-based failure detection and identification is relatively simple since the innovation sequence and its covariance matrix can be obtained directly from the Kalman filter without additional computation. Also the method suffers no degradation prior detection.

6.3 Reliability Analysis

1.0.0 Residual Analysis

Given a linear measurement system model defined by

$$Ax = 1 + v \quad 1 \sim N(\mu, \sigma^2 P_1^{-1}) \text{ or } 1 \sim N(\mu, \sigma^2 Q_1) \quad (6.42)$$

where A is $n \times m$ matrix and $\text{Rank}(A) = m$, the corresponding least squares estimate of x becomes

$$\hat{x} = (A^T P_1 A)^{-1} A^T P_1 1 \quad (6.43)$$

We can also prove that

$$C_{\hat{x}} = \sigma^2 (A^T P_1 A)^{-1} = \sigma^2 Q_{\hat{x}} \quad (6.44)$$

$$\begin{aligned} v = \hat{1} - 1 &= A\hat{x} - 1 = A(A^T P_1 A)^{-1} A^T P_1 1 - 1 \\ &= [A(A^T P_1 A)^{-1} A^T - Q_1] P_1 1 \end{aligned} \quad (6.45)$$

$$C_v = \sigma^2 [Q_1 - A(A^T P_1 A)^{-1} A^T] = \sigma^2 Q_v \quad (6.46)$$

Applying equation 6.43 to equation 6.42, we have

$$v = -(Q_v P_1) 1 \quad (6.47)$$

Equation 6.47 is an important equation that has described the relationship between the observations and residuals. Most outlier analysis methods are based on testing the residuals. $Q_v P_1$ is an idempotent matrix and therefore we have

$$\text{tr}(Q_v P_1) = \sum_{i=1}^n (Q_v P_1)_{ii} = \sum_{i=1}^n r_i = r = n - m .$$

$r_i = (Q_v P_1)_{ii}$ is often called redundancy number. From equation 6.47, we are able to describe

- 1) the impact of all observations on a specific residual

$$v_i = -[(Q_v P_1)1]_i = -\sum_{j=1}^n [(Q_v P_1)_{ij} 1_j] \quad (6.48)$$

- 2) the impact of a specific observation on all residuals

$$v_j = -(Q_v P_1)_{ji} 1_i \quad j = 1, 2, \dots, n \quad (6.49)$$

- 3) the impact of a specific observation on its own residual

$$v_i = -(Q_v P_1)_{ii} 1_i = -r_i 1_i \quad (6.50)$$

2.0.0 Reliability Analysis

There are several problems associated with any statistic test:

- 1) A statistical test always requires define a probability of type-I error α , i.e. the probability of rejecting a true hypothesis.
- 2) The power of a statistical test is usually defined by $\gamma = 1 - \beta$ where β is the probability of type-II error, i.e., the probability of accepting a wrong hypothesis.
- 3) The power γ in fact depends not only on the predefined α but also the actual observation and stochastic models. The latter of course depends on the problem in your hand.

To ensure if a statistical test is actually conducted or a statistical decision is made under sufficient power, a theory called reliability analysis was developed by Baarda (1968) for that purpose. Reliability analysis consists of two components: internal reliability and external reliability.

- 1) Internal reliability: describes the minimum size of a blunder that can be detected with predefined power.
- 2) External reliability: describes the maximum impact of any undetectable blunder in 1) on the estimates.

They are derived based on equation 6.40 and therefore it often assumes that only a single blunder exists in the entire data set.

- **Internal reliability:**

Give the type-I error and test power, namely, α and γ threshold can be determined for w_i from tables of standard normal distribution:

$$\nabla_0 w_i = \delta_0 = \delta_0(\alpha, \gamma) \quad (6.51)$$

Applying equation 6.51 to equation 6.17, we have

$$\nabla_0 l_i = \frac{\sigma \delta_0}{\sqrt{e_i^T P_i Q_v P_i e_i}} \quad (6.52)$$

With uncorrelated observations, we have

$$\nabla_0 l_i = \frac{\sigma \delta_0}{p_i \sqrt{r_i}} = \sigma_{l_i} \frac{\delta_0}{\sqrt{r_i}} \quad (6.53)$$

where $\nabla_0 l_i$ is called the minimum detectable error (blunder) with given α and γ .

- **External reliability:**

For any error that could not be detected by given α and γ , the error then remains in the observations. Therefore it is a good idea to assess the maximum possible impact on the estimate of the unknown parameter x .

