# Contents

Acknowledgments ........................................... 1

Introduction ................................................. 2

1 Observations and Random Errors ......................... 6
  1.1 Univariate Case ......................................... 6
    1.1.1 Expectation and Variance Propagation .............. 7
    1.1.2 Mean Squared Error ................................ 8
  1.2 Multivariate Case ....................................... 9
    1.2.1 The Cauchy-Schwartz Inequality and the Correlation Matrix 11

2 The Model of Direct Observations .......................... 12
  2.1 Model Definition ........................................ 12
    2.1.1 Terminology ......................................... 13
  2.2 The Least-Squares Solution .............................. 15
  2.3 Best Linear Uniformly Unbiased Estimate ............... 17
  2.4 Estimated Variance Component ......................... 19
  2.5 Effects of a Wrongly Chosen Weight Matrix .......... 19
    2.5.1 Effect on the Parameter Estimate .................. 19
    2.5.2 Effect on the Cofactor Matrix for the Estimated Parameter 20
    2.5.3 Effect on the Estimated Variance Component ..... 20
    2.5.4 Effect on the Estimated Dispersion ............... 21
  2.6 Summary Formulas ...................................... 21

3 The Gauss-Markov Model .................................... 24
  3.1 The Least-Squares Solution Within the GMM ............ 25
    3.1.1 Correlation of Adjusted Observations and Predicted Residuals 27
    3.1.2 $P$-Weighted Norm of the Residual Vector .......... 28
  3.2 Estimated Variance Component ......................... 28
  3.3 Summary Formulas ...................................... 30
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>10.4.2</td>
<td>Variance Testing</td>
<td>111</td>
</tr>
<tr>
<td>10.4.3</td>
<td>$F$-distribution</td>
<td>112</td>
</tr>
<tr>
<td>10.5</td>
<td>Hypothesis Testing on the Estimated Parameters</td>
<td>113</td>
</tr>
<tr>
<td>10.6</td>
<td>Checking an Individual Element (or 2-D or 3-D Point) in the Parameter Vector</td>
<td>114</td>
</tr>
<tr>
<td>10.6.1</td>
<td>Non-central $F$-distribution</td>
<td>115</td>
</tr>
<tr>
<td>10.7</td>
<td>Detection of a Single Outlier in the GMM</td>
<td>115</td>
</tr>
</tbody>
</table>

A Useful Matrix Relations and Identities

Bibliography

120

124
Acknowledgments

The compiler of this document would like to thank Dr. Shengjie Ge for his assistance with typesetting an early version with \LaTeX. Also, Mr. Xiankun Wang, MSc., is acknowledged for his careful reading of the text and for pointing out several typographical errors that were corrected in the edition distributed in January, 2010. Dr. Dru Smith suggested additional formulas and the inclusion of summary formulas found at the end of most chapters beginning in 2018. Questions posed by Dr. Jihye Park resulted in improvements to the first edition posted in 2019. A thorough review by Prof. Chris Jekeli in early 2019 resulted in many improvements. Constructive comments from Prof. Clyde Goad were also helpful. Of course, without the extraordinary teaching of Prof. Burkhard Schaffrin, these notes would not even exist.
Introduction

This document is primarily based on notes taken by Kyle Snow in Geodetic Science adjustments courses GS650 and GS651 taught by Burkhard Schaffrin at The Ohio State University in 1997 and 1998. The appendix contains several matrix properties and identities used throughout the text. A bibliography at the end includes referenced material and material for suggested reading.

Notation

A few comments about the notation used in this document may be helpful. Matrices are displayed in uppercase. Vectors are lowercase and are set in bold-face type (bold face is not used for any other symbols). Scalar variables are generally lowercase. Greek letters are used for unknown, non-random parameters, while Latin letters are used for unknown, random variables. Symbols denoting estimates of non-random variables use Greek letters with a hat on top, while predictions of random variables are shown as Latin letters with tildes on top. Tables 1 and 2 list variables, mathematical operators, and abbreviations used herein.

Table 1: Variables and mathematical operators

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$A$</td>
<td>coefficient (design) matrix in the Gauss-Markov Model</td>
</tr>
<tr>
<td>$B$</td>
<td>coefficient matrix in the Model of Condition Equations</td>
</tr>
<tr>
<td>$c$</td>
<td>right-side vector in the system of normal equations $N\xi = c$</td>
</tr>
<tr>
<td>$C{\cdot}$</td>
<td>covariance operator</td>
</tr>
<tr>
<td>$D{\cdot}$</td>
<td>dispersion operator</td>
</tr>
<tr>
<td>$\text{diag}[\cdot]$</td>
<td>a diagonal matrix with diagonal elements comprised of $[\cdot]$</td>
</tr>
<tr>
<td>$\text{dim}$</td>
<td>the dimension of a matrix</td>
</tr>
<tr>
<td>$e$</td>
<td>unknown random error vector for the observations</td>
</tr>
<tr>
<td>$\hat{e}$</td>
<td>predicted random error (residual) vector for the observations</td>
</tr>
</tbody>
</table>

Continued on next page
<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$e_0$</td>
<td>unknown random error vector associated with stochastic constraints</td>
</tr>
<tr>
<td>$\tilde{e}_0$</td>
<td>predicted random error (residual) vector for $e_0$</td>
</tr>
<tr>
<td>$E{\cdot}$</td>
<td>expectation operator</td>
</tr>
<tr>
<td>$H_0$</td>
<td>null hypothesis</td>
</tr>
<tr>
<td>$H_A$</td>
<td>alternative hypothesis</td>
</tr>
<tr>
<td>$K$</td>
<td>constraint matrix used in the Gauss-Markov Model with (stochastic) constraints</td>
</tr>
<tr>
<td>$m$</td>
<td>number of unknown parameters</td>
</tr>
<tr>
<td>$\text{MSE}{\cdot}$</td>
<td>mean squared error operator</td>
</tr>
<tr>
<td>$n$</td>
<td>number of observations</td>
</tr>
<tr>
<td>$N$</td>
<td>normal-equations matrix in the system of normal equations $N\hat{\xi} = c$</td>
</tr>
<tr>
<td>$N(\cdot)$</td>
<td>the nullspace (kernel) of a matrix or the normal distribution, depending on the context</td>
</tr>
<tr>
<td>$P$</td>
<td>weight matrix for the observations</td>
</tr>
<tr>
<td>$P_0$</td>
<td>weight matrix for stochastic constraints</td>
</tr>
<tr>
<td>$q$</td>
<td>rank of the coefficient (design) matrix $A$</td>
</tr>
<tr>
<td>$Q$</td>
<td>cofactor matrix for the observations</td>
</tr>
<tr>
<td>$Q_{\tilde{e}}$</td>
<td>cofactor matrix for the predicted random errors (residuals)</td>
</tr>
<tr>
<td>$r$</td>
<td>redundancy of data model</td>
</tr>
<tr>
<td>$\mathbb{R}$</td>
<td>the field of real numbers</td>
</tr>
<tr>
<td>$\mathbb{R}(\cdot)$</td>
<td>the range (column) space of a matrix</td>
</tr>
<tr>
<td>$\operatorname{rk}$</td>
<td>the rank of a matrix</td>
</tr>
<tr>
<td>$\operatorname{tr}$</td>
<td>the trace of a matrix</td>
</tr>
<tr>
<td>$U$</td>
<td>matrix of eigenvectors</td>
</tr>
<tr>
<td>$w$</td>
<td>constant vector in the Model of Condition Equations</td>
</tr>
<tr>
<td>$y$</td>
<td>vector of observations (possibly in linearized form)</td>
</tr>
<tr>
<td>$z$</td>
<td>vector of constraints used in the Gauss-Markov Model with stochastic constraints</td>
</tr>
<tr>
<td>$\alpha$</td>
<td>significance level for statistical tests</td>
</tr>
<tr>
<td>$\alpha$</td>
<td>observation coefficient vector in the Model of Direct Observations</td>
</tr>
<tr>
<td>$\beta$</td>
<td>a quantity associated with the power of a statistical test</td>
</tr>
<tr>
<td>$\chi^2$</td>
<td>chi-square statistical distribution</td>
</tr>
<tr>
<td>$\delta$</td>
<td>a small deviation or non-random error, as in $\delta P$ denoting a non-random error in matrix $P$</td>
</tr>
<tr>
<td>$\Phi$</td>
<td>Lagrange target function</td>
</tr>
<tr>
<td>$\eta$</td>
<td>unit vector used in the Outlier Detection Model</td>
</tr>
<tr>
<td>$\kappa_0$</td>
<td>vector of specified constants used in the Gauss-Markov Model with constraints</td>
</tr>
<tr>
<td>$\lambda$</td>
<td>unknown vector of Lagrange multipliers</td>
</tr>
<tr>
<td>$\hat{\lambda}$</td>
<td>estimated vector of Lagrange multipliers</td>
</tr>
</tbody>
</table>

*Continued on next page*
<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\mu$, $\mu$</td>
<td>the expected value of a non-random variable, could be a scalar $\mu$ or vector $\mu$</td>
</tr>
<tr>
<td>$\hat{\mu}$, $\hat{\mu}$</td>
<td>the estimate of a non-random variable</td>
</tr>
<tr>
<td>$\bar{y}$</td>
<td>vector of adjusted observations</td>
</tr>
<tr>
<td>$\nu$</td>
<td>statistical degrees of freedom</td>
</tr>
<tr>
<td>$\theta$</td>
<td>the orientation of a confidence ellipse</td>
</tr>
<tr>
<td>$\sigma_0^2$</td>
<td>unknown variance component</td>
</tr>
<tr>
<td>$\hat{\sigma}_0^2$</td>
<td>estimated variance component</td>
</tr>
<tr>
<td>$\Sigma$</td>
<td>dispersion (or covariance) matrix for the observations</td>
</tr>
<tr>
<td>$\tau$</td>
<td>vector of ones (also called “summation vector”)</td>
</tr>
<tr>
<td>$\Omega$</td>
<td>(weighted) sum of squared residuals (unconstrained case)</td>
</tr>
<tr>
<td>$\xi$</td>
<td>vector of unknown parameters</td>
</tr>
<tr>
<td>$\hat{\xi}$</td>
<td>estimated parameter vector</td>
</tr>
</tbody>
</table>

Continued from previous page

Table 2: List of abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>BLUUE</td>
<td>Best Linear Uniformly Unbiased Estimate</td>
</tr>
<tr>
<td>BLIP</td>
<td>Best Linear Prediction</td>
</tr>
<tr>
<td>cdf</td>
<td>cumulative distribution function</td>
</tr>
<tr>
<td>GHM</td>
<td>Gauss-Helmert Model</td>
</tr>
<tr>
<td>GMM</td>
<td>Gauss-Markov Model</td>
</tr>
<tr>
<td>LESS</td>
<td>LEast-Squares Solution</td>
</tr>
<tr>
<td>MSE</td>
<td>Mean Squared Error</td>
</tr>
<tr>
<td>pdf</td>
<td>probability density function</td>
</tr>
<tr>
<td>rms</td>
<td>root mean square</td>
</tr>
</tbody>
</table>
Chapter 1

Observations and Random Errors

In this chapter, we present some properties of random variables, which are also called stochastic variables by some authors (e.g., Bjerhammar, 1973). More particularly, we focus on variables that represent random observation errors. Such errors also have been called accidental errors ibid, pg. 5. Though we cannot know what values random errors will take on, we may state what we expect their values to be, and we may also specify their level of deviation or variance about their expected values.

In the following sections, the notions of expectation and dispersion are defined mathematically. We first start with the univariate case, where only one random error is considered. Then we proceed to the multivariate case, where a vector of $n$ random errors is considered.

1.1 Univariate Case

The univariate case deals with a single random variable, i.e., it treats a scalar quantity rather than a vector quantity. Let us introduce the continuous random variable $e$ with a given probability density function (pdf) $f(e_t)$, where $e_t$ represents a realization of $e$, i.e., a possible value that $e$ might take on (see Chapter 10 for more about pdf’s). The probabilistic mean (or average) of $e$ is the value that we expect $e$ to take on. We denote the expectation of $e$ as $\mu_e$ and define it as follows:

$$\mu_e := E\{e\} = \int_{-\infty}^{\infty} e_t f(e_t) \, de_t, \tag{1.1}$$

where $E$ is called expectation operator. Equation (1.1) is also called the first moment of $e$. If the random variable $e$ represents measurement error, then, ideally, $E\{e\} = 0$. If $E\{e\} \neq 0$, we say that the measurement error is biased.
The dispersion, or variance, of $e$ is denoted by $\sigma_e^2$ and is defined by

$$\sigma_e^2 := E\{(e - \mu_e)^2\} = \int_{-\infty}^{\infty} (e_t - \mu_e)^2 f(e_t) \, de_t. \tag{1.2}$$

Equation (1.2) is also called the second centralized moment of $e$. The dispersion operator, $D\{e\}$, is also used to denote the dispersion (variance) of $e$, but usually we reserve this notation for the multivariate case. The terms dispersion and variance are used interchangeably throughout these notes. The square root of variance is called standard deviation.

Variance is an indicator of how closely the values taken on by the random variable $e$ are to the expected value of $e$. It is reflective of measurement precision and is inversely proportional to it. Thus, a small variance indicates high precision, and a large variance indicates low precision. A succinct expression for the expectation and variance of the random variable $e$, when $e$ is assumed to be unbiased random measurement error, is

$$e \sim (0, \sigma_e^2). \tag{1.3}$$

The expression (1.3) is said in words as “$e$ is distributed with zero mean and sigma-sub-$e$-squared variance.” Note that (1.3) does not specify a pdf for $e$ but only its expectation and dispersion (or variance).

### 1.1.1 Expectation and Variance Propagation

Consider the observation equation

$$y = \mu + e, \ e \sim (0, \sigma_e^2), \tag{1.4}$$

where $y$ is an observation (measurement), $\mu$ is an unknown observable, and $e$ accounts for the random error inherent in the observation $y$. We want to find the expectation and variance of $y$. In other words, we want to know how the expectation and variance propagate from the random variable $e$ to the random variable $y$. Note that $\mu$ is a constant, or non-random, variable. The expectation of a constant is the constant itself; i.e., $E\{\mu\} = \mu$.

For the expectation of $y$, we have

$$E\{y\} = \int_{-\infty}^{\infty} (\mu + e_t) f(e_t) \, de_t, \tag{1.5a}$$

where $e_t$ was defined in the preceding section as a value that the random variable $e$ can take on. The expectation operator is linear. Thus, the expectation of the sum of random variables is the sum of their individual expectations. And, as noted already, $\mu$ is a constant variable. Therefore

$$E\{y\} = \mu \int_{-\infty}^{\infty} f(e_t) \, de_t + \int_{-\infty}^{\infty} e_t f(e_t) \, de_t = \mu + E\{e\} = \mu + 0 = \mu. \tag{1.5b}$$
After introducing \(y_t\), as was done for \(e_t\) above, the dispersion (variance) of \(y\) is defined by

\[
D\{y\} = \int_{-\infty}^{\infty} (y_t - E\{y\})^2 f(y_t) dy_t =
\]

\[
= \int_{-\infty}^{\infty} (\mu + e_t - \mu)^2 f(e_t) de_t = \int_{-\infty}^{\infty} e_t^2 f(e_t) de_t = \sigma_y^2. \quad (1.6)
\]

Another useful formula for the dispersion is derived as follows:

\[
D\{y\} = E\{(y - E\{y\})^2\} =
\]

\[
= E\{y^2 - 2yE\{y\} + E\{y\}^2\} =
\]

\[
= E\{y^2 - 2\mu y + \mu^2\} =
\]

\[
= E\{y^2\} - 2\mu^2 + \mu^2 =
\]

\[
= E\{y^2\} - \mu^2 \Rightarrow \quad (1.7b)
\]

\[
D\{y\} = E\{y^2\} - E\{y\}^2 = \sigma_y^2 \quad (1.7c)
\]

Given constants \(\alpha\) and \(\gamma\), the above formulas for expectation and dispersion can be summarized by the following properties:

\[
E\{\alpha y + \gamma\} = \alpha E\{y\} + \gamma, \quad (1.8a)
\]

\[
D\{\alpha y + \gamma\} = \alpha^2 D\{y\}. \quad (1.8b)
\]

Equation (1.8b) represents the law of error propagation (covariance propagation) in its simplest form. It shows that, in contrast to the expectation, the dispersion operator is not linear. Furthermore, it shows that dispersion is not affected by a constant offset.

### 1.1.2 Mean Squared Error

The mean squared error, or MSE, of \(y\) is the expectation of the square of the difference of \(y\) and its true value \(\mu\). It is defined as

\[
\text{MSE}\{y\} = E\{(y - \mu)^2\} \quad (1.9)
\]

(compare to (1.7a)). It is useful to express the MSE as a combination of the dispersion and a (squared) bias term. This is done via the following derivation:

\[
\text{MSE}\{y\} = E\{(y - \mu)^2\} = E\{\{(y - E\{y\}) - (\mu - E\{y\})\}^2\} =
\]

\[
= E\{(y - E\{y\})^2 - 2(y - E\{y\})(\mu - E\{y\}) + (\mu - E\{y\})^2\} =
\]

\[
= E\{(y - E\{y\})^2\} - 2E\{(y - E\{y\})(\mu - E\{y\})\} + E\{(\mu - E\{y\})^2\}. \quad (1.10)
\]
Note that while \( y \) is a random variable, \( E\{y\} \) is not. So, in the middle term, the expectation operator only applies to \( y \). Therefore, we may continue with

\[
\text{MSE}\{y\} = D\{y\} - 2(E\{y\} - E\{y\})(\mu - E\{y\}) + (\mu - E\{y\})^2 = D\{y\} + \beta^2, \tag{1.11}
\]

where bias is defined formally as

\[
\beta := E\{\mu - y\} = \mu - E\{y\}. \tag{1.12}
\]

Thus, we see that the dispersion of \( y \) and the MSE of \( y \) are only equal in the absence of bias, or in other words, only if indeed \( \mu = E\{y\} \).

We noted previously that dispersion (variance) is an indicator of precision. In contrast, MSE is a measure of accuracy; it includes both dispersion and bias terms. In general, it is harder to meet accuracy standards than precision standards. We can typically increase precision by making more observations (though this may come with additional costs in time and resources); however it may be very difficult to determine the origin of bias.