Applying equation 6.52 to equation 6.2 results in the following maximum impact of the undetectable error $\nabla_0 l_i$ on the parameter x .

$$\nabla_0 x_i = (A^T P_i A)^{-1} A^T P_i \nabla_0 l_i \quad i = 1, 2, \dots, n \quad (6.54)$$

Often the following quadratic form of $\nabla_0 l_i$ is used to define the external reliability as a quantity to measure the impact due to minimum undetectable blunders.

$$\bar{\delta}_{0i} = \frac{1}{\sigma} \sqrt{\nabla_0 x_i^T Q_x^{-1} \nabla_0 x_i} \quad (6.55)$$

The above reliability analysis approach can be applied to the case of Kalman Filtering. For testing statistic given in equation 6.40, the minimum detectable bias can be defined by

$$(\nabla v_k)_i = \frac{\delta_0}{\sqrt{e_i^T (C_{vk}^{(-)})^{-1} e_i}} \quad i = 1, 2, \dots, n \quad (6.56)$$

More details can be found in Teunissen and Salzmann (1989).

2.0 Robust Statistical Testing

The data snooping procedure is based on the testing of least squares residuals. The statistic defined in equation 6.17, however, has been recognized as not robust. The cause is due to the smoothness effect of a conventional least squares adjustment procedure. In other words, a single blunder in the measurements tends to spread or smooth its effect into all observations. As the result of this, a large residual is not necessary to say that the corresponding observation contains a blunder. In the same way, a small residual also does not necessarily mean that the corresponding observation is blunder free. The situation becomes even worse when there exist multiple blunders in the measurements. Generally speaking, the data snooping statistic in equation 6.17 is suitable only for single blunder identification and it is still a challenge to develop effective statistic for multiple blunder identification (Gao, 1992).

In order for statistic capable of multiple blunder identification, the test procedure must be robustified. In the following, a robust test statistic is introduced in which the residuals will not be affected by the so-called smoothness effect from least squares (Gao et al., 1992). All the blunders can then be more reliably identified than the conventional data snooping method.

It has been shown that the methods so far are usually only suitable for the detection of one blunder. Besides, the residuals in the statistic may not represent the actual discrepancies of the corresponding observations due to the smoothing effect of the least-squares adjustment. In other words, the blunders will be spread (distributed) to other (good and bad) observations and thus the sizes of the actual blunders will be distorted. As a result, incorrect decisions may be derived from the statistical test, i.e., a good observation may be rejected or a bad observation may not be detected at all (see the two examples). This problem becomes more acute with an increasing number of observations and unknown parameters.

Due to the problem pointed out in the previous section, a robust test procedure is thus proposed. In the method, the theories of robust estimation and statistical testing are integrated by combining their respective strengths and hereby providing a reliable procedure for the detection and identification of multiple blunders. First, the residuals are calculated through robust estimation. The robust residuals are then used to formulate the test statistic. The detection and the isolation of the blunders can thus be greatly improved compared with the conventional least-squares method. The presence of blunders not only can be detected but also the blunders themselves can be precisely isolated.

The L_1 norm method, which minimizes the sum of the absolute residuals, has been recognized as an effective robust method. The discussions, however, have been limited to obtaining a solution for the unknown parameters. The major advantage of the method is that the solution is not sensitive to blunders. In other words, reasonably good estimates for the unknowns are obtained even with the presence of blunders. A disadvantage of the method is the resultant low accuracy stemming from a unique solution; in other words, the method only searches for the u necessary observation and neglects all redundant observations. This can hardly be accepted in survey practice since surveyors usually make redundant observations for the purpose of obtaining high accuracy and reliability through high redundancy. In summary, the redundant observations in the L_1 norm method are only treated as possible candidates when searching for the u necessary observations with smallest absolute errors.

Mathematically, the L_1 norm methods can be described as follows:

$$\min \sum |v_i| \quad (6.57)$$

subject to $Ax = l + v$

which can be easily transformed into a standard linear programming problem. Sophisticated algorithms exist for the solution of this linear programming problem like the simplex method and the computer software is readily available. The unique solution for the unknowns is

$$\tilde{x} = (A_1)^{-1}l_1 \quad (6.58)$$

where l_1 contains the u observations l_1 selected from l by the algorithm; and A_1 = the corresponding design matrix.