Finally, we note that the square root of MSE is commonly called rms (root mean square). Thus, strictly speaking, standard deviation and rms are only equivalent in the absence of bias.

## 1.2 Multivariate Case

The multivariate case deals with multiple random variables, which are collected in a column vector. For example, multiple observations of the observable \( \mu \) in (1.4) can be expressed in the following system of equations:

\[
y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = \tau \mu + e = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} \mu + \begin{bmatrix} e_1 \\ \vdots \\ e_n \end{bmatrix}, \tag{1.13}
\]

where \( \tau \) is the “summation vector" defined as \( \tau := [1, \ldots, 1]^T \). In the case of unbiased observations, the expectation of the random error vector \( e \) is written as

\[
E\{ \begin{bmatrix} e_1 \\ \vdots \\ e_n \end{bmatrix} \} = \begin{bmatrix} E\{e_1\} \\ \vdots \\ E\{e_n\} \end{bmatrix} = \begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix}. \tag{1.14}
\]

And, for the dispersion of each element \( e_j \) of \( e \) we have

\[
D\{e_j\} = E\{(e_j - E\{e_j\})^2\} = E\{e_j^2\} = \sigma_j^2. \tag{1.15}
\]

For the multivariate case, we introduce the concept of covariance, which is a measure of similar behavior between random variables, e.g., between elements \( e_j \) and \( e_k \) of \( e \). Formally, the definition of covariance is

\[
C\{e_j, e_k\} = \sigma_{jk} := E\{(e_j - E\{e_j\})(e_k - E\{e_k\})\}. \tag{1.16a}
\]
Obviously,
\[ C\{e_j, e_k\} = C\{e_k, e_j\}. \quad (1.16b) \]
For a random error vector \( e \) having zero expectation, the covariance reduces to
\[ C\{e_j, e_k\} = E\{e_j e_k\}, \quad (1.16c) \]
since \( E\{e_j\} = E\{e_k\} = 0 \). Even though we see from the definition of the covariance (1.16a) that it does not depend on bias, in practice we often find that bias appears as positive correlation (see (1.21) for the definition of correlation coefficient).

Two random variables are said to be independent if their joint probability distribution is equal to the product of their individual probability distributions. Mathematically, this is written as
\[ f\{e_j, e_k\} = f(e_j) f(e_k) \iff e_j \text{ and } e_k \text{ are independent}. \quad (1.17) \]
If two random variables are independent, their covariance is zero. The converse is not true unless the random variables are jointly normally distributed.

In light of the concept of covariance, the dispersion of a vector of random variables is represented by a matrix. The \( j \)th diagonal element of the matrix is denoted by \( \sigma^2_j \) and the \( j, k \) off-diagonal term is written as \( \sigma_{jk} \). The matrix is called a covariance matrix and is represented by \( \Sigma \). Due to (1.16b), the covariance matrix is symmetric. An explicit representation of the covariance matrix \( \Sigma \) is given by
\[
D\{y_{n \times 1}\} = \Sigma_{n \times n} = \begin{bmatrix}
\sigma^2_1 & \sigma_{12} & \cdots & \sigma_{1n} \\
\sigma_{21} & \sigma^2_2 & \cdots & \sigma_{2n} \\
\vdots & \vdots & \ddots & \vdots \\
\sigma_{n1} & \sigma_{n2} & \cdots & \sigma^2_n
\end{bmatrix} = \Sigma^T. \quad (1.18)
\]
Obviously, if the random variables are uncorrelated, the covariance matrix is diagonal.

An important property of a covariance matrix is that it must be at least positive-semidefinite. A matrix is positive-semidefinite if, and only if, all of its eigenvalues are non-negative. For many applications in geodetic science, the covariance matrix is positive-definite, which means that all its eigenvalues are greater than zero. The following statements hold for any positive-definite matrix \( \Sigma \):
- \( \alpha^T \Sigma \alpha = 0 \Rightarrow \alpha = 0 \).
- \( \Sigma \) is a nonsingular matrix (also called a regular matrix).
- All eigenvalues of \( \Sigma \) are positive and non-zero.
- All principle submatrices of \( \Sigma \) are also positive-definite.

In the following chapters, we usually factor out of the covariance matrix \( \Sigma \) a scalar term denoted by \( \sigma^2_0 \), called a variance component, with the resulting matrix called the cofactor matrix. We label the cofactor matrix as \( Q \); its inverse is labeled \( P \) and is called weight matrix. The relations between these terms are written mathematically as
\[ \Sigma = \sigma^2_0 Q = \sigma^2_0 P^{-1}. \quad (1.19) \]
1.2.1 The Cauchy-Schwartz Inequality and the Correlation Matrix

The Cauchy-Schwartz inequality is given by

\[
C(e_j, e_k) = \int \int (e_t_j(e_t) d(e_t)_j, (e_t)_k) f((e_t)_j) d(e_t)_j d(e_t)_k = \sigma_{jk} \leq \sqrt{\int (e_t_j)^2 f((e_t)_j) d(e_t)_j} \int (e_t_k)^2 f((e_t)_k) d(e_t)_k = \sqrt{\sigma_j^2 \sigma_k^2}. \tag{1.20}
\]

The above inequality leads to the notion of a correlation coefficient, defined as

\[
\rho_{jk} := \frac{\sigma_{jk}}{\sqrt{\sigma_j^2 \sigma_k^2}}, \text{ with } -1 \leq \rho_{jk} \leq 1. \tag{1.21}
\]

Analogous to the covariance matrix, we may form a matrix of correlation coefficients. Such a matrix is called a correlation matrix and is defined as

\[
R_{n \times n} := \begin{bmatrix}
1 & \rho_{12} & \ldots & \rho_{1n} \\
\rho_{21} & 1 & \ldots & \rho_{2n} \\
\vdots & \vdots & \ddots & \vdots \\
\rho_{n1} & \rho_{n2} & \ldots & 1
\end{bmatrix} = R^T. \tag{1.22}
\]

Given a covariance matrix \(\Sigma\), the correlation matrix can be generated easily by

\[
R = \text{diag}\left(\frac{1}{\sigma_1}, \ldots, \frac{1}{\sigma_n}\right) \cdot \Sigma \cdot \text{diag}\left(\frac{1}{\sigma_1}, \ldots, \frac{1}{\sigma_n}\right). \tag{1.23}
\]

A further discussion on observations and random errors is given in section 2.1.1 in the context of data models and least-squares adjustments.
Chapter 2

The Model of Direct Observations

2.1 Model Definition

When an unknown parameter $\mu$ can be observed directly, the Model of Direct Observations can be formed for the data by

$$y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} \mu + e_1 \\ \vdots \\ \mu + e_n \end{bmatrix} = \tau \mu + e,$$  \hspace{1cm} (2.1a)

$$e \sim (0, \sigma_0^2 Q), \quad Q := P^{-1}.\hspace{1cm} (2.1b)$$

The terms in the data model are defined as follows:

$\mathbf{y}$ is a given $n \times 1$ vector of observations with random properties.

$\mu$ is an unknown, non-random parameter to be estimated.

$\tau$ is an $n \times 1$ vector of ones ("summation vector"), i.e., $\tau := [1, \ldots, 1]^T$.

$e$ is an $n \times 1$ vector of unknown, random errors to be predicted.

$Q$ is a given $n \times n$ cofactor matrix associated with $e$. It is symmetric, positive-definite, and non-random.

$P$ is an $n \times n$ positive-definite weight matrix, being the inverse of $Q$.

$\sigma_0^2$ is an unknown, non-random variance component that can be estimated.

Equation (2.1a) is called observation equations, while (2.1b) provides the stochastic model for the random observational errors. Together, these two equations comprise the full data model.
2.1.1 Terminology: Observations, Redundancy, Residuals and Their Minimization

The observation vector $y$ was described above as a given quantity. It is given in the sense that it consists of measurements (observations) that are typically made and recorded in the field. The measurements are considered to be a physical realization of an observable — the quantity ("the thing") being observed. An observable could be a dimension of an element of a physical object or a relationship between its elements, such as an angle between two connected edges of a geodetic network, the end points of which being accessible monuments in the ground. Or, and observable could be a property of an immaterial object, such as the phase of an electromagnetic wave. Another example of an observable is the length of a bridge from a mark scribed in concrete at its beginning to another at its end; then, an associated observation could be a distance measured and recorded with a surveyor’s total station between those two marks. Being a measurement of an observable, an observation is a numerical value with an associated unit of measurement.

Even though the vector of observations $y$ is given, it has random properties due to unavoidable random errors inherent both in making observations and in the instruments used to make them. These random errors are unknown quantities, and they are accounted for in the observation equations (2.1a) by the random error vector $e$. Thus, we can say that we know the value of an observation, but we do not know the value of its random error constituent. However, we have already seen that we can say something about the expected values of the random errors (i.e., $E(e) = 0$). Likewise, a statement can be made about the expectation of the observations, viz.

$$
\mu_y := E\{y\} = E\{\tau \mu + e\} = \tau \mu.
$$

We may think of the vector $\mu_y$ as the vector of true observations, the values of which are unknown, though they can be estimated via

$$
\hat{\mu}_y = \tau \hat{\mu},
$$

where $\hat{\mu}$ is an estimate of the unknown parameter $\mu$. The vector $\hat{\mu}_y$ is called the vector of adjusted observations.

Because the given observations, $y$, contain unknown random errors represented by $e$, we cannot possibly expect that $y$ will equal $\tau \mu$, though we may usually hope that at least $y \approx \tau \mu$. The inequality $y \neq \tau \mu$ should be immediately evident from the symbols, since they imply that $y$ is random and $\tau \mu$ is not. (Recall the use of Latin characters for random variables and Greek characters for non-random variables as discussed on page 2.) The rule eluded to here is that when one side of an equation results in a random quantity, so must the other side. The incongruency reflected in $y \neq \tau \mu$ is rectified in (2.1a) by the addition of $e$ on the right side. But practically speaking, $e$ is not much help, since it is unknown. This is where least-squares adjustment theory and techniques can come to the rescue. For if there are more observations than parameters in the model (i.e., more than one observation for model (2.1)), we have a chance to use also these redundant observations to predict values for $e$ using a predictor derived from the principle of least-squares adjustment (see below for a brief discussion on predictors and estimators).
The number of independent, redundant observations is called the redundancy of the model. Another term for it is degrees of freedom, sometimes abbreviated df in the statistical literature.

The vector of predicted random errors is denoted by \( \hat{e} \), and it is also called the vector of residuals. We note that the sign of \( e \) in the model (2.1) is a matter of convention. Some authors show it as a positive term when placed on the left side of the equation. We prefer to show it as positive on the right side, which seems consistent with the majority of authors over the past twenty years or so. It also seems intuitive to us to think of subtracting the random errors from the observations in order to arrive at the true values, at least analytically. And, the equation \( y - e = \tau \mu \) is consistent with the above stated rule that when the right side of an equation results in a random quantity, so must the left side. Here, subtracting \( e \) from \( y \) strips \( y \) of its randomness, resulting in the non-random quantity \( \tau \mu = \mu_y \).

The idea behind least-squares adjustment is to predict the residuals so that the (weighted) sum of their squares is minimized, while still satisfying the relation shown in (2.1a), but now with the predicted random errors \( \hat{e} \) and the estimated parameter \( \hat{\mu} \) rather than their corresponding “true,” but unknown, quantities. That is, the relation

\[
y = \tau \hat{\mu} + \hat{e}
\]

must hold after the adjustment, and the (weighted) sum of squared residuals \( \Omega := \hat{e}^T P \hat{e} \) must be as small as possible. That is both the objective and the outcome of least-squares adjustments.

It is the data that are being adjusted in least-squares adjustments. They are adjusted so that the inconsistent equation \( y \neq \tau \mu \) is replaced by the consistent equation \( \hat{\mu}_y = \tau \hat{\mu} \). So, we speak of adjusted data, predicted residuals, and estimated parameters as the outcomes of a least-squares solution, which is derived in the next section.

**Estimate vs. estimator** In these notes we have hardly distinguished between the terms estimate and estimator. This is partly because sometimes the same symbol works for both terms depending on the context, though perhaps sometimes we should have used estimator when we used estimate, and the same can be said for predictor and prediction. Rather than define these terms in our own words, we prefer to quote directly from Tukey (1987, p. 633).

An estimator is a function of the observations, a specific way of putting them together. It may be specified by an arithmetic formula, like \( \bar{y} = \Sigma x_i / n \), or by words alone, as in directions for finding a sample median by ordering and counting. We distinguish between the estimator and its value, an estimate, obtained from the specific set of data. The variance estimator, \( s^2 = \Sigma(x_i - \bar{x})^2 / (n - 1) \), yields the estimate 7 from the three observations 2, 3, 7. We say \( s^2 \) is an estimator for \( \sigma^2 \), and we call \( \sigma^2 \) the estimand. In the numerical example, 7 estimates \( \sigma^2 \).
2.2 The Least-Squares Solution

In order to minimize $\Omega := \hat{e}^T P \hat{e}$ while satisfying (2.3) we form the Lagrange target function

$$\Phi(e, \lambda, \mu) := e^T P e + 2\lambda^T (y - \tau \mu - e),$$

(2.4)

where $\lambda$ is an unknown $m \times 1$ vector of Lagrange multipliers. The target function is made stationary with respect to the unknown terms $e$, $\lambda$, and $\mu$ when its first partial derivatives are set equivalent to zero, which is reflected in the following Euler-Lagrange necessary conditions:

$$\frac{1}{2} \frac{\partial \Phi}{\partial e} = \frac{1}{2} \left[ \frac{\partial \Phi}{\partial e_j} \right]_{n \times 1} = P \hat{e} - \hat{\lambda} \doteq 0,$$  \hspace{1cm} (2.5a)

$$\frac{1}{2} \frac{\partial \Phi}{\partial \lambda} = \frac{1}{2} \left[ \frac{\partial \Phi}{\partial \lambda_j} \right]_{n \times 1} = y - \tau \hat{\mu} - \hat{e} \doteq 0,$$  \hspace{1cm} (2.5b)

$$\frac{1}{2} \frac{\partial \Phi}{\partial \mu} = \tau^T \hat{\lambda} \doteq 0.$$  \hspace{1cm} (2.5c)

These necessary conditions are sometimes called first-order conditions due to the involvement of first partial derivatives. The sufficient condition for minimization is satisfied by the fact that the second partial derivative of $\Phi$ is $\frac{\partial^2 \Phi}{\partial e \partial e^T} = 2P$, where the weight matrix $P$ is positive definite according to (2.1). Therefore, the solution to the system of equations (2.5) yields the minimum of $\Phi$, and thus the weighted sum of squared residuals (weighted SSR) $\Omega = \hat{e}^T P \hat{e}$ is minimized. See Appendix A for comments on derivatives of quadratic functions with respect to column vectors.

Throughout these notes, we use a hat to denote an estimate of a non-random variable, whereas a tilde denotes a prediction of a random variable. The hat and tilde marks were introduced into (2.5) to distinguish between the unknown variables of the target function (2.4) and those particular quantities that satisfy the necessary conditions. This reflects that $\hat{e}$, $\hat{\lambda}$, and $\hat{\mu}$ cannot take on just any values but rather only those that result from setting the first partial derivatives of the target function to zero (denoted by the \doteq sign), which explains why it would not be logical to introduce the hat and tilde symbols in (2.4). Also note that, for the vector $\hat{e}$, we use the terms residual and predicted random error synonymously.

Now we must solve the system of equations (2.5) to obtain the least-squares solution (LESS) as follows:

$$\hat{\lambda} = P \hat{e} = P(y - \tau \hat{\mu})$$  \hspace{1cm} using (2.5a) and (2.5b) \hspace{1cm} (2.6a)

$$\tau^T \hat{\lambda} = \tau^T P y - (\tau^T P \tau) \hat{\mu} = 0$$  \hspace{1cm} using (2.6a) and (2.5c) \hspace{1cm} (2.6b)

Equations (2.6a) and (2.6b) lead to

$$\hat{\mu} = \frac{\tau^T P y}{\tau^T P \tau},$$  \hspace{1cm} (2.7)
for the estimate of the unknown parameter \( \mu \). And, from (2.5b), we have

\[
\tilde{e} = y - \tau \hat{\mu}
\]  
(2.8)

for the prediction of the random error vector \( e \). As stated already, the prediction \( \tilde{e} \) is also called residual vector. We say that the quantities \( \hat{\mu}, \tilde{e}, \) and \( \hat{\lambda} \) belong to a Least-Squares Solution (LESS) within the Model of Direct Observations (2.1).

The vectors \( \tau \) and \( \tilde{e} \) are said to be \( P \)-orthogonal since

\[
\tau^TP\tilde{e} = \tau^T[P(y - \tau \hat{\mu})] = \tau^T[P[I_n - \tau(\tau^TP\tau)^{-1}\tau^TP]y = \tau^T[Py] - \tau^TP\tau(\tau^TP\tau)^{-1}\tau^TPy = 0. \]  
(2.9)

The adjusted observations, \( \tau \hat{\mu} \), on the right side of (2.8) can also be expressed as

\[
\hat{\mu}_y := \tau \hat{\mu} = y - \tilde{e}. \]  
(2.10)

Obviously, since \( \tau^TP\tilde{e} = 0 \), we also have

\[
(\tau \hat{\mu})^TP\tilde{e} = \hat{\mu}_y^TP\tilde{e} = 0. \]  
(2.11)

Equation (2.11) reveals an important characteristic of LESS; viz., the vector of adjusted observations and the vector of \( P \)-weighted residuals are orthogonal to one another. From a geometric point of view, the orthogonal relationship between these vectors means that the vector of observations \( y \) and the vector of adjusted observations \( \hat{\mu}_y \) are as close as possible to each other (considering the weights in \( P \)), which is exactly what we require from a least-squares adjustment: a minimum adjustment of the data that will satisfy the given model.

In addition to solving for the estimated parameter \( \hat{\mu} \) and the predicted random error vector \( \tilde{e} \), we are typically interested in their dispersions (variances), which are an indicator of their precisions. To compute their dispersions, we apply the law of covariance propagation. First, for the dispersion of the estimated parameter \( \hat{\mu} \) we have

\[
D\{\hat{\mu}\} = \frac{\tau^TP}{\tau^T\tau}D\{y\} \frac{P\tau}{\tau^T\tau} = \frac{\tau^TP(\sigma_0^2P^{-1})P\tau}{\tau^T\tau} \Rightarrow
\]  

\[
D\{\hat{\mu}\} = \frac{\sigma_0^2}{\tau^T\tau}. \]  
(2.12)

The \( n \times n \) dispersion matrix for the residual vector \( \tilde{e} \) is derived by

\[
D\{\tilde{e}\} = D\{[I_n - \tau(\tau^TP\tau)^{-1}\tau^TP]y\} = [I_n - \tau(\tau^TP\tau)^{-1}\tau^T]D\{y\}[I_n - P\tau(\tau^TP\tau)^{-1}\tau^T] = \sigma_0^2[P^{-1} - \tau(\tau^TP\tau)^{-1}\tau^T][I_n - P\tau(\tau^TP\tau)^{-1}\tau^T] = \sigma_0^2[P^{-1} - \tau(\tau^TP\tau)^{-1}\tau^T] - \sigma_0^2\tau(\tau^TP\tau)^{-1}\tau^T + \sigma_0^2\tau(\tau^TP\tau)^{-1}\tau^TP\tau(\tau^TP\tau)^{-1}\tau^T \Rightarrow
\]
\[
D(\hat{e}) = \sigma^2_0 [P^{-1} - \tau (\tau^T P \tau)^{-1} \tau^T].
\]  

(2.13)

It turns out that the last matrix in (2.13) involves the dispersion of the adjusted observations, since

\[
D(\hat{\mu}_y) = \tau D(\hat{\mu}) \tau^T = \sigma^2_0 \tau (\tau^T P \tau)^{-1} \tau^T.
\]  

(2.14)

Formally, neither (2.12) nor (2.13) nor (2.14) can be computed, since the variance component \(\sigma^2_0\) is unknown, though it can be replaced by its estimate shown in (2.22).

From (2.13) we see that the dispersion (variance) of the \(j\)th element of \(\hat{e}\) is

\[
\sigma^2_{\hat{e}_j} = \sigma^2_0 \left( \sigma^2_{jj} - \frac{1}{\tau^T P \tau} \right),
\]  

(2.15)

where \(\sigma^2_{jj}\) is the \(j\)th diagonal element of \(P^{-1}\), and \(\sigma^2_0\) is the variance component from the model (2.1). Thus it is apparent that the variance of the \(j\)th element of the residual vector \(\hat{e}\) is smaller than the variance of the corresponding \(j\)th element of the true, but unknown, random error vector \(e\).

### 2.3 Best Linear Uniformly Unbiased Estimate

Here we take a statistical approach to estimating the unknown parameter \(\mu\). We want to find an estimate for \(\mu\), expressed as a linear combination of the observations \(y\), that extracts the “best” information from the data. The estimate is denoted by \(\hat{\mu}\) and is characterized as the Best Linear Uniformly Unbiased Estimate (BLUUE) of \(\mu\).

1. **Linear criterion:** The linear criterion states that the estimated parameter must be a linear combination of the data contained in \(y\), i.e.

   \[ \hat{\mu} = \alpha^T y, \]

   where \(\alpha\) is an unknown vector to be determined.

2. **Uniformly Unbiased criteria:** An unbiased estimator is one for which its expectation is equal to the true, but unknown, quantity it estimates.

   \[ \mu = E(\hat{\mu}) = E(\alpha^T y) = \alpha^T E(y) = \alpha^T E(\tau \mu + e) = \alpha^T \tau \mu, \]

   for any \( \mu \in \mathbb{R} \leftrightarrow \alpha^T \tau = 1. \)

   Requiring this condition to hold for any \( \mu \in \mathbb{R} \) satisfies the “uniform” criterion, whereas the requirement that \(\alpha^T \tau = 1\) satisfies the “unbiased” criterion.

3. **Best criterion:** The best criterion requires minimum MSE(\(\hat{\mu}\)), or, equivalently, minimum dispersion, since \(\hat{\mu}\) is unbiased. Mathematically, the criterion reads

   \[ \min D(\hat{\mu}), \text{ where } D(\hat{\mu}) = D(\alpha^T y) = \alpha^T D(y) \alpha = \sigma^2_0 \alpha^T Q \alpha, \]

   subject to \(\tau^T \alpha = 1.\)
Accordingly, a Lagrange target function is formed by

\[ \Phi(\alpha, \lambda) := \alpha^T Q \alpha + 2\lambda (\tau^T \alpha - 1). \]  

(2.16)

The necessary conditions for stationarity are provided by the Euler-Lagrange equations, which are written as

\[ \frac{1}{2} \frac{\partial \Phi}{\partial \alpha} = Q \hat{\alpha} + \tau \hat{\lambda} \equiv 0, \]  

(2.17a)

\[ \frac{1}{2} \frac{\partial \Phi}{\partial \lambda} = \tau^T \hat{\alpha} - 1 \equiv 0. \]  

(2.17b)

The sufficient condition for minimization is satisfied by \( \partial^2 \Phi / (\partial \alpha \partial \alpha^T) = 2Q \), which is a positive definite matrix according to (2.1). Solving (2.17a) and (2.17b) simultaneously yields

\[ \hat{\alpha} = -Q^{-1} \tau = -P \tau \hat{\lambda} \]  

using (2.17a),

\[ 1 = \tau^T \hat{\alpha} = -\tau^T P \tau \hat{\lambda} \Rightarrow \hat{\lambda} = \frac{-1}{\tau^T P \tau} \]  

using (2.17b) and (2.18a).

Substituting (2.18b) into (2.18a) we get

\[ \hat{\alpha} = (\tau^T P \tau)^{-1} P \tau. \]  

(2.18c)

Finally, substituting the transpose of (2.18c) into the linear requirement \( \hat{\mu} = \alpha^T y \) yields the BLUUE of \( \mu \) as

\[ \hat{\mu} = \frac{\tau^T P y}{\tau^T P \tau}. \]  

(2.19)

Equation (2.19) agrees with (2.7) derived for LESS. Thus we see that the LESS and the BLUUE are equivalent within the Model of Direct Observations.

We may also prove mathematically that (2.19) fulfills the weighted LESS principle by showing that the \( P \)-weighted residual norm for any other solution is larger than that obtained via BLUUE, which we do in the following: Suppose \( \tilde{\mu} \) is any other estimate for \( \mu \), then

\[ \tilde{e}^T P \tilde{e} = (y - \tau \tilde{\mu})^T P (y - \tau \tilde{\mu}) = \]

\[ = \left[ (y - \tau \tilde{\mu}) - \tau (\hat{\mu} - \tilde{\mu}) \right]^T P \left[ (y - \tau \tilde{\mu}) - \tau (\hat{\mu} - \tilde{\mu}) \right] = \]

\[ = (y - \tau \hat{\mu})^T P (y - \tau \hat{\mu}) - 2(\hat{\mu} - \tilde{\mu}) \tau^T P (y - \tau \tilde{\mu}) + (\tau^T P \tau) (\hat{\mu} - \tilde{\mu})^2 = \]

\[ = (y - \tau \hat{\mu})^T P (y - \tau \hat{\mu}) + (\tau^T P \tau) (\hat{\mu} - \tilde{\mu})^2 \geq \]

\[ \geq (y - \tau \tilde{\mu})^T P (y - \tau \tilde{\mu}) \]

Q.E.D

We have used the \( P \)-orthogonality relation (2.9) in the third line of the proof.
2.4 Estimated Variance Component

The variance component \( \sigma_0^2 \) is an unknown quantity in model (2.1), though in practice its expected value may often be one. The variance component can be estimated as a function of the \( P \)-weighted norm \( \Omega \) of the residual vector \( \tilde{e} \), which is also called the sum of squared residuals (SSR). It is defined by

\[
\Omega := \tilde{e}^T P \tilde{e},
\]

which, together with the degrees of freedom (also called the redundancy of the model)

\[
r := n - 1,
\]

leads to the formula

\[
\hat{\sigma}_0^2 := \frac{\tilde{e}^T P \tilde{e}}{n - 1}
\]

for the estimated variance component. Though it is not derived here, the estimated variance component is derived within the more general Gauss-Markov Model in Section 3.2.

2.5 Effects of a Wrongly Chosen Weight Matrix in the Model of Direct Observations

Assume that the weight matrix \( P \) has been wrongly chosen by an amount \( \delta P \), where \( \delta P \) is assumed to be a small, positive semi-definite matrix that is uncorrelated with \( P \). Consequently, we have

\[
(P \rightarrow (P + \delta) \Rightarrow \hat{\mu} \rightarrow (\hat{\mu} + \delta\hat{\mu}) \quad \text{and} \quad D\{\hat{\mu}\} \rightarrow D\{\hat{\mu} + \delta\hat{\mu}\} \quad \text{and} \quad \delta\sigma_0^2 \rightarrow \delta\sigma_0^2 + \delta\hat{\sigma}_0^2.
\]

2.5.1 Effect on the Parameter Estimate

The following shows the effect of a wrongly chosen weight matrix on the estimated parameter \( \hat{\mu} \):

\[
(\hat{\mu} + \delta\hat{\mu}) = \frac{\tau^T (P + \delta P)y}{\tau^T (P + \delta P)\tau} \Rightarrow \\
\delta\mu = \frac{\tau^T (P + \delta P)y}{\tau^T (P + \delta P)\tau} \hat{\mu} = \frac{\tau^T (P + \delta P)y}{\tau^T (P + \delta P)\tau} \cdot \frac{\tau^T P\tau}{\tau^T P\tau} - \left( \frac{\tau^T P\tau}{\tau^T P\tau} \right) \frac{\tau^T (P + \delta P)\tau}{\tau^T (P + \delta P)\tau} = \\
= \frac{\tau^T P\tau \tau^T P\tau}{\tau^T (P + \delta P)\tau} + \tau^T \delta P y \tau^T P\tau - \tau^T P \tau \tau^T \delta P\tau - \tau^T P \tau \tau^T \delta P\tau = \\
= \frac{\tau^T \delta P y}{\tau^T (P + \delta P)\tau} - \frac{\tau^T \delta P \tau \hat{\mu}}{\tau^T (P + \delta P)\tau} = \frac{\tau^T \delta P (y - \tau \hat{\mu})}{\tau^T (P + \delta P)\tau}.
\]
Finally, we arrive at
\[ \delta \mu = \frac{\tau^T \delta P}{\tau^T (P + \delta P) \tau} \hat{e}. \] (2.23)

2.5.2 Effect on the Cofactor Matrix for the Estimated Parameter

The following shows the effect of a wrongly chosen weight matrix on the cofactor matrix \( Q_\mu \) for the estimated parameter \( \hat{\mu} \), where \( D(\hat{\mu}) = \sigma^2_0 Q_\mu \) is the dispersion of \( \hat{\mu} \):

\[
\begin{align*}
\delta Q_\mu &= (Q_\mu + \delta Q_\mu) - Q_\mu = \frac{1}{\tau^T (P + \delta P) \tau} - \frac{1}{\tau^T P \tau} \\
&= \frac{\tau^T P \tau - \tau^T (P + \delta P) \tau}{(\tau^T P \tau) \tau^T (P + \delta P) \tau} = -\frac{\tau^T \delta P \tau}{(\tau^T P \tau)^2}.
\end{align*}
\]

Thus we have
\[ \delta Q_\mu = -\frac{\tau^T \delta P \tau}{\tau^T (P + \delta P) \tau} Q_\mu. \] (2.24)

2.5.3 Effect on the Estimated Variance Component

The following shows the effect of a wrongly chosen weight matrix on the estimated variance component: First note that

\[
\begin{align*}
e^T P \hat{e} &= (y^T - \hat{\mu}^T) P (y - \tau \hat{\mu}) \\
&= y^T P (y - \tau \hat{\mu}) - \hat{\mu} (\tau^T P y = \tau^T (P + \delta P) y \tau^T P \tau) \\
&= y^T P y - y^T P \tau \hat{\mu} = y^T P y - \tau^T P y \hat{\mu} = y^T P y - \hat{\mu}^2 \tau^T P \tau.
\end{align*}
\]

Following the above logic, we have
\[
(n - 1)(\delta^2_0 + \delta^2_\mu) = y^T (P + \delta P) y - \tau^T (P + \delta P) y (\hat{\mu} + \delta \hat{\mu}) \Rightarrow
\]
\[
(n - 1) \delta^2_\mu = y^T (P + \delta P) y - \tau^T (P + \delta P) y (\hat{\mu} + \delta \hat{\mu}) - y^T P y (\tau^T P \tau)^{-1} \tau^T P \tau.
\]

(Note: the last term will cancel one of the four terms in the binomial product.)

\[
\begin{align*}
&= y^T (\delta P) y - \tau^T (P + \delta P) y (\hat{\mu} + \delta \hat{\mu}) - (\tau^T P y) \delta \hat{\mu} = \\
&= y^T (\delta P) y - \hat{\mu} \tau^T (P + \delta P) y - \tau^T (P + \delta P) y \delta \hat{\mu} = \\
&= (y^T - \hat{\mu} \tau^T)(\delta P) y - \tau^T (P + \delta P) y \delta \hat{\mu} = \\
&= \hat{e}^T (\delta P) y - \tau^T (P + \delta P) y \tau^T (\delta P) \hat{e} =
\end{align*}
\]

(Note that the previous results for \( \delta \hat{\mu} \) have been substituted in the line above.)

\[ = y^T (\delta P) \hat{e} - (\hat{\mu} + \delta \hat{\mu}) \tau^T (\delta P) \hat{e} =
\]
\[
(\text{Using } y^T (\delta P) \tilde{e} = (\hat{\mu} \tau^T + \tilde{e}^T) \delta P \tilde{e} = \tilde{e}^T \delta P \tilde{e} + \hat{\mu} \tau^T \delta P \tilde{e})
\]

\[
= \tilde{e}^T (\delta P) \tilde{e} - \delta \hat{\mu} \tau^T (\delta P) \tilde{e} =
\]

\[
= \tilde{e}^T (\delta P) \tilde{e} - (\delta \hat{\mu})^2 \tau^T (P + \delta P) \tau
\]

Finally, we arrive at

\[
\delta \hat{\sigma}^2_0 = \frac{1}{n - 1} [\tilde{e}^T (\delta P) \tilde{e} - (\delta \hat{\mu})^2 \tau^T (P + \delta P) \tau]. \tag{2.25}
\]

### 2.5.4 Effect on the Estimated Dispersion

The effect of a wrongly chosen weight matrix on the estimated dispersion of \( \hat{\mu} \) is obviously given by

\[
D \{ \hat{\mu} + \delta \hat{\mu} \} = (\delta \hat{\sigma}^2_0 + \delta \hat{\sigma}^2_0) \cdot D \{ \hat{\mu} + \delta \hat{\mu} \} = (\delta \hat{\sigma}^2_0 + \delta \hat{\sigma}^2_0) \cdot (Q_{\hat{\mu}} + \delta Q_{\hat{\mu}}). \tag{2.26}
\]

### 2.6 Summary Formulas for the Least-Squares Solution Within the Model of Direct Observations

The Model of Direct Observations is given by

\[
\begin{bmatrix}
y_1 \\
\vdots \\
y_n
\end{bmatrix}
= \begin{bmatrix}
\mu + e_1 \\
\vdots \\
\mu + e_n
\end{bmatrix}
= \tau \mu + e,
\]

\[
e \sim (0, \sigma^2_0 Q), \quad Q := P^{-1}.
\]

Table 2.1: Summary formulas for the LESS within the Model of Direct Observations

<table>
<thead>
<tr>
<th>Quantity</th>
<th>Formula</th>
<th>Eq.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model redundancy</td>
<td>( r = n - 1 )</td>
<td>(2.21)</td>
</tr>
<tr>
<td>Estimated parameter</td>
<td>( \hat{\mu} = (\tau^T P y) / (\tau^T P \tau) )</td>
<td>(2.7)</td>
</tr>
<tr>
<td>Dispersion of estimated</td>
<td>( D { \hat{\mu} } = \sigma^2_0 / (\tau^T P \tau) )</td>
<td>(2.12)</td>
</tr>
<tr>
<td>parameter</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Vector of predicted residuals</td>
<td>( \tilde{e} = y - \tau \hat{\mu} )</td>
<td>(2.8)</td>
</tr>
</tbody>
</table>

Continued on next page
<table>
<thead>
<tr>
<th>Quantity</th>
<th>Formula</th>
<th>Eq.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dispersion matrix for residuals</td>
<td>$D(\tilde{e}) = \sigma_0^2 \cdot [P^{-1} - \tau \tau^T P \tau^{-1} \tau^T]$</td>
<td>(2.13)</td>
</tr>
<tr>
<td>Sum of squared residuals (SSR)</td>
<td>$\Omega = \tilde{e}^T P \tilde{e}$</td>
<td>(2.20)</td>
</tr>
<tr>
<td>Estimated variance component</td>
<td>$\sigma_0^2 = (\tilde{e}^T P \tilde{e})/(n - 1)$</td>
<td>(2.22)</td>
</tr>
<tr>
<td>Vector of adjusted observations</td>
<td>$\hat{\mu}_y = y - \tilde{e}$</td>
<td>(2.10)</td>
</tr>
<tr>
<td>Dispersion matrix for adjusted observations</td>
<td>$D(\hat{\mu}_y) = \sigma_0^2 \tau \tau^T P \tau^{-1} \tau^T$</td>
<td>(2.14)</td>
</tr>
</tbody>
</table>

*Continued from previous page*
The Gauss-Markov Model (GMM) is the underlying data model for many of the topics that follow. In presentation of the model, it is assumed that the observation equations (3.1a) have been linearized, if necessary. The model is written as follows:

\[ y = A \xi + e, \quad \text{rk} A = m, \]

\[ e \sim (0, \sigma_0^2 P^{-1}). \]  

In the case of linearization, \( y \) is a vector of observations minus “zeroth-order” terms; \( A \) is a known \( n \times m \) coefficient matrix (also called design or information matrix, or Jacobian matrix if partial derivatives are involved) relating the observations to the unknown parameters; \( \xi \) is a vector of unknown parameters to estimate (corrections to initial values in the case of linearization), and \( e \) is a vector of random observation errors, having zero expectation. Equation (3.1a) requires the \( n \times m \) coefficient matrix \( A \) to have full column rank.