Clearly, the residuals of the u selected observations are zero and they are the observations that are assumed not to contain blunders. This assumption is valid from the practical point of view given that blunders normally are in the

small minority. The residuals of the remaining observations (not included in I_1) and their covariance matrix are respectively equal to the following:

$$\tilde{\mathbf{v}}_2 = \mathbf{I}_2 - \mathbf{A}_2 \tilde{\mathbf{x}} = \mathbf{I}_2 - \mathbf{A}_2 (\mathbf{A}_1)^{-1} \mathbf{I}_1 \quad (6.59)$$

and

$$\mathbf{C}_{\tilde{\mathbf{v}}_2} = \mathbf{C}_{I_2} + \mathbf{A}_2 (\mathbf{A}_1)^{-1} \mathbf{C}_{I_1} [\mathbf{A}_2 (\mathbf{A}_1)^{-1}]^T - \mathbf{A}_2 (\mathbf{A}_1)^{-1} \mathbf{C}_1 - \mathbf{C}_{I_2} [\mathbf{A}_2 (\mathbf{A}_1)^{-1}]^T \quad (6.60)$$

When the two groups of observations are uncorrelated (i.e. $\mathbf{C}_{I_2} = 0$), we have

$$\mathbf{C}_{\tilde{\mathbf{v}}_2} = \mathbf{C}_{I_2} + \mathbf{A}_2 (\mathbf{A}_1)^{-1} \mathbf{C}_{I_1} [\mathbf{A}_2 (\mathbf{A}_1)^{-1}]^T \quad (6.61)$$

The $\tilde{\mathbf{v}}$ can be considered as the predicted residuals based on the robust solution $\tilde{\mathbf{x}}$ and these observations may contain blunders. The blunder detection scheme is described next. A global test statistic is formulated as follows:

$$H_0 : \mathbf{E}(\tilde{\mathbf{v}}_2) = \mathbf{0} \quad H_a : \mathbf{E}(\tilde{\mathbf{v}}_2) \neq \mathbf{0}$$

$$T' = \tilde{\mathbf{v}}_2^T \mathbf{C}_{\tilde{\mathbf{v}}_2}^{-1} \tilde{\mathbf{v}}_2 \Big|_{H_0} \sim \chi^2(0) \quad (6.62)$$

$$T' = \tilde{\mathbf{v}}_2^T \mathbf{C}_{\tilde{\mathbf{v}}_2}^{-1} \tilde{\mathbf{v}}_2 \Big|_{H_a} \sim \chi^2(\lambda_2^2) \quad (6.63)$$

The local test statistic is as follows:

$$H_0 : \mathbf{E}[(\tilde{\mathbf{v}}_2)_i] = 0 \quad H_a : \mathbf{E}[(\tilde{\mathbf{v}}_2)_i] = (\nabla_2)_i$$

$$w'_i = \frac{(\tilde{\mathbf{v}}_2)_i}{\sqrt{(\mathbf{C}_{\tilde{\mathbf{v}}_2})_{ii}}} \Big|_{H_0} \sim N(0,1) \quad (6.64)$$

$$w'_i = \frac{(\tilde{\mathbf{v}}_2)_i}{\sqrt{(\mathbf{C}_{\tilde{\mathbf{v}}_2})_{ii}}} \Big|_{H_a} \sim N(\delta_i, 1) \quad (6.65)$$

where

$$\lambda_2^2 = \nabla_2^T \mathbf{C}_{\tilde{\mathbf{v}}_2}^{-1} \nabla_2 \quad (6.66)$$

$$\delta_i = \frac{(\nabla_2)_i}{\sqrt{(\mathbf{C}_{\tilde{\mathbf{v}}_2})_{ii}}} \quad (6.67)$$

In short, the advantages of the aforementioned test procedure can be summarized as follows:

- 1) The algorithm provides a robust basis for the calculation for the residuals, i.e., the predicted residual vector \mathbf{v}_2 is robust in that it represents the actual discrepancies of the observations. Subsequently, the test statistic becomes a robust statistic and thus the detection and isolation of blunders can be greatly improved as shown in the examples herein.
- 2) The method turns the problem of multiple-blunder detection into a one-dimensional test procedure. This avoids the difficulty of having to specify beforehand the possible blunders in the context of alternative hypothesis testing used by the conventional methods. On the other hand, the method also simplifies the test procedure as only one global test and a one-dimensional test are involved.

REFERENCES

- Teunissen, P.J.G. and M.A. Salzmann (1989). A Recursive Slippage Test for Use in State-Space Filtering. *Manuscripta Geodetica*, Vol. 14, No. 6.
- Gao, Y., Krakiwsky, E.J., and Czompo, J. (1992). "Robust Testing Procedure for Detection of Multiple Blunders", *Journal of Surveying Engineering*, ASCE, Vol. 118, No. 1.
- Gao, Y. (1992). *A Robust Quality Control System for GPS Navigation and Kinematic Positioning*, PhD Thesis, Department of Geomatics Engineering, The University of Calgary, Calgary, Alberta.