The \( n \times n \) matrix \( P \) is symmetric. It contains weights of the observations, which may be correlated. The inverse of \( P \) shown in (3.1) implies that \( P \) is a positive-definite matrix; this inverse matrix is called the cofactor matrix and is denoted by \( Q \). The symbol \( \sigma_0^2 \) represents a variance component, which is considered unknown but can be estimated. The dispersion matrix \( D\{e\} = \sigma_0^2 P^{-1} \) is called the variance-covariance matrix, or simply the covariance matrix, and is also denoted by \( \Sigma \). In summary, we have the following relation between the dispersion, weight, and cofactor matrices of the unknown, random error vector \( e \):

\[ D\{e\} = \Sigma = \sigma_0^2 Q = \sigma_0^2 P^{-1}. \]  

(3.2)

The redundancy \( r \) of the model (3.1a) is defined as

\[ r := n - \text{rk} A = n - m. \]  

(3.3)
Redundancy is also called degrees of freedom in the context of statistical testing discussed in Chapter 10.

The GMM shown in (3.1) has two main components. The first component, (3.1a), contains the observation equations $y = A\xi + e$, which show the functional relation between the observations, their random errors, and the unknown parameters that are to be estimated. The second component, (3.1b), shows a stochastic model, $e \sim (0, \sigma_0^2 P^{-1})$, which expresses the expectation and dispersion of the random errors. These quantities are also called the first and second moments, respectively, of the random error vector $e$.

If the rank of matrix $A$ is less than the number of unknown parameters to estimate, we say that the problem is rank deficient. Such a problem cannot be solved based on the observations alone; additional information about the unknown parameters must be provided. The problem of rank deficiency is covered in Chapter 6 and, much more thoroughly, in the notes for the advanced adjustments course.

3.1 The Least-Squares Solution Within the Gauss-Markov Model

We now derive the LEast-Squares Solution (LESS) for the parameter estimate $\hat{\xi}$ and the predicted random error (residual) vector $\hat{e}$, with their associated dispersion matrices, under the assumption that the coefficient matrix $A$ has full column rank. For convenience, we define the $m \times m$ matrix $N$ and the $m \times 1$ vector $c$ as

$$[N, c] := A^T P [A, y]. \tag{3.4}$$

The objective of least-squares minimization is to minimize the $P$-weighted sum of squared residuals, or, equivalently, to minimize the $P$-weighted random errors in the model (3.1). Thus the Lagrange target function

$$\Phi(\xi) := (y - A\xi)^T P (y - A\xi) = \text{stationary} \tag{3.5}$$

should be minimized. Forming the the Euler-Lagrange necessary conditions (or first-order conditions) leads directly to the least-squares normal equations

$$\frac{1}{2} \frac{\partial \Phi}{\partial \xi} = (A^T PA) \hat{\xi} - A^T Py = N \hat{\xi} - c = 0. \tag{3.6}$$

The sufficient condition is satisfied by $(1/2) \cdot (\partial^2 \Phi / \partial \xi \partial \xi^T) = N$, which is positive-definite since matrix $A$ has full column rank according to (3.1a). Equation (3.6) leads to the least-squares solution (LESS)

$$\hat{\xi} = N^{-1} c \tag{3.7}$$

for the unknown parameter vector $\xi$, with its expectation computed by

$$E\{\hat{\xi}\} = N^{-1} E\{c\} = N^{-1} A^T P E\{y\} = N^{-1} A^T PA \xi = \xi. \tag{3.8}$$
The predicted random error vector (also called residual vector) is then given by

\[ \tilde{e} = y - \hat{A}\xi = (I_n - AN^{-1}A^TP)y, \]  

with expectation

\[ E\{\tilde{e}\} = (I_n - AN^{-1}A^TP)E\{y\} = (I_n - AN^{-1}A^TP)A\xi = A\xi - A\xi = \mathbf{0}. \]  

The expectation of the given observation vector is expressed as

\[ E\{y\} = \mu_y, \]

where \( \mu_y \) is the true, but unknown, vector of observables. Thus we write the vector of adjusted observations as

\[ \hat{\mu}_y = y - \tilde{e} = A\hat{\xi}, \]

with expectation

\[ E\{\hat{\mu}_y\} = AE\{\hat{\xi}\} = A\xi. \]

Equations (3.8), (3.10) and (3.12) show that the estimated parameters, the residuals, and the adjusted observations, respectively, are unbiased.

The corresponding dispersion matrices are computed by using the law of covariance propagation. The dispersion of the estimated parameters is computed by

\[ D\{\hat{\xi}\} = D\{N^{-1}A^TPy\} = (N^{-1}A^TP)D\{y\}(PAN^{-1}) = N^{-1}A^TP(\sigma_0^2P^{-1})PAN^{-1} \Rightarrow D\{\hat{\xi}\} = \sigma_0^2N^{-1}. \]

And, the dispersion of the residual vector \( \tilde{e} \) is

\[ D\{\tilde{e}\} = (I_n - AN^{-1}A^TP)y(I_n - PAN^{-1}A^T) = (I_n - AN^{-1}A^TP)(\sigma_0^2P^{-1})(I_n - PAN^{-1}A^T) = \sigma_0^2(I_n - AN^{-1}A^TP)(P^{-1} - AN^{-1}A^T) \Rightarrow D\{\tilde{e}\} = \sigma_0^2(P^{-1} - AN^{-1}A^T) = D\{y\} - D\{A\xi\} =: \sigma_0^2Q_{\tilde{e}}, \]

where the matrix

\[ Q_{\tilde{e}} := P^{-1} - AN^{-1}A^T \]

is the cofactor matrix of the residual vector \( \tilde{e} \). Equations (3.14a) to (3.14c) reveal that the variances of the residuals are smaller than the corresponding variances of the observations, since the matrix product \( AN^{-1}A^T \) is positive-definite. Finally, the dispersion of the vector of adjusted observations is computed by

\[ D\{\hat{\mu}_y\} = AD\{\hat{\xi}\}A^T = \sigma_0^2AN^{-1}A^T. \]
Summarizing the above equations, the respective distributions for the estimated parameter vector, the residual vector, and the vector of adjusted observations are succinctly expressed by

\[
\hat{\xi} \sim (\xi, \sigma_0^2 N^{-1}),
\]
\[
\hat{e} \sim (0, \sigma_0^2 [P^{-1} - AN^{-1}A^T]) =: \sigma_0^2 Q \hat{e},
\]
\[
\hat{\mu}_y \sim (A\xi, \sigma_0^2 AN^{-1}A^T).
\]

Since the variance component \( \sigma_0^2 \) is an unknown quantity, the dispersions shown in (3.16) cannot be computed unless either \( \sigma_0^2 \) is estimated or a value is specified for it. In the case where the estimated variance component is used in lieu of the true, but unknown, variance component, we speak of an estimated dispersion matrix for the estimated parameter vector, which is provided by

\[
\hat{D}\{\hat{\xi}\} = \sigma_0^2 N^{-1},
\]

with obvious extension to other quantities, such as \( \hat{D}\{\hat{e}\} \) and \( \hat{D}\{\hat{\mu}_y\} \). See Section 3.2 for the derivation of the variance component estimate \( \hat{\sigma}_0^2 \), the formula for which is given in (3.26).

### 3.1.1 Correlation of Adjusted Observations and Predicted Residuals

Equation (3.14b) implies that the covariance between the vector of adjusted observations \( \hat{\mu}_y = A\hat{\xi} \) and the vector of residuals \( \hat{e} \) is zero. Since, according to (3.7) and (3.9), both vectors are a function of the random vector \( y \), this can also be shown by applying the law of covariance propagation as follows:

\[
C\{A\hat{\xi}, \hat{e}\} = AN^{-1}A^T P \cdot D\{y\} \cdot (I_n - AN^{-1}A^T P)^T =
\]
\[
= \sigma_0^2 [AN^{-1}A^T - AN^{-1}(A^T PA) N^{-1}A^T] =
\]
\[
= \sigma_0^2 [AN^{-1}A^T - AN^{-1}A^T] = 0.
\]

Also, we have the following covariance between the adjusted and original observations:

\[
C\{A\hat{\xi}, y\} = AN^{-1}A^T PD\{y\} = \sigma_0^2 AN^{-1}A^T PP^{-1} =
\]
\[
= \sigma_0^2 AN^{-1}A^T = D\{A\hat{\xi}\}.
\]

Zero correlation does not necessarily imply statistical independence, though the converse does hold. Analogous to (10.9a), the adjusted observations and predicted residuals are not statistically independent unless the expectation of their product is equal to the product of their expectations. The following shows that this property is not satisfied: Since the trace of a scalar product is the scalar product itself, we start with

\[
E\{(A\hat{\xi})^T \hat{e}\} = E\{\text{tr} \hat{\xi}^T A^T (I_n - AN^{-1}A^T P) y\}.
\]
But the trace is invariant with respect to a cyclic transformation (see (A.5)). Thus,

\[ E\{(A\hat{\xi})^T \hat{e}\} = E\{\text{tr}(A^T - A^T AN^{-1} A^T P)y\hat{\xi}^T\} = \]

\[ = \text{tr}(A^T - A^T AN^{-1} A^T P)E\{y\hat{\xi}^T\} \neq 0 = E\{(A\hat{\xi})^T\}E\{\hat{e}\}, \text{ since } E\{\hat{e}\} = 0. \]

### 3.1.2 \(P\)-Weighted Norm of the Residual Vector

The \(P\)-weighted norm of the residual vector \(\hat{e}\) is an important quantity that can be used to check the overall ("global") fit of the adjustment. The norm is defined as

\[ \Omega := \hat{e}^T P \hat{e}, \quad (3.20) \]

and it is guaranteed to be a minimum, since \(\hat{e}\) was obtained by minimizing \(e^T Pe\) (cf. (3.5)). In the special case where \(P = I_n\), the quadratic form \(\Omega\) is often called the sum of squared residuals, or SSR, in the statistical literature. We use the term SSR in the following chapters even when \(P\) is not the identity matrix. Substituting (3.9) into (3.20) leads to some commonly used alternative forms for \(\Omega\).

\[ \hat{e}^T P \hat{e} = (y - A\hat{\xi})^T P(y - A\hat{\xi}) = \]

\[ = y^T Py - y^T P A\hat{\xi} - \hat{\xi}^T A^T P y + \hat{\xi}^T A^T PA\hat{\xi} = \]

\[ = y^T Py - 2\hat{\xi}^T \hat{\xi} + \hat{\xi}^T \hat{\xi} = \quad (3.21a) \]

\[ = y^T Py - c^T \hat{\xi} = \quad (3.21b) \]

\[ = y^T Py - c^T N^{-1} c = \quad (3.21c) \]

\[ = y^T Py - (N\hat{\xi})^T N^{-1} N \hat{\xi} = \]

\[ = y^T Py - \hat{\xi}^T N \hat{\xi} = \quad (3.21d) \]

\[ = y^T (P - PAN^{-1} A^T P) y \quad (3.21e) \]

Note that the target function (3.5) could have been written explicitly as a function of the random error vector \(e\) with the introduction of a vector of Lagrange multipliers \(\lambda\) as follows:

\[ \Phi(e, \xi, \lambda) = e^T Pe - 2\lambda^T (y - A\xi + e) = \text{stationary}. \quad (3.22) \]

This approach leads to the estimate of Lagrange multipliers as \(-\hat{\lambda} = P\hat{e}\) and thus leads to yet another expression for the \(P\)-weighted norm

\[ \Omega = \hat{e}^T P \hat{e} = -\hat{\xi}^T \hat{\lambda} = \hat{\lambda}^T P^{-1} \hat{\lambda}. \quad (3.23) \]

### 3.2 Estimated Variance Component Within the Gauss-Markov Model

As stated in Section 2.4, the variance component \(\sigma_0^2\) is an unknown quantity in the Gauss-Markov Model (GMM). We now present the derivation of the estimated
variance component $\hat{\sigma}_0^2$. As defined in (3.1), the dispersion matrix for the random error vector $e$ is $D\{e\} = \sigma_0^2 Q$. Also, by definition of dispersion we have $D\{e\} = E\{(e - E\{e\})(e - E\{e\})^T\}$. But, for the error vector $E\{e\} = 0$; therefore $D\{e\} = E\{ee^T\} = \sigma_0^2 Q = \sigma_0^2 P^{-1}$.

The following steps lead to an expression for the variance component $\sigma_0^2$ in terms of the quadratic product $e^T Pe$.

\[
E\{ee^T\} = \sigma_0^2 Q \quad \text{(by definition)}
\]
\[
PE\{ee^T\} = \sigma_0^2 I_n \quad \text{(multiply both sides by } P\text{)}
\]
\[
\text{tr} PE\{ee^T\} = \sigma_0^2 \text{tr } I_n = n\sigma_0^2 \quad \text{(apply the trace operator)}
\]
\[
\text{tr} E\{Pee^T\} = n\sigma_0^2 \quad \text{(move the constant matrix } P\text{ into the expectation)}
\]
\[
E\{\text{tr } Pe e^T\} = n\sigma_0^2 \quad \text{(interchange the trace and expectation operators—both linear)}
\]
\[
E\{\text{tr } e^T Pe\} = n\sigma_0^2 \quad \text{(the trace is invariant with respect to a cyclic transformation)}
\]
\[
E\{e^T Pe\} = n\sigma_0^2 \quad \text{(a quadratic product is a scalar; trace of scalar is scalar itself)}
\]
\[
\sigma_0^2 = E\{\frac{e^T Pe}{n}\} \quad \text{(dividing through by } n\text{ and placing } n\text{ inside } E\{\cdot\}\text{)}
\]
\[
\tilde{\sigma}_0^2 = \frac{e^T Pe}{n} \quad \text{(define a symbol for the term inside } E\{\cdot\}\text{)}
\]
\[
E\{\tilde{\sigma}_0^2\} = \sigma_0^2 \quad \text{(by substitution)}
\]

Thus we can say that $(e^T Pe)/n$ provides an unbiased representation of $\sigma_0^2$. However, we do not actually know the true random error vector $e$, but we do know its predicted value $\tilde{e}$.

We now work with the residual vector $\tilde{e}$ to find an unbiased estimate of $\sigma_0^2$. Combining steps similar to those explained above, we can write

\[
E\{\tilde{e}^T P \tilde{e}\} = \text{tr } E\{\tilde{e}^T P \tilde{e}\} = \text{tr } E\{\tilde{e} \tilde{e}^T\} P = \text{tr } D\{\tilde{e}\} P. \quad (3.24)
\]

According to (3.14a), the dispersion of the residual vector is $D\{\tilde{e}\} = \sigma_0^2 (P^{-1} - AN^{-1}A^T)$. Substituting this result into (3.24) gives

\[
E\{\tilde{e}^T P \tilde{e}\} = \text{tr } \sigma_0^2 (P^{-1} - AN^{-1}A^T) P = \sigma_0^2 (\text{tr } I_n - \text{tr } AN^{-1}A^T P) = \sigma_0^2 (\text{tr } I_n - \text{tr } N^{-1}A^T PA) = \sigma_0^2 (n - \text{rk } N) = \sigma_0^2 (n - \text{rk } A). \quad \text{(using } A.5\text{)}
\]

Finally, we arrive at

\[
E\{\frac{\tilde{e}^T P \tilde{e}}{n - \text{rk } A}\} = \sigma_0^2. \quad (3.25)
\]

Now, we simply label the argument of the expectation operator on the left side of (3.25) as $\hat{\sigma}_0^2$, which allows us to write the expression for the estimated variance.
component as

\[ \hat{\sigma}_0^2 = \frac{\tilde{e}^T P \tilde{e}}{n - \text{rk} A}. \]  

(3.26)

Obviously, \( \hat{\sigma}_0^2 \) is a uniformly unbiased estimate of \( \sigma_0^2 \), since

\[ E\{\hat{\sigma}_0^2\} = \sigma_0^2. \]  

In the case of the Model of Direct Observations, we replace \( A \) with \( \tau \), which has rank of 1, and thus we have \( \hat{\sigma}_0^2 := \tilde{e}^T P \tilde{e} / (n - 1) \), which verifies (2.22). Alternative expressions for \( \hat{\sigma}_0^2 \) can be reached by use of (3.21) and (3.23).

The above derivations imply the following relationship between

\[ E\{e^T P e\} \]

and

\[ E\{\tilde{e}^T P \tilde{e}\} \]

\[ n \]

\[ \text{rk} A \]

\[ \Rightarrow \]  

(3.27a)

\[ E\{\tilde{e}^T P \tilde{e}\} < E\{e^T P e\} \]  

(3.27b)

According to Grafarend and Schaffrin (1993, pg. 103), and Schaffrin (1997b), the dispersion, and estimated dispersion, respectively, of \( \hat{\sigma}_0^2 \) are given by

\[ D\{\hat{\sigma}_0^2\} = (n - m)^{-1} \cdot 2(\hat{\sigma}_0^2)^2 \]  

(3.28)

and

\[ \hat{D}\{\hat{\sigma}_0^2\} = (n - m)^{-1} \cdot 2(\hat{\sigma}_0^2)^2, \]  

(3.29)

where it is assumed that \( m = \text{rk} A \).

### 3.3 Summary Formulas for the Least-Squares Solution Within the Gauss-Markov Model

The Gauss-Markov Model with full column-rank coefficient matrix \( A \) is given by

\[ y_{n \times 1} = A_{n \times m} \xi + e, \quad e \sim (0, \sigma_0^2 P^{-1}), \quad \text{rk} A = m. \]

Table 3.1: Summary formulas for the LESS within the Gauss-Markov Model

<table>
<thead>
<tr>
<th>Quantity</th>
<th>Formula</th>
<th>Eq.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model redundancy</td>
<td>( r = n - \text{rk} A )</td>
<td>(3.3)</td>
</tr>
<tr>
<td>Vector of estimated parameters</td>
<td>( \hat{\xi} = N^{-1}c, \quad [N, c] := A^T P[A, y] )</td>
<td>(3.7)</td>
</tr>
</tbody>
</table>

Continued on next page
<table>
<thead>
<tr>
<th>Quantity</th>
<th>Formula</th>
<th>Eq.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dispersion matrix for estimated parameters</td>
<td>$D{\hat{\xi}} = \sigma_0^2 \cdot N^{-1}$</td>
<td>(3.13)</td>
</tr>
<tr>
<td>Vector of predicted residuals</td>
<td>$\hat{e} = y - A\hat{\xi} = (I_n - AN^{-1}A^T P)y$</td>
<td>(3.9)</td>
</tr>
<tr>
<td>Dispersion matrix for residuals</td>
<td>$D{\hat{e}} = \sigma_0^2 \cdot (P^{-1} - AN^{-1}A^T)$</td>
<td>(3.14a)</td>
</tr>
<tr>
<td>Sum of squared residuals (SSR)</td>
<td>$\Omega = \hat{e}^T P \hat{e}$</td>
<td>(3.20)</td>
</tr>
<tr>
<td>Estimated variance component</td>
<td>$\hat{\sigma}^2_0 = (\hat{e}^T P \hat{e})/(n - rk A)$</td>
<td>(3.26)</td>
</tr>
<tr>
<td>Vector of adjusted observations</td>
<td>$\hat{\mu}_y = y - \hat{e}$</td>
<td>(3.11)</td>
</tr>
<tr>
<td>Dispersion matrix for adjusted observations</td>
<td>$D{\hat{\mu}_y} = \sigma_0^2 \cdot AN^{-1}A^T$</td>
<td>(3.15)</td>
</tr>
</tbody>
</table>

Continued from previous page
Chapter 4

The Model of Condition Equations

In the least-squares adjustment within the Model of Condition Equations, the unknown parameters $\xi$ are not estimated directly, rather the random error vector $e$ is predicted. This approach might be taken if the parameters are of no particular interest, or it might be done to make the problem easier to formulate. An example of the latter is the adjustment of leveling networks, where the parameters (heights of the stations) are of primary interest, but because closed “level loops” within the network sum to zero (a necessary condition), it is convenient to difference the observations along these loops before performing the adjustment (see level-loop example in Section 4.3). Another motivation for using the Model of Condition Equations is that the size of the matrix to invert in the least-squares solution (LESS) may be smaller than that in the corresponding LESS within the Gauss-Markov Model (GMM).

Let the $r \times n$ matrix $B$ represent a difference operator such that when it is applied to the $n \times 1$ observation equations $y = A\xi + e$, the parameters are eliminated. More specifically, we require that $BA = 0$, which implies that $By = B(A\xi + e) = Be$. Therefore, by applying the difference operator $B$, the GMM is transformed to the following Model of Condition Equations:

\[
\begin{align*}
\mathbf{w} &:= B y = Be, \quad e \sim (0, \sigma_0^2 P^{-1}), \\
r &:= n - q = \text{rk } B,
\end{align*}
\] (4.1a)

where the variable $r$ denotes the redundancy of the model, and the rank of the $n \times m$ matrix $A$ from the GMM is $\text{rk } A := q \leq m$. Equation (4.1b) requires that matrix $B$ has full row-rank. Moreover, it shows that the redundancy of the model is not changed from that of the GMM by eliminating parameters.
4.1 The Least-Squares Solution Within the Model of Condition Equations

The least-squares criterion for minimizing the \((P\)-weighted, squared) norm of the observation residuals is written as

\[
\min e^T Pe \text{ subject to } w = Be,
\]

for which the Lagrange target function

\[
\Phi(e, \lambda) := e^T Pe + 2\lambda^T (w - Be)
\]

can be written, which must be made stationary with respect to the unknown terms \(e\) and \(\lambda\). Here, \(\lambda\) is an \(r \times 1\) vector of Lagrange multipliers. Taking the first partial derivatives of (4.3) leads to the Euler-Lagrange necessary conditions

\[
\frac{1}{2} \frac{\partial \Phi}{\partial e} = P\tilde{e} - B^T \hat{\lambda} \equiv 0,
\]

\[
\frac{1}{2} \frac{\partial \Phi}{\partial \lambda} = w - B\tilde{e} \equiv 0.
\]

The sufficient condition, required to ensure a minimum is reached, is satisfied by \(\frac{\partial \Phi^2}{\partial e \partial e^T} = 2P\), which is positive definite since the weight matrix \(P\) is invertible according to (4.1a). The simultaneous solution of (4.4a) and (4.4b) leads to the Best Linear Prediction (BLIP) of \(e\) as derived in the following: Equation (4.4a) leads to

\[
\tilde{e} = P^{-1} B^T \hat{\lambda}.
\]

Then, (4.4b) and (4.5a) allows

\[
w = B\tilde{e} = (BP^{-1} B^T) \hat{\lambda} \Rightarrow
\]

\[
\hat{\lambda} = (BP^{-1} B^T)^{-1} w \Rightarrow
\]

\[
\hat{e} = P^{-1} B^T (BP^{-1} B^T)^{-1} w,
\]

finally leading to the predicted random error vector

\[
\hat{e} = P^{-1} B^T (BP^{-1} B^T)^{-1} By.
\]

Note that the matrix product \(BP^{-1} B^T\) results in a symmetric, positive definite matrix of size \(r \times r\), since \(B\) has full row rank. The predicted random error vector \(\hat{e}\) is also called the residual vector. The expectation of the given observation vector is expressed as \(E\{y\} = \mu_y\), where \(\mu_y\) is the true, but unknown, vector of observables. Thus we write the vector of adjusted observations as

\[
\hat{\mu}_y = y - \hat{e}.
\]
The square of the $P$-weighted residual norm $\Omega$, also called the sum of squared residuals (SSR), is computed by

$$\Omega = \hat{e}^T P \hat{e} = \hat{e}^T B^T \hat{\lambda} = w^T \hat{\lambda} = w^T (BP^{-1}B^T)^{-1} w,$$  \hspace{1cm} (4.7)

leading to the estimated variance component

$$\hat{\sigma}^2_0 = \frac{\hat{e}^T P \hat{e}}{r_k B} = \frac{\hat{e}^T P \hat{e}}{r}.$$ \hspace{1cm} (4.8)

In words, it is described as the squared $P$-weighted residual norm divided by the degrees of freedom (redundancy) of the model.

Apply the law of error propagation, the dispersion of the residual vector is computed by

$$D\{\hat{e}\} = P^{-1}B^T(BP^{-1}B^T)^{-1}B \cdot D\{y\} \cdot B^T(BP^{-1}B^T)^{-1}BP^{-1} =$$

$$= P^{-1}B^T(BP^{-1}B^T)^{-1}B(\sigma^2_0 P^{-1})B^T(BP^{-1}B^T)^{-1}BP^{-1} \Rightarrow$$

$$D\{\hat{e}\} = \sigma^2_0 \cdot P^{-1}B^T(BP^{-1}B^T)^{-1}BP^{-1}. \hspace{1cm} (4.9)$$

As we did earlier within the GMM (Section 3.1.1), we compute the covariance between the residual vector $\hat{e}$ and the vector adjusted observations $\hat{\mu}_y = y - \hat{e}$ as follows:

$$C\{\hat{\mu}_y, \hat{e}\} = C\{[I - P^{-1}B^T(BP^{-1}B^T)^{-1}B]y, P^{-1}B^T(BP^{-1}B^T)By\} =$$

$$= [I - P^{-1}B^T(BP^{-1}B^T)^{-1}B] \cdot D\{y\} \cdot [P^{-1}B^T(BP^{-1}B^T)^{-1}B]^T =$$

$$= [I - P^{-1}B^T(BP^{-1}B^T)^{-1}B] \cdot \sigma^2_0 P^{-1} \cdot B^T(BP^{-1}B^T)^{-1}BP^{-1} =$$

$$= \sigma^2_0 [P^{-1}B^T(BP^{-1}B^T)^{-1}BP^{-1} - P^{-1}B^T(BP^{-1}B^T)^{-1}BP^{-1}B^T \cdot (BP^{-1}B^T)^{-1}BP^{-1}] = 0 \hspace{1cm} (4.10)$$

Thus, it has been shown that the residuals and adjusted observations are uncorrelated, and therefore the dispersion of the adjusted observations can be written as

$$D\{\hat{\mu}_y\} = D\{y\} - D\{\hat{e}\} = \sigma^2_0 [P^{-1} - P^{-1}B^T(BP^{-1}B^T)^{-1}BP^{-1}]. \hspace{1cm} (4.11)$$

Note that $B$ is not a unique matrix, but regardless of how $B$ is chosen the results of the adjustment will be the same, provided the following necessary conditions for $B$ are satisfied:

(i) **Dimensionality:** $\text{rk } B = n - \text{rk } A = n - q = r$, which means that $\text{rk } B + \text{rk } A = (n - q) + q = n$.

(ii) **Orthogonality:** $BA = 0$. 

33
4.2 Equivalence Between LESS Within the Gauss-Markov Model and the Model of Condition Equations

To show the equivalence between the least-squares adjustments within the GMM and the Model of Condition Equations, it must be shown that the predicted random error vectors (residuals) from both adjustments are equivalent. The residual vector $\tilde{e}$ from each adjustment can be expressed as a projection matrix times the true random error vector $e$ (or equivalently, times the observation vector $y$) as shown below.

The residual vector within the GMM can be written as

$$\tilde{e} = [I_n - AN^{-1}A^T P]e.$$  (4.12)

And the residual vector within the Model of Condition Equations can be written as

$$\tilde{e} = [P^{-1}B^T(BP^{-1}B^T)^{-1}B]e.$$  (4.13)

Note that the right sides of (4.12) and (4.13) cannot actually be computed since $e$ is unknown, but the equations do hold since, for the GMM,

$$\tilde{e} = [I_n - AN^{-1}A^T P]y = [I_n - AN^{-1}A^T P](A\xi + e) = [A\xi - AN^{-1}(A^T PA)\xi] + [I_n - AN^{-1}A^T P]e \Rightarrow \tilde{e} = [I_n - AN^{-1}A^T P]e,$$  (4.14)

and, for the Model of Condition Equations,

$$\tilde{e} = P^{-1}B^T(BP^{-1}B^T)^{-1}By = P^{-1}B^T(BP^{-1}B^T)^{-1}B(A\xi + e) \Rightarrow \tilde{e} = [P^{-1}B^T(BP^{-1}B^T)^{-1}B]e,$$  (4.15)

using the fact that $BA = 0$.

To show that (4.12) and (4.13) are equivalent, it must be shown that the range spaces and the nullspaces are equivalent for their respective projection matrices $[I_n - AN^{-1}A^T P]$ and $[P^{-1}B^T(BP^{-1}B^T)^{-1}B]$.

(i) Equivalent range spaces: Show that

$$\mathcal{R}[I_n - AN^{-1}A^T P] = \mathcal{R}[P^{-1}B^T(BP^{-1}B^T)^{-1}B].$$

Proof: Since $A^T PP^{-1}B^T = A^T B^T = 0$, then

$$[I_n - AN^{-1}A^T P][P^{-1}B^T(BP^{-1}B^T)^{-1}B]z = [P^{-1}B^T(BP^{-1}B^T)^{-1}B]z - 0 \text{ for any } z \in \mathbb{R}^n,$
which, with the help of (A.6), implies that
\[ \mathcal{R}[P^{-1}B^T(BP^{-1}B)^{-1}B] \subset \mathcal{R}[I_n - AN^{-1}A^T P]. \]

Also:
\[
\dim \mathcal{R}[P^{-1}B^T(BP^{-1}B)^{-1}B] = \\
= \text{rk}[P^{-1}B^T(BP^{-1}B)^{-1}B] = \\
= \text{tr}[P^{-1}B^T(BP^{-1}B)^{-1}B] = \text{using (A.12)} \\
= \text{tr}[BP^{-1}B^T(BP^{-1}B)^{-1}] = \text{using (A.5)} \\
= \text{tr}(I_r) = r.
\]

Furthermore:
\[
\dim \mathcal{R}[I_n - AN^{-1}A^T P] = \\
= \text{rk}(I_n - AN^{-1}A^T P) = \\
= \text{tr}(I_n - AN^{-1}A^T P) = \text{using (A.12)} \\
= \text{tr}(I_n) - \text{tr}(N^{-1}A^T PA) = \text{using (A.5)} \\
= n - \text{rk} N = n - \text{rk} A = \\
= n - q = r,
\]

which implies that
\[ \mathcal{R}[I_n - AN^{-1}A^T P] = \mathcal{R}[P^{-1}B^T(BP^{-1}B)^{-1}B], \quad (4.16) \]

since one range space contains the other and both have the same dimension.

(ii) Equivalent Nullspaces: Show that
\[ \mathcal{N}[I_n - AN^{-1}A^T P] = \mathcal{N}[P^{-1}B^T(BP^{-1}B)^{-1}B]. \]

Proof:
First show that \( \mathcal{N}[I_n - AN^{-1}A^T P] = \mathcal{R}(A) \).
\[ [I_n - AN^{-1}A^T P]A\alpha = 0 \text{ for all } \alpha \]
\[ \Rightarrow \mathcal{R}(A) \subset \mathcal{N}[I_n - AN^{-1}A^T P] \]

Also:
\[ \dim \mathcal{R}(A) = \text{rk} A = q \]

and
\[ \dim \mathcal{N}[I_n - AN^{-1}A^T P] = \\
= n - \dim \mathcal{R}[I_n - AN^{-1}A^T P] = n - r = q \]
\[ \Rightarrow \mathcal{N}[I_n - AN^{-1}A^T P] = \mathcal{R}(A) \]
\[ [P^{-1}B^T (BP^{-1}B^T)^{-1}B] A = 0, \]

since \( BA = 0 \). The preceding development implies that
\[ \mathcal{R}(A) = \mathcal{N}[I_n - AN^{-1}A^T P] \subset \mathcal{N}[P^{-1}B^T (BP^{-1}B^T)^{-1}B]. \]

We showed in part (i) that the dimensions of the range spaces of the respective projection matrices are equivalent. And, since \( \dim \mathcal{N}(\cdot) = n - \dim \mathcal{R}(\cdot) \), it follows that the dimension of the nullspaces of the respective projection matrices are also equivalent. Here in part (ii), we have shown that the nullspace of the projection matrix from the GMM is contained within the nullspace of the projection matrix from the Model of Condition Equations. But, if one space is a subset of another and both spaces have the same dimension, then the subspaces are equivalent. Therefore, we can indeed say that the projections are the same, and thus the adjustments are equivalent.

4.3 Numerical Example – a Small Leveling Network

The following example is borrowed from Mikhail and Gracie (1981, problem 4-8). It involves a leveling network comprised of two closed loops as shown in Figure 4.1. The data are listed in Table 4.1.

\[ A \]
\[ B \]
\[ 18 \text{ km} \]
\[ y_1 \]
\[ 12 \text{ km} \]
\[ y_2 \]
\[ 20 \text{ km} \]
\[ y_3 \]
\[ 22 \text{ km} \]
\[ y_5 \]
\[ 8 \text{ km} \]
\[ y_4 \]

Figure 4.1: Example leveling network

In a leveling network, one condition equation can be written for each closed loop. Connecting observations in a counter-clockwise order, two condition equations may be written as
\[ (y_1 - e_1) + (y_2 - e_2) + (y_3 - e_3) = 0, \]
\[ -(y_3 - e_3) + (y_4 - e_4) + (y_5 - e_5) = 0, \]
Table 4.1: Leveling network data

<table>
<thead>
<tr>
<th>Line</th>
<th>Element of $y$</th>
<th>Observed elevation difference (m)</th>
<th>Length (km)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$A$ to $B$</td>
<td>$y_1$</td>
<td>$-12.386$</td>
<td>$18$</td>
</tr>
<tr>
<td>$B$ to $C$</td>
<td>$y_2$</td>
<td>$-11.740$</td>
<td>$12$</td>
</tr>
<tr>
<td>$C$ to $A$</td>
<td>$y_3$</td>
<td>$24.101$</td>
<td>$20$</td>
</tr>
<tr>
<td>$C$ to $D$</td>
<td>$y_4$</td>
<td>$-8.150$</td>
<td>$8$</td>
</tr>
<tr>
<td>$D$ to $A$</td>
<td>$y_5$</td>
<td>$32.296$</td>
<td>$22$</td>
</tr>
</tbody>
</table>

or, in matrix form, as

$$\begin{bmatrix} 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & -1 & 1 & 1 \end{bmatrix} \begin{bmatrix} -12.386 \\ -11.740 \\ 24.101 \\ -8.150 \\ 32.296 \end{bmatrix} = Be.$$  

The observations weights are inversely proportional to the distances in km, so that

$$P^{-1} = 10^{-6} \cdot \text{diag}(18, 12, 20, 8, 22) \cdot \text{m}^2$$

appears to be a reasonable weight matrix. The residuals are then computed by

$$\tilde{e} = P^{-1}B^T(BP^{-1}B^T)^{-1}By = \begin{bmatrix} -0.003 \\ -0.002 \\ -0.020 \\ 0.007 \\ 0.018 \end{bmatrix} \cdot \text{m}.$$  

The redundancy of the model is given by $r = \text{rk} \, B = 2$. The adjusted observations
are computed by

$$\hat{\mu}_y = y - \tilde{\epsilon} = \begin{bmatrix} -12.383 \\ -11.738 \\ 24.121 \\ -8.157 \\ 32.278 \end{bmatrix} \cdot \text{m.}$$

The dispersion matrix for the residuals is

$$D\{\tilde{\epsilon}\} = \sigma_0^2 \cdot P^{-1} B^T (BP^{-1}B^T)^{-1}BP^{-1} =$$

$$= \sigma_0^2 \cdot \begin{bmatrix} 0.000077 & 0.000051 & 0.000051 & 0.000014 & 0.000038 \\ 0.000051 & 0.000034 & 0.000034 & 0.000009 & 0.000025 \\ 0.0000051 & 0.000034 & 0.0000114 & -0.000023 & -0.000063 \\ 0.000014 & 0.000009 & -0.000023 & 0.000015 & 0.000042 \\ 0.000038 & 0.000025 & -0.000063 & 0.000042 & 0.0000115 \end{bmatrix} \text{m}^2.$$

The weighted sum of squared residuals is $\Omega := \tilde{\epsilon}^T P \tilde{\epsilon} = (6.454972)^2$, leading to the estimated variance component $\hat{\sigma}_0^2 = \Omega/r = (4.564355)^2$. The estimated variance component may be used to compute the estimated dispersion matrix for the residuals as

$$\hat{D}\{\tilde{\epsilon}\} = \hat{\sigma}_0^2 \cdot P^{-1} B^T (BP^{-1}B^T)^{-1}BP^{-1} =$$

$$= \begin{bmatrix} 0.00016 & 0.00011 & 0.00011 & 0.00003 & 0.00008 \\ 0.00011 & 0.00007 & 0.00007 & 0.00002 & 0.00005 \\ 0.00011 & 0.00007 & 0.00024 & -0.00005 & -0.00013 \\ 0.00003 & 0.00002 & -0.00005 & 0.00003 & 0.00009 \\ 0.00008 & 0.00005 & -0.00013 & 0.00009 & 0.00024 \end{bmatrix} \text{m}^2.$$

38
4.4 Summary Formulas for the Least-Squares Solution Within the Model of Condition Equations

The Model of Condition Equations is given by

\[ \begin{align*}
    w_{r \times 1} & := B y_{r \times n} = Be, \quad e \sim (0, \sigma_0^2 P^{-1}), \\
    r & := \text{rk} B.
\end{align*} \]

Table 4.2: Summary formulas for the LELESS within the Model of Condition Equations

<table>
<thead>
<tr>
<th>Quantity</th>
<th>Formula</th>
<th>Eq. No.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model redundancy</td>
<td>( r = \text{rk} B )</td>
<td>(4.1b)</td>
</tr>
<tr>
<td>Vector of predicted residuals</td>
<td>( \tilde{e} = P^{-1} B^T (BP^{-1}B^T)^{-1} By )</td>
<td>(4.5e)</td>
</tr>
<tr>
<td>Dispersion matrix for residuals</td>
<td>( D{\tilde{e}} = \sigma_0^2 P^{-1} B^T (BP^{-1}B^T)^{-1} BP^{-1} )</td>
<td>(4.9)</td>
</tr>
<tr>
<td>Sum of squared residuals (SSR)</td>
<td>( \Omega = \tilde{e}^T P \tilde{e} )</td>
<td>(4.7)</td>
</tr>
<tr>
<td>Estimated variance component</td>
<td>( \hat{\sigma}_0^2 = \Omega / r )</td>
<td>(4.8)</td>
</tr>
<tr>
<td>Vector of adjusted observations</td>
<td>( \hat{\mu}_y = y - \tilde{e} )</td>
<td>(4.6)</td>
</tr>
<tr>
<td>Dispersion matrix for adjusted observations</td>
<td>( D{\hat{\mu}_y} = \sigma_0^2 P^{-1} - D{\tilde{e}} )</td>
<td>(4.11)</td>
</tr>
</tbody>
</table>
The Gauss-Markov Model with Constraints

When *prior information* about the values of certain parameters, or about functional relationships between them, is known before the adjustment, those quantities can be maintained through the adjustment by application of constraints. For example, one may already know the height difference between two points in a leveling network that is to be adjusted, or it could be that the azimuth between two points in a 2D network to be adjusted must maintain a specified value. In both cases, the prior information can be preserved through constraints added to the Gauss-Markov Model (GMM). We say that such information is known *a priori*. The term *a priori* is a Latin phrase that literally means “from the earlier.” In geodetic science, it refers to knowledge or information possessed before an experiment is conducted or an adjustment is computed.

One case where constraints might be useful is when the design matrix $A$ does not have full column-rank, implying that the inverse $N^{-1}$ of the normal equation matrix does not exist, which means that the parameters of the model cannot be estimated using (3.7). This problem can occur, for example, when network observations must be adjusted in the estimation of point coordinates, but the observations themselves do not provide complete information about the network datum (i.e., its size, shape, orientation, and origin). For example, distance measurements provide information about the scale of a network, and angle measurements provide information about its shape. But neither measurement type provides information about the origin or orientation of the network figure, which is necessary for estimating coordinates of network points. In such a 2-D case, applying a constraint on two coordinates (i.e., on two parameters) and one azimuth (a function of four parameters) would provide the missing information. In this case, the specified constraint values could be somewhat arbitrary, but we still may speak of them as being “known” (i.e., specified a priori) in the context of adjustments with constraints.
The Gauss-Markov Model (GMM) with constraints is written as
\[ y_{n \times 1} = A_{n \times m} \xi_{m \times 1} + e_{n \times 1}, \quad e \sim (0, \sigma^2_0 P^{-1}), \quad \text{rk} \ A =: q \leq \min\{m, n\}, \] (5.1a)
\[ \kappa_0 = K_{l \times m} \xi_{m \times 1}, \quad \text{rk} \ K =: l \geq m - q, \quad \text{rk}[A^T, K^T] = m. \] (5.1b)
The variables are as defined on page 24, but now with the addition of a known \( l \times m \) coefficient matrix \( K \) and an \( l \times 1 \) vector of specified constants \( \kappa_0 \). Symbols for the normal equations were introduced in (3.4) and are repeated here for convenience:
\[ [N, c] = A^T P [A, y]. \] (5.2)
Note that, in contrast to the model in (3.1), the coefficient matrix \( A \) in (5.1a) is not required to have full column-rank, in which case the matrix inverse \( N^{-1} \) would not exist. However, the specified rank conditions imply that \((N+K^T K)^{-1}\) exists, and, if \( N^{-1} \) exists, so does \((KN^{-1}K^T)^{-1}\). This is because the range space of \([A^T, K^T]\) spans \( \mathbb{R}^m \) as implied by the rank condition stated in (5.1b). The redundancy of the system is computed by
\[ r := n - m + \text{rk} K = n - m + l. \] (5.3)
Introducing an \( l \times 1 \) vector of Lagrange multipliers \( \lambda \), the Lagrange target function to minimize is
\[ \Phi(\xi, \lambda) := (y - A\xi)^T P(y - A\xi) - 2\lambda^T (\kappa_0 - K\xi) = \text{stationary}. \] (5.4)
Its first partial derivatives are taken to form the following Euler-Lagrange necessary conditions:
\[ \frac{1}{2} \frac{\partial \Phi}{\partial \xi} = N\dot{\xi} - c + K^T \dot{\lambda} = 0, \] (5.5a)
\[ \frac{1}{2} \frac{\partial \Phi}{\partial \lambda} = -\kappa_0 + K\dot{\xi} = 0. \] (5.5b)
In matrix form (5.5a) and (5.5b) are expressed as
\[ \begin{bmatrix} N & K^T \\ K & 0 \end{bmatrix} \begin{bmatrix} \dot{\xi} \\ \dot{\lambda} \end{bmatrix} = \begin{bmatrix} c \\ \kappa_0 \end{bmatrix}, \] (5.6)
where the vector on the left side contains \( m + l \) unknowns to be estimated.
The sufficient condition, required for minimization, is satisfied by
\[ (1/2)(\partial^2 \Phi / \partial \xi \partial \xi^T) = N, \]
which is positive-(semi)definite. We refer to the matrix on the left side of (5.6) as the least-squares normal equation matrix. It is invertible if, and only if, \( \text{rk}[A^T, K^T] = m \). This rank condition means that, for the normal equation matrix,
• among the first \( m \) columns, \( m - l \) must be linearly independent, and
• the additional \( l \) columns are complementary, meaning that when combined with the first \( m - l \) columns they span \( \mathbb{R}^m \).

In the following, we consider two cases: (1) \( N \) is invertible (nonsingular or regular), and (2) \( N \) is singular.

**Case 1:** \( N \) is invertible, implying that matrix \( A \) has full column-rank, i.e., \( \text{rk} A = m \). Equations (5.5a) and (5.5b) imply

\[
\hat{\xi} = N^{-1}(c - K^T \hat{\lambda}), \quad \kappa_0 = K \hat{\xi} = KN^{-1}c - KN^{-1}K^T \hat{\lambda}
\]

\[
\Rightarrow \hat{\lambda} = -(KN^{-1}K^T)^{-1}(\kappa_0 - KN^{-1}c),
\]

finally leading to the LESS

\[
\hat{\xi} = N^{-1}c + N^{-1}K^T(KN^{-1}K^T)^{-1}(\kappa_0 - KN^{-1}c).
\]

The vector difference \( \kappa_0 - KN^{-1}c \) in (5.7d) is called a *vector of discrepancies*. It shows the mismatch between the vector of specified constants \( \kappa_0 \) and a linear combination (as generated by the matrix \( K \)) of the solution without constraints (i.e., \( N^{-1}c \)). The estimated vectors \( \hat{\xi} \) and \( \hat{\lambda} \) may also be presented in terms of the inverse of the matrix in (5.6), viz.

\[
\begin{bmatrix} \hat{\xi} \\ \hat{\lambda} \end{bmatrix} = \begin{bmatrix} N^{-1} - N^{-1}K^T(KN^{-1}K^T)^{-1}KN^{-1} & N^{-1}K^T(KN^{-1}K^T)^{-1} \\ (KN^{-1}K^T)^{-1} & -(KN^{-1}K^T)^{-1} \end{bmatrix} \begin{bmatrix} c \\ \kappa_0 \end{bmatrix}.
\]

Also, note that the expectation of the estimated vector of Lagrange multipliers is derived by

\[
E\{\hat{\lambda}\} = -E\{(KN^{-1}K^T)^{-1}(\kappa_0 - KN^{-1}c)\} = (KN^{-1}K^T)^{-1}[KN^{-1}A^TPE\{y\} - \kappa_0] = (KN^{-1}K^T)^{-1}(K\xi - \kappa_0) = 0 = E\{\hat{\lambda}\}.
\]

**Case 2:** \( N \) is singular (i.e., not invertible), implying that matrix \( A \) does not have full column-rank, i.e., \( \text{rk} A < m \). Multiplying equation (5.5b) by \( K^T \) and adding the result to (5.5a), leads to

\[
(N + K^TK)\hat{\xi} = c + K^T(\kappa_0 - \hat{\lambda}) \Rightarrow \hat{\xi} = (N + K^TK)^{-1}c + (N + K^TK)^{-1}K^T(\kappa_0 - \hat{\lambda}).
\]

Then from (5.5b) and (5.10) we have

\[
\kappa_0 = K\hat{\xi} = K(N + K^TK)^{-1}c + K(N + K^TK)^{-1}K^T(\kappa_0 - \hat{\lambda}) \Rightarrow \]
(\kappa_0 - \hat{\lambda}) = \left[ K \left( N + K^T K \right)^{-1} K^T \right]^{-1} \left[ \kappa_0 - K \left( N + K^T K \right)^{-1} c \right]. \quad (5.11)

Substituting (5.11) into (5.10) leads to the Least-Squares Solution (LESS)

$$\hat{\xi} = \left( N + K^T K \right)^{-1} c + \left( N + K^T K \right)^{-1} K^T \left[ K \left( N + K^T K \right)^{-1} K^T \right]^{-1} \left[ \kappa_0 - K \left( N + K^T K \right)^{-1} c \right].$$ \quad (5.12)

The form of (5.12) is identical to (5.7d) except that all occurrences of matrix \(N\) in (5.7d) have been replaced by \(N + K^T K\) in (5.12). Of course, (5.12) can be used for both the singular and nonsingular cases.

Also, note that the expectation of vector difference \(\kappa_0 - \hat{\lambda}\) is derived by

$$E\{\kappa_0 - \hat{\lambda}\} = E\{ \left[ K \left( N + K^T K \right)^{-1} K^T \right]^{-1} \left[ \kappa_0 - K \left( N + K^T K \right)^{-1} c \right] \} = \left[ K \left( N + K^T K \right)^{-1} K^T \right]^{-1} K \left[ I_m - \left( N + K^T K \right)^{-1} N \right] \xi = \left[ K \left( N + K^T K \right)^{-1} K^T \right]^{-1} K \left[ I_m - \left( N + K^T K \right)^{-1} N \right] \xi = \left[ K \left( N + K^T K \right)^{-1} K^T \right]^{-1} K \left( N + K^T K \right)^{-1} K^T \xi = K \xi = E\{ \kappa_0 - \hat{\lambda} \}. \quad (5.13)$$

We now compute the formal dispersion of the vector of estimated parameters \(\hat{\xi}\) for both cases. From (5.6) we have

$$\begin{bmatrix} \hat{\xi} \\ \hat{\lambda} \end{bmatrix} = \begin{bmatrix} N & K^T \\ K & 0 \end{bmatrix}^{-1} \begin{bmatrix} c \\ \kappa_0 \end{bmatrix}, \quad (5.14)$$

which, from the law of covariance propagation, implies that

$$D\{ \xi \} \begin{bmatrix} N & K^T \\ K & 0 \end{bmatrix}^{-1} D\{ \kappa_0 \} \begin{bmatrix} N & K^T \\ K & 0 \end{bmatrix}^{-1} = \sigma_0^2 \begin{bmatrix} N & K^T \\ K & 0 \end{bmatrix}^{-1} \left[ \begin{bmatrix} N & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} N & K^T \\ K & 0 \end{bmatrix}^{-1} \right]. \quad (5.15)$$

Here, the symmetry of the normal-equation matrix and the fact that \(\kappa_0\) is a non-random vector have been considered. Upon algebraic reduction of (5.15), we find that

$$\begin{bmatrix} D\{ \xi \} & X \\ X & -D\{ \hat{\lambda} \} \end{bmatrix} = \sigma_0^2 \begin{bmatrix} N & K^T \\ K & 0 \end{bmatrix}^{-1}, \quad (5.16)$$
where the symbol $X$ represents a term of no particular interest. Note that $X \neq C\{\xi, \lambda\} = 0$. The inverse on the right side of (5.16) reveals the following dispersion matrices for cases 1 and 2, respectively:

**Case 1:**

\[
D\{\hat{\xi}\} = \sigma_0^2 \left[ N^{-1} - N^{-1}K^T(KN^{-1}K^T)^{-1}KN^{-1} \right] \quad (5.17a)
\]

\[
D\{\hat{\lambda}\} = \sigma_0^2 (KN^{-1}K^T)^{-1} \quad (5.17b)
\]

**Case 2:**

\[
D\{\hat{\xi}\} = \sigma_0^2 \left( N + K^T K \right)^{-1} - \sigma_0^2 \left( N + K^T K \right)^{-1}K^T \cdot \left[ K (N + K^T K)^{-1}K^T \right]^{-1}K (N + K^T K)^{-1} \quad (5.18a)
\]

\[
D\{\hat{\lambda}\} = \sigma_0^2 \left\{ \left[ K (N + K^T K)^{-1}K^T \right]^{-1} - I_l \right\} \quad (5.18b)
\]

As with the parameter estimates, the dispersion matrices for both cases have a similar form, with every occurrence of $N$ in Case 1 being replaced by $N + K^T K$ in Case 2. Also note that the dispersions in (5.17a) and (5.18a) are nothing more than the coefficient matrices multiplying the vector $c$ in (5.7d) and (5.12), respectively, multiplied by the (unknown) variance component $\sigma_0^2$. Finally, it is clear from the above that the constraints reduce the dispersion matrix of $\hat{\xi}$ compared to the corresponding dispersion matrix within the GMM (without constraints) derived in Chapter 3 (cf. (3.13)).

For completeness, we show the inverse of the block matrix on the right side of (5.16), where, for the sake of brevity, the notation $N_k := (N + K^T K)$ is introduced.

\[
\begin{bmatrix}
N & K^T \\
K & 0
\end{bmatrix}^{-1} = \begin{bmatrix}
N_k^{-1} - N_k^{-1}K^T(KN_k^{-1}K^T)^{-1}KN_k^{-1} & N_k^{-1}K^T(KN_k^{-1}K^T)^{-1} \\
(KN_k^{-1}K^T)^{-1}KN_k^{-1} & I_l - (KN_k^{-1}K^T)^{-1}
\end{bmatrix} \quad (5.19)
\]

Aside from the term $N^{-1}$ in (5.8) being replaced by $N_k^{-1}$ in (5.19), note that the matrix inverse contains the term $I_l$ in the lower-right block, which does not appear in (5.8).

In either case, the residual vector $\hat{e}$ and vector of adjusted observations $\hat{\mu}_y$ may be obtained in a straightforward way after the estimation of the parameters by use of the formulas

\[
\hat{e} = y - A\hat{\xi}; \quad (5.20)
\]

and

\[
\hat{\mu}_y = y - \hat{e}. \quad (5.21)
\]

Here, $\hat{\mu}_y$ is also interpreted as an estimate of the true, and thus unknown, vector of observables $\mu_y$, where $E(y) = \mu_y$. 

45
The dispersion matrix for the residual vector \( \tilde{e} \) can be derived from application of the law of covariance propagation as follows: Since
\[
D\{\tilde{e}\} = D\{y - A\hat{\xi}\} = D\{y\} + AD\{\hat{\xi}\}A^T - 2C\{y, A\hat{\xi}\},
\] (5.22)
we start by deriving the covariance matrix \( C\{y, A\hat{\xi}\} \). For case 1 we have
\[
C\{y, A\hat{\xi}\} = I_n \cdot D\{y\} \cdot \{A[N^{-1}A^TP - N^{-1}K^T(KN^{-1}K^T)^{-1}KN^{-1}A^TP]\}^T = \sigma^2_0 \cdot \{P^{-1} - A[N^{-1} - N^{-1}K^T(KN^{-1}K^T)^{-1}KN^{-1}]A^T\} = AD\{\hat{\xi}\}A^T = D\{A\hat{\xi}\} = C\{y, A\hat{\xi}\}. \] (5.23a)

Then, by substituting (5.23d) into (5.22), we arrive at
\[
D\{\tilde{e}\} = D\{y\} - AD\{\hat{\xi}\}A^T \Rightarrow
D\{\tilde{e}\} = \sigma^2_0 \cdot \{P^{-1} - A[N^{-1} - N^{-1}K^T(KN^{-1}K^T)^{-1}KN^{-1}]A^T\} \] (5.24a)
and
\[
D\{\tilde{e}\} = \sigma^2_0 \cdot [P^{-1} - AN^{-1}A^T + AN^{-1}K^T(KN^{-1}K^T)^{-1}KN^{-1}A^T]. \] (5.24c)

Note that (5.24c) reveals that the dispersion matrix for the residuals within the GMM with constraints is larger than that for the GMM without constraints (cf. (3.14a)). For case 2, one only needs to replace the matrix \( N^{-1} \) with \((N + K^T K)^{-1}\) in formulas (5.23) and (5.24).

Obviously, the dispersion matrix for the adjusted observations is written as
\[
D\{\hat{\mu}_y\} = D\{y - \tilde{e}\} = D\{A\hat{\xi}\} = AD\{A\hat{\xi}\}A^T. \] (5.25)

5.1 Estimated Variance Component

The estimated variance component for the GMM with constraints is derived similar to that for the GMM without constraints as shown in Section 3.2. The estimation is based on the principle
\[
\frac{\sigma^2_0}{\tilde{e}^T \tilde{P} \tilde{e}} = \frac{\sigma^2_0}{E\{\tilde{e}^T \tilde{P} \tilde{e}\}}. \] (5.26)

Furthermore, for the purpose of validating the constraints, we wish to decompose the quadratic form \( \tilde{e}^T \tilde{P} \tilde{e} \) into the sum \( \tilde{e}^T \tilde{P} \tilde{e} = \Omega + R \). In the following, we derive these components for both cases 1 and 2.

5.1.1 Case 1 — Matrix \( N \) is invertible \( \Rightarrow \text{rk} A = m \)
\[
\tilde{e}^T \tilde{P} \tilde{e} = (y - A\hat{\xi})^T P(y - A\hat{\xi}) = \]

respectively.

Obviously, both Ω and the constraints LESS within the GMM without constraints, whereas turns out that they are also uncorrelated. The variable Ω is associated with the hypothesis testing as discussed in Chapter 10.

Thus we have decomposed the quadratic form \( \tilde{e}^T P \tilde{e} \) into components Ω and R. From (5.28b) we see that R is always positive, and thus the constraints will increase the value of \( \tilde{e}^T P \tilde{e} \). The variables Ω and R are used for hypothesis testing as discussed in Chapter 10.

We now derive the expectation of \( \tilde{e}^T P \tilde{e} \).

\[
E\{\tilde{e}^T P \tilde{e}\} = E\{\Omega\} + E\{R\} =
\]

\[
= (n - m)\sigma_0^2 + E\{\tilde{\lambda}^T (KN^{-1}K^T) \tilde{\lambda}\} = \text{using (3.25) for } E\{\Omega\}
\]

\[
= (n - m)\sigma_0^2 + \text{tr}[(KN^{-1}K^T)E\{\tilde{\lambda} \tilde{\lambda}^T\}] =
\]

\[
= (n - m)\sigma_0^2 + \text{tr}[(KN^{-1}K^T)(D\{\tilde{\lambda}\} + E\{\tilde{\lambda} \ E\{\tilde{\lambda}\}^T\}] =
\]

\[
(\text{with } E\{\tilde{\lambda}\} = 0 \text{ and } D\{\tilde{\lambda}\} = \sigma_0^2 (KN^{-1}K^T)^{-1})
\]

\[
= (n - m)\sigma_0^2 + \text{tr}[(KN^{-1}K^T)\sigma_0^2 (KN^{-1}K^T)^{-1}]
\]

\[
= (n - m + 1)\sigma_0^2
\]

(5.29)
Substitution of (5.27) and (5.29) into (5.26) yields the following formula for the estimated variance component:

\[
\hat{\sigma}_0^2 = \frac{(y - AN^{-1}c)^T P(y - AN^{-1}c)}{n - m + l} + \frac{(\kappa_0 - KN^{-1}c)^T (KN^{-1}K^T)^{-1}(\kappa_0 - KN^{-1}c)}{n - m + l}.
\]

(5.30)

Other useful forms of \(\hat{\epsilon}^T P\hat{\epsilon}\) are derived below starting with (5.27).

\[
\hat{\epsilon}^T P\hat{\epsilon} = (y - AN^{-1}c)^T P(y - AN^{-1}c) + \hat{\lambda}^T (KN^{-1}K^T)\hat{\lambda} =
\]

\[
y^T P y - c^T N^{-1} c - (\kappa_0 - \hat{\lambda})^T (\kappa_0 - \hat{\lambda}) = \text{using (5.7c)}
\]

\[
y^T P y - c^T N^{-1} (c - K^T \hat{\lambda}) - \kappa_0^T \hat{\lambda} = \text{using (5.7a)}
\]

\[
y^T P y - c^T \hat{\xi} - \kappa_0^T \hat{\lambda} =
\]

\[
y^T P (y - A\hat{\xi}) - \kappa_0^T \hat{\lambda} =
\]

\[
y^T P \hat{\epsilon} - \kappa_0^T \hat{\lambda}
\]

(5.31)

5.1.2 Case 2 — Matrix \(N\) is singular \(\Rightarrow \text{rk} A < m\)

\[
\hat{\epsilon}^T P\hat{\epsilon} =
\]

\[
= \{y - A(N + K^T K)^{-1} [c + K^T (\kappa_0 - \hat{\lambda})]\}^T P \{y - A(N + K^T K)^{-1} \cdot \left[c + K^T (\kappa_0 - \hat{\lambda})\right]\}
\]

\[
= y^T P y - y^T PA(N + K^T K)^{-1} [c + K^T (\kappa_0 - \hat{\lambda})] - [c + K^T (\kappa_0 - \hat{\lambda})]^T \cdot (N + K^T K)^{-1} A^T P y + [c + K^T (\kappa_0 - \hat{\lambda})]^T (N + K^T K)^{-1}.
\]

\[
\cdot (A^T PA + K^T K - K^T K) (N + K^T K)^{-1} [c + K^T (\kappa_0 - \hat{\lambda})] =
\]

\[
y^T P y - c^T (N + K^T K)^{-1} [c + K^T (\kappa_0 - \hat{\lambda})] - [c + K^T (\kappa_0 - \hat{\lambda})]^T \cdot (N + K^T K)^{-1} c + [c + K^T (\kappa_0 - \hat{\lambda})]^T (N + K^T K)^{-1} [c + K^T (\kappa_0 - \hat{\lambda})] -
\]

\[
- [c + K^T (\kappa_0 - \hat{\lambda})]^T (N + K^T K)^{-1} K^T K (N + K^T K)^{-1} [c + K^T (\kappa_0 - \hat{\lambda})] =
\]

\[
y^T P y - c^T (N + K^T K)^{-1} [c + K^T (\kappa_0 - \hat{\lambda})] -
\]

\[
- c^T (N + K^T K)^{-1} c - (\kappa_0 - \hat{\lambda})^T K (N + K^T K)^{-1} c +
\]

\[
+ c^T (N + K^T K)^{-1} [c + K^T (\kappa_0 - \hat{\lambda})] + (\kappa_0 - \hat{\lambda})^T K (N + K^T K)^{-1} c +
\]

\[
+ (\kappa_0 - \hat{\lambda})^T K (N + K^T K)^{-1} K^T (\kappa_0 - \hat{\lambda}) - \hat{\xi}^T K^T K \hat{\xi} =
\]

\[
y^T P y - c^T (N + K^T K)^{-1} c + (\kappa_0 - \hat{\lambda})^T [K (N + K^T K)^{-1} K^T] (\kappa_0 - \hat{\lambda}) -
\]

\[
- \kappa_0^T \kappa_0
\]

(5.32)
Now we compute the expectation for $\hat{e}^T P \hat{e}$.

$$E\{\hat{e}^T P \hat{e}\} =$$

$$= E\{y^T P y - \hat{e}^T (N + K^T K)^{-1} c\} +$$

$$\quad + (\kappa_0 - \lambda)^T \left[ K (N + K^T K)^{-1} K^T \right] (\kappa_0 - \lambda) - \kappa_0^T \kappa_0 \} =$$

$$= E\{y^T P \left[ y - A(N + K^T K)^{-1} c \right] \} + E\{(\kappa_0 - \lambda)^T \left[ K (N + K^T K)^{-1} K^T \right] \cdot (\kappa_0 - \lambda) \} - E\{\kappa_0^T \kappa_0 \} =$$

$$= \text{tr} [I_n - A(N + K^T K)^{-1} A^T P] E\{yy^T\} + \text{tr} [K(N + K^T K)^{-1} K^T].$$

(5.34b)

Finally, substituting (5.32) and (5.33) into (5.26) yields

$$\sigma_0^2 = \frac{y^T P y - \hat{e}^T (N + K^T K)^{-1} c}{(n - m + l)} +$$

$$\quad + \frac{(\kappa_0 - \lambda)^T \left[ K (N + K^T K)^{-1} K^T \right] (\kappa_0 - \lambda) - \kappa_0^T \kappa_0 \} (n - m + l), \quad \text{(5.34a)}$$

or, by use of (5.11) and with $N_k := N + K^T K$ for compactness,

$$\sigma_0^2 = \frac{y^T P y - \hat{e}^T N_k^{-1} c + (\kappa_0 - K N_k^{-1} c)^T (K N_k^{-1} K^T) (\kappa_0 - K N_k^{-1} c) - \kappa_0^T \kappa_0 \} (n - m + l).$$

(5.34b)
or
\[
\hat{\sigma}^2_0 = \frac{\hat{e}^T P \hat{e}}{(n - m + l)},
\]
(5.34c)

We cannot directly identify \( \Omega \) and \( R \) in (5.32) as we could in case 1. Therefore, we define \( \Omega \) as
\[
\Omega = (y - AN^{-1}c)^T P (y - AN^{-1}c),
\]
(5.35)

and \( R \) as
\[
R = \hat{e}^T P \hat{e} - \Omega,
\]
(5.36)

where \( \hat{e}^T P \hat{e} \) is given in (5.32). The symbol \( N^{-1} \) in (5.35) stands for a generalized inverse of the matrix \( N \). While generalized inverses are beyond the scope of these notes, the following generalized inverse is shown so that readers unfamiliar with the topic can still make use of equation (5.35). First, assume that the matrix \( N \) and vector \( c \) have been partitioned as follows:
\[
N_{m \times m} = \begin{bmatrix}
N_{11} & N_{12} \\
N_{21} & N_{22}
\end{bmatrix}
\quad \text{and} \quad
c_{m \times 1} = \begin{bmatrix}
c_1 \\
c_2
\end{bmatrix},
\]
(5.37)

where the \( q \times q \) submatrix \( N_{11} \) has full rank \( q \), i.e., \( \text{rk} N_{11} = q := \text{rk} N \). Note that such a partitioning can always be formed, even if the parameters in \( \xi \) must be reordered to do so. Then, the \( m \times m \) matrix \( G := \begin{bmatrix}
N_{11}^{-1} & 0 \\
0 & 0
\end{bmatrix} \) is a generalized inverse of \( N \) and thus can be used in (5.35) for \( N^{-1} \), which simplifies that equation to
\[
\Omega = y^T P y - c_1^T N_{11}^{-1} c_1 \quad \text{if} \quad \text{rk} N_{11} = \text{rk} N.
\]
(5.38)

The following ratio is formed for both cases 1 and 2 for the purposes of hypothesis testing (see Chapter 10 for more details on hypothesis testing):
\[
T := \frac{R/(l - m + q)}{\Omega/(n - q)} \sim F(l - m + q, n - q), \quad \text{with} \quad q := \text{rk}(A).
\]
(5.39)

The hypothesis test is then
\[
H_0 : K\xi = \kappa_0 \quad \text{versus} \quad H_A : K\xi \neq \kappa_0.
\]
(5.40)

For some chosen significance level \( \alpha \),

\begin{align*}
\text{Accept } H_0 : & \quad \text{if } T \leq F_{,l - m + q,n - q} \\
\text{Reject } H_0 : & \quad \text{if } T > F_{,l - m + q,n - q},
\end{align*}

where \( F_{,l - m + q,n - q} \) is taken from a table of critical values for the \( F \)-distribution.

Note that the redundancy \( r_2 := n - q \) represents the degrees of freedom for the system of equations if no constraints were applied, whereas the redundancy \( r_1 := l - m + q \) represents the increase in degrees of freedom due to the constraints. In the case that matrix \( A \) has full column rank (i.e., \( \text{rk} A = q = m \)), then the redundancies reduce to \( r_1 := l \) and \( r_2 := n - m \), respectively.
5.2 Summary Formulas for the Least-Squares Solution Within the Gauss-Markov Model with Constraints

The Gauss-Markov Model with constraints is given by

\[
y_{n \times 1} = A_{n \times m} \xi_{m \times 1} + e_{n \times 1}, \quad e \sim (0, \sigma_0^2 P^{-1}), \quad \text{rk} A = q \leq \{m, n\}, \\
\kappa_0 = K_{l \times m} \xi_{m \times 1}, \quad \text{rk} K = l \geq m - q, \quad \text{rk}[A^T, K^T] = m.
\]

<table>
<thead>
<tr>
<th>Quantity</th>
<th>Formula</th>
<th>Eq.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model redundancy</td>
<td>[ r = n - m + \text{rk} K = n - m + l ]</td>
<td>(5.3)</td>
</tr>
<tr>
<td>Vector of estimated parameters, when ( \text{rk} A = m )</td>
<td>[ \hat{\xi} = N^{-1} c + N^{-1} K^T (KN^{-1} K^T)^{-1} (\kappa_0 - K N^{-1} c) ]</td>
<td>(5.7d)</td>
</tr>
<tr>
<td>Dispersion matrix for estimated parameters, when ( \text{rk} A = m )</td>
<td>[ D{\hat{\xi}} = \sigma_0^2 [N^{-1} - N^{-1} K^T (KN^{-1} K^T)^{-1} K N^{-1}] ]</td>
<td>(5.17a)</td>
</tr>
<tr>
<td>Vector of estimated parameters, when ( \text{rk} A &lt; m )</td>
<td>[ \hat{\xi} = (N + K^T K)^{-1} c + (N + K^T K)^{-1} K^T [K(N + K^T K)^{-1} K^T]^{-1} ] [ \kappa_0 - K(N + K^T K)^{-1} c ]</td>
<td>(5.12)</td>
</tr>
<tr>
<td>Dispersion matrix for estimated parameters, when ( \text{rk} A &lt; m )</td>
<td>[ D{\hat{\xi}} = \sigma_0^2 [N_K^{-1} - N_K^{-1} K^T (K N_K^{-1} K^T)^{-1} K N_K^{-1}] ] [ \text{with } N_K = N + K^T K ]</td>
<td>(5.18a)</td>
</tr>
<tr>
<td>Vector of predicted residuals</td>
<td>[ \hat{e} = y - A \hat{\xi} ]</td>
<td>(5.20)</td>
</tr>
<tr>
<td>Dispersion matrix for residuals, when ( \text{rk} A = m )</td>
<td>[ D{\hat{e}} = \sigma_0^2 {P^{-1} - A [N^{-1} - N^{-1} K^T (K N^{-1} K^T)^{-1} K N^{-1}] A^T } ]</td>
<td>(5.24b)</td>
</tr>
<tr>
<td>Quantity</td>
<td>Formula</td>
<td>Eq.</td>
</tr>
<tr>
<td>------------------------------------------------------------------------</td>
<td>---------------------------------------------------------------------------------------------------</td>
<td>--------</td>
</tr>
<tr>
<td>Dispersion matrix for residuals, when ( \text{rk} A &lt; m )</td>
<td>( D{\tilde{e}} = \sigma_0^2 \cdot (P^{-1} - A{(N + KTK)^{-1} - (N + KTK)^{-1}K(N + KTK)^{-1}K}A^T) )</td>
<td>(5.24b)</td>
</tr>
<tr>
<td>Sum of squared residuals (SSR)</td>
<td>( \text{SSR} = \tilde{e}^T P \tilde{e} )</td>
<td>(5.31)</td>
</tr>
<tr>
<td>Estimated variance component</td>
<td>( \hat{\sigma}_0^2 = (\tilde{e}^T P \tilde{e})/r )</td>
<td>(5.34c)</td>
</tr>
<tr>
<td>Vector of adjusted observations</td>
<td>( \hat{\mu}_y = y - \tilde{e} )</td>
<td>(5.21)</td>
</tr>
<tr>
<td>Dispersion matrix for adjusted observations</td>
<td>( D{\hat{\mu}_y} = A \cdot D{\hat{\xi}} \cdot A^T )</td>
<td>(5.25)</td>
</tr>
</tbody>
</table>

*Continued from previous page*
Chapter 6

Introduction of Datum Information to Treat the Rank-Deficient Gauss-Markov Model

A rank-deficient Gauss-Markov Model (GMM) is one in which the rank of the design matrix $A$ is less than its number of columns. This means that at least one column of $A$ is either a scalar multiple of a different column, or it can be expressed as a linear combination of other columns. Such rank deficiency is expressed mathematically as $\text{rk} A < m$, where $m$ is the number of columns of $A$. It implies that $\text{rk} N = \text{rk} A^T PA < m$, which means that the unknown parameters cannot be estimated by (3.7). Put another way, a rank-deficient model is one in which there are more parameters than can be estimated from the data. In fact, the rank of the design matrix reveals the number of estimable parameters of the model.

Rank deficiency often arises in the context of network adjustments where station coordinates must be estimated and the observations do not contain sufficient information to define the underlying coordinate system, also called datum in this context. Thus we speak of a network datum deficiency. As noted in Chapter 5, a 2-D network where only angles and distance have been measured provides an example of a datum deficiency of three, owing to the unknown origin and orientation parameters of the network. A method for introducing datum information is provided in the following sections.

Consider the following (linearized) GMM with rank-deficient design matrix $A$:

$$ y = A\xi + e, \quad e \sim (0, \sigma_e^2 P^{-1}), \quad \text{rk} A =: q < m. \quad (6.1a) $$

We can partition the matrix $A$ as

$$ A_{n \times m} = \begin{bmatrix} A_1 & A_2 \end{bmatrix}_{n \times q} \begin{bmatrix} A_2 \end{bmatrix}_{n \times (m-q)}, \quad \text{with} \quad \text{rk} A_1 = q := \text{rk} A, \quad (6.1b) $$
so that matrix $A_1$ has full column-rank. A compatible partitioning of the parameter vector $\xi$, i.e.,

$$
\xi = \begin{bmatrix}
\xi_1 \\
\xi_2
\end{bmatrix}_{q \times 1}
$$

(6.1c)

leads to the following system of partitioned normal equations:

$$
\begin{bmatrix}
A^T_1 \\
A^T_2
\end{bmatrix} P \begin{bmatrix}
A_1 \\
A_2
\end{bmatrix} \begin{bmatrix}
\hat{\xi}_1 \\
\hat{\xi}_2
\end{bmatrix} =
\begin{bmatrix}
A^T_1 P A_1 \\
A^T_2 P A_2
\end{bmatrix} \begin{bmatrix}
\hat{\xi}_1 \\
\hat{\xi}_2
\end{bmatrix} =
\begin{bmatrix}
A^T_1 P y \\
A^T_2 P y
\end{bmatrix}
$$

(6.2)

The sub-scripted terms in (6.2) may be defined more succinctly as

$$
[N_{ij}, c_i] := A^T_i P [A_j, y], \text{ for } i,j \in \{1,2\}.
$$

(6.3)

Defining a datum for $m-q$ parameters means that values for them must be specified. Mathematically, a datum is defined by $\hat{\xi}_2 \rightarrow \xi_2^0$, where $\xi_2^0$ is known. The rank of $A_1$ given in (6.1b) implies that the inverse of the $q \times q$ matrix $N_{11}$ exists. Therefore, from the top row of (6.2), and with a given datum $\xi_2^0$ substituted for $\hat{\xi}_2$, we can write

$$
N_{11} \hat{\xi}_1 = c_1 - N_{12} \xi_2^0 \Rightarrow 
\hat{\xi}_1 = N_{11}^{-1}(c_1 - N_{12} \xi_2^0).
$$

(6.4b)

Equation (6.4b) shows that the datum can be chosen or modified after observations have been made and matrix $N_{11}$ has been inverted. Moreover, since the only random component in (6.4b) is $c_1$, we have

$$
D\{\hat{\xi}_1\} = \sigma^2_0 N_{11}^{-1}
$$

(6.5)

for the dispersion of the vector of estimated parameters $\hat{\xi}_1$.

The predicted random error (residual) vector and its dispersion are then defined as follows:

$$
\hat{e} = y - A \hat{\xi} = y - \begin{bmatrix}
A_1 \\
A_2
\end{bmatrix} \begin{bmatrix}
\hat{\xi}_1 \\
\xi_2^0
\end{bmatrix} = y - A_1 \hat{\xi}_1 - A_2 \xi_2^0,
$$

(6.6a)

$$
D\{\hat{e}\} = D\{y\} - D\{A_1 \hat{\xi}_1\} = \sigma^2_0 (P^{-1} - A_1 N_{11}^{-1} A^T_1).
$$

(6.6b)
Note that $C\{y, \xi_1\} = 0$, which is implied by (6.6b). After computing the residuals, it is straightforward to compute the vector of adjusted observations and its dispersion matrix, respectively, by

$$
\hat{\mu}_y = y - \hat{e} = A_1 \hat{\xi}_1 + A_2 \xi_0^0, \quad (6.7a)
$$

$$
D\{\hat{\mu}_y\} = D\{A_1 \hat{\xi}_1\} = \sigma_0^2 \cdot A_1 N_{11}^{-1} A_1^T. \quad (6.7b)
$$

Here, $\hat{\mu}_y$ is also interpreted as an estimate of the true, and thus unknown, vector of observables $\mu_y$, where $E\{y\} = \mu_y$.

The sum of squared residuals (SSR) is given by

$$
\Omega = \hat{e}^T P \hat{e}, \quad (6.8)
$$

while the redundancy of the model is provided by

$$
r = n - rk A = n - q. \quad (6.9)
$$

Substituting (6.6a) into (6.8), and considering (6.4a), leads to

$$
\hat{\sigma}_0^2 = \frac{\hat{e}^T P \hat{e}}{r} = \frac{y^T P y - c_1^T \hat{\xi}_1 - c_2^T \xi_0^0}{n - q} \quad (6.10)
$$

as an estimate for the unknown variance component $\sigma_0^2$. Here, the relation $\hat{\xi}_1^T N_{11} \hat{\xi}_1 + \hat{\xi}_2^T N_{12} \hat{\xi}_2 = \xi_1^T c_1$ has been used. However, since $rk A_1 = rk A = q$, the $n \times (m - q)$ submatrix $A_2$ must be in the column space of the $n \times q$ matrix $A_1$ so that $A_2 = A_1 L$ for some $q \times (m - q)$ matrix $L$. Therefore,

$$
N_{12} = A_1^T P A_2 = A_1^T P A_1 L = N_{11} L \Rightarrow \quad (6.11a)
$$

$$
N_{11}^{-1} N_{12} = L. \quad (6.11b)
$$

With this result, and using (6.4b), we have

$$
c_1^T \hat{\xi}_1 + c_2^T \xi_0^0 = y^T P A_1 (N_{11}^{-1} c_1 - N_{11}^{-1} N_{12} \xi_0^0) + y^T P A_2 \xi_0^0 =
= y^T P A_1 (N_{11}^{-1} c_1 - L \xi_0^0) + y^T P A_2 \xi_0^0 =
= y^T P A_1 N_{11}^{-1} c_1 - y^T P (A_1 L) \xi_0^0 + y^T P A_2 \xi_0^0 =
= y^T P A_1 N_{11}^{-1} c_1 = c_1^T N_{11}^{-1} c_1, \quad (6.12)
$$

which, upon substitution into (6.10), leads to

$$
\hat{\sigma}_0^2 = \frac{y^T P y - c_1^T N_{11}^{-1} c_1}{n - q} \quad (6.13)
$$

as an alternative form for the estimated variance component.

It is instructive to compare the dispersion of $\hat{\xi}_1$ shown in (6.5) with the corresponding dispersion in the case that matrix $A$ has full row rank, i.e., $rk A = m$. In
the full-rank case, we could invert the coefficient matrix of (6.2) and find the upper $q \times q$ block of the inverse, scaled by $\sigma_0^2$, to be the dispersion of $\hat{\xi}_1$. Referring to (A.14) for the inverse of the partitioned matrix $N$, we find

$$D\{\hat{\xi}_1\} = \sigma_0^2 [N_{11}^{-1} + N_{11}^{-1}N_{12}(N_{22} - N_{21}N_{11}^{-1}N_{12})^{-1}N_{21}N_{11}^{-1}] = \sigma_0^2 (N_{11} - N_{12}N_{22}^{-1}N_{21})^{-1} = \sigma_0^2 N_{11}^{-1} = D\{\hat{\xi}_1\}.$$  \hspace{1cm} (6.14)

The smaller dispersion in the last line of (6.14) shows that if a datum is introduced (increase in information), the unknown parameters $\xi$ are estimated with smaller variance.

6.1 Generation of Equivalent Condition Equations

We may also wish to transform the rank-deficient model (6.1a) into a Model of Condition Equations. To do so, consider the further splitting of the rank-deficient matrix $A$ defined in (6.1b) as follows:

$$A = \begin{bmatrix} A_1 & A_2 \end{bmatrix} = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}, \hspace{1cm} (6.15a)$$

$$\dim(A_{11}) = q \times q \text{ and } \dim(A_{22}) = (n - q) \times (m - q). \hspace{1cm} (6.15b)$$

Also, we have $\text{rk} A_{11} = q := \text{rk} A$. And, based on the definition of $L$ in the preceding section, we may write

$$A_2 = \begin{bmatrix} A_{12} \\ A_{22} \end{bmatrix} = A_1 L = \begin{bmatrix} A_{11} \\ A_{21} \end{bmatrix} L. \hspace{1cm} (6.16)$$

The matrix $B$ within the Model of Condition Equations could be defined as

$$B := A_{21}A_{11}^{-1} - I_{n-q}, \hspace{1cm} (6.17)$$

with $r$ being the redundancy of the model as shown in (6.9).

As discussed in Chapter 4, two conditions must be satisfied in order to reach an equivalent Model of Condition Equations:

i. dimensionality condition,

ii. orthogonality condition.

The first condition requires that the dimensions of the column spaces of $A$ and $B$ sum to the number of observations $n$. The second condition requires that the rows
of matrix $B$ are orthogonal to the columns of $A$, i.e., $BA = 0$. Taken together, these conditions mean that $A$ and $B^T$ are orthogonal complements in $n$-dimensional space, or, stated more succinctly,

$$\mathcal{R}(A) \oplus \mathcal{R}(B^T) = \mathbb{R}^n. \quad (6.18)$$

Both conditions i and ii are satisfied for (6.17) as shown below.

i. Dimensionality condition:

$$\text{rk } B = n - q = n - \text{rk } A \Rightarrow \text{rk } A + \text{rk } B = n \quad (6.19a)$$

ii. Orthogonality condition:

$$BA = B \begin{bmatrix} A_1 \\ A_2 \end{bmatrix} = BA \begin{bmatrix} I_q \\ L \end{bmatrix}, \quad (6.19b)$$

but

$$BA_1 = \begin{bmatrix} A_21 A_{11}^{-1} \\ -I_{n-q} \end{bmatrix} \begin{bmatrix} A_{11} \\ A_{21} \end{bmatrix} = A_{21} A_{11}^{-1} A_{11} - A_{21} = 0 \Rightarrow \quad (6.19c)$$

$$BA = 0 \quad (6.19d)$$

Note that as long as the rank of matrix $A$ is known, we can always generate a splitting of $A$ as shown in (6.15a); however, we may need to reorder the columns of $A$ (tantamount to reordering the elements of the parameter vector) to ensure that $A_{11}$ has full column rank.

6.2 Summary Formulas for the Introduction of Datum Information for the Least-Squares Solution Within the Rank Deficient Gauss-Markov Model

The rank deficient Gauss-Markov Model is given by

$$y_{n \times 1} = \begin{bmatrix} A_1 \\ n \times q \\ A_2 \\ n \times (m-q) \end{bmatrix} \begin{bmatrix} \xi_1 \\ q \times 1 \\ \xi_2 \\ (m-q) \times 1 \end{bmatrix} + e, \quad e \sim (0, \sigma_0^2 P^{-1}),$$

$$\text{rk } A =: q < m \quad \text{and} \quad \text{rk } A_1 = q.$$
<table>
<thead>
<tr>
<th>Quantity</th>
<th>Formula</th>
<th>Eq. No.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model redundancy</td>
<td>$r = n - rk A = n - q$</td>
<td>(6.9)</td>
</tr>
<tr>
<td>Vector of estimated parameters, with given $\xi_2^0$</td>
<td>$\hat{\xi}<em>i = N</em>{11}^{-1}(c_1 - N_{12} \xi_2^0)$</td>
<td>(6.4b)</td>
</tr>
<tr>
<td>Dispersion matrix for estimated parameters</td>
<td>$D{\hat{\xi}<em>1} = \sigma_0^2 N</em>{11}^{-1}$</td>
<td>(6.5)</td>
</tr>
<tr>
<td>Vector of predicted residuals, with given $\xi_2^0$</td>
<td>$\bar{e} = y - A\hat{\xi} = y - A_1\hat{\xi}_1 - A_2 \xi_2^0$</td>
<td>(6.6b)</td>
</tr>
<tr>
<td>Dispersion matrix for residuals</td>
<td>$D{\bar{e}} = \sigma_0^2 (P^{-1} - A_1 N_{11}^{-1} A_1^T)$</td>
<td>(6.6b)</td>
</tr>
<tr>
<td>Sum of squared residuals (SSR)</td>
<td>$\Omega = \bar{e}^T P \bar{e}$</td>
<td>(6.8)</td>
</tr>
<tr>
<td>Estimated variance component, with given $\xi_2^0$</td>
<td>$\sigma_0^2 = (\bar{e}^T P \bar{e})/r = (y^T Py - c_1^T \hat{\xi}_1 - c_2^T \xi_2^0)/(n - q)$</td>
<td>(6.10)</td>
</tr>
<tr>
<td>Vector of adjusted observations</td>
<td>$\hat{\mu}_y = y - \bar{e} = A_1\hat{\xi}_1 + A_2 \xi_2^0$</td>
<td>(6.7a)</td>
</tr>
<tr>
<td>Dispersion matrix for adjusted observations</td>
<td>$D{\hat{\mu}<em>y} = \sigma_0^2 A_1 N</em>{11}^{-1} A_1^T$</td>
<td>(6.7b)</td>
</tr>
</tbody>
</table>
The Gauss-Markov Model with Stochastic Constraints

The Gauss-Markov Model (GMM) with stochastic constraints is similar in form to the GMM with constraints shown in Chapter 5, with one important difference: the constraints in the stochastic case are specified with some level of uncertainty, expressed in the form of a given weight matrix $P_0$, or an associated cofactor matrix $Q_0 := P_0^{-1}$. The model reads

$$
\begin{align*}
\mathbf{y}_{n \times 1} &= \mathbf{A}_{n \times m} \mathbf{\xi} + \mathbf{e}, \quad \text{rk} \mathbf{A} =: q \leq \min\{m, n\}, \\
\mathbf{z}_{0\times1} &= \mathbf{K}_{l \times m} \mathbf{\xi} + \mathbf{e}_0, \quad \text{rk} \mathbf{K} =: l \geq m - q,
\end{align*}
$$

(7.1a)

$$
\begin{bmatrix}
\mathbf{e} \\
\mathbf{e}_0
\end{bmatrix} \sim \left( \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \sigma^2_0 \begin{bmatrix} P^{-1} & 0 \\ 0 & P_0^{-1} \end{bmatrix} \right).
$$

(7.1c)

Note that in this model there is no correlation between the random error vectors $\mathbf{e}$ and $\mathbf{e}_0$. Also, the unknown variance component $\sigma^2_0$ is common to both cofactor matrices $P^{-1}$ and $P_0^{-1}$. However, there may be correlations within one or both of the cofactor matrices, just not between them. Depending on the application, the data in the vector $\mathbf{y}$ can be thought of as new information, while the constraint information in the vector $\mathbf{z}_0$ can be thought of as prior information (for example, $\mathbf{z}_0$ could contain coordinates estimated from a previous adjustment, now considered as prior information).

It is required that the column space of the augmented matrix $[\mathbf{A}^T \mid \mathbf{K}^T]$ spans $\mathbb{R}^m$, which holds when the rank condition

$$
\text{rk}[\mathbf{A}^T \mid \mathbf{K}^T] = m
$$

(7.2)

is satisfied.
The least-squares solution (LESS) for the unknown parameters \( \xi \) within model (7.1) may be derived by minimizing the Lagrange target function

\[
\Phi(\xi, \lambda) = e^T Pe + 2\lambda^T (K\xi - z_0) - \lambda^T P_0^{-1} \lambda = \text{stationary.} \tag{7.3}
\]

Here we simply consider (7.1) as an extended GMM and apply the addition theory of normal equations as follows:

\[
\begin{bmatrix} P & 0 \\ 0 & P_0 \end{bmatrix} \begin{bmatrix} A \\ K \end{bmatrix} \cdot \hat{\xi} = \begin{bmatrix} P & 0 \\ 0 & P_0 \end{bmatrix} \begin{bmatrix} y \\ z_0 \end{bmatrix}
\]

or

\[
(N + K^T P_0 K) \hat{\xi} = c + K^T P_0 z_0, \tag{7.4}
\]

where

\[
[N, c] := A^T P[A, y]. \tag{7.5}
\]

In the case where the matrix \( N \) is invertible, the Sherman-Morrisson-Woodbury-Schur formula (A.7) may be used to invert the matrix on the left side of (7.4) as in the following:

\[
\hat{\xi} = (N + K^T P_0 K)^{-1}(c + K^T P_0 z_0) = \tag{7.6a}
\]

\[
= \left[ N^{-1} - N^{-1} K^T (P_0^{-1} + KN^{-1} K^T)^{-1} KN^{-1} \right] (c + K^T P_0 z_0) =
\]

\[
= N^{-1} c + N^{-1} K^T P_0 z_0 + N^{-1} K^T (P_0^{-1} + KN^{-1} K^T)^{-1}.
\]

\[
\cdot (-KN^{-1} c - KN^{-1} K^T P_0 z_0) =
\]

\[
= N^{-1} c + N^{-1} K^T (P_0^{-1} + KN^{-1} K^T)^{-1}[(P_0^{-1} + KN^{-1} K^T) P_0 z_0 - KN^{-1} c -
\]

\[
- KN^{-1} K^T P_0 z_0] \Rightarrow
\]

\[
\hat{\xi} = N^{-1} c + N^{-1} K^T (P_0^{-1} + KN^{-1} K^T)^{-1} (z_0 - KN^{-1} c). \tag{7.6b}
\]

Thus, the LESS (7.6b) can be viewed as a weighted average between the prior and the new information. The vector \( z_0 - KN^{-1} c \) is referred to as a vector of discrepancies. The solution can also be recognized as an update to the solution \( \xi = N^{-1} c \) within the GMM (3.1). It is also interesting to express it as an update to the LESS within the GMM with “fixed” constraints (5.1). This can be done by changing the symbols \( \hat{\xi} \) and \( \kappa_0 \) in (5.7d) to \( \hat{\xi}_K \) and \( z_0 \), respectively, solving for \( N^{-1} c \) in terms of these renamed variables, and substituting into (7.6b), which yields the following:

\[
\hat{\xi} = \hat{\xi}_K + N^{-1} K^T [(P_0^{-1} + KN^{-1} K^T)^{-1} - (KN^{-1} K^T)^{-1}] (z_0 - KN^{-1} c). \tag{7.7}
\]

By applying the laws of covariance propagation to (7.6a), the dispersion of the vector of estimated parameters \( \xi \) is computed as follows:

\[
D\{\hat{\xi}\} = (N + K^T P_0 K)^{-1} D\{c + K^T P_0 z_0\} (N + K^T P_0 K)^{-1} =
\]

\[
60
\]
\[
\begin{aligned}
&= \sigma_0^2 (N + K^T P_0 K)^{-1} (N + K^T P_0 K) (N + K^T P_0 K)^{-1} \Rightarrow \\
D\{\xi\} = \sigma_0^2 (N + K^T P_0 K)^{-1} = \sigma_0^2 [N^{-1} - N^{-1} K^T (P_0^{-1} + KN^{-1} K^T)^{-1} K N^{-1}].
\end{aligned}
\] (7.8)

The subtraction in (7.8) implies that our knowledge of the parameters has improved (variance decreased) by supplying the additional prior information, provided the estimated variance component \(\hat{\sigma}_0^2\) does not change much in doing so. Indeed, if the new data is consistent with the old, \(\hat{\sigma}_0^2\) is not expected to change very much. In contrast, \(\hat{\sigma}_0^2\) is expected to increase if there is inconsistency between the old and new information. In such a case, it may be advised to introduce a second variance component, with one is associated with the weight matrix \(P\) and the other with \(P_0\). This is the purpose of the variance component model, which is introduced in the advanced adjustments notes.

We now present the residual vectors \(\tilde{e}\) and \(\hat{e}_0\) (also called predicted random error vectors). The residual vector \(\tilde{e}\) for the observations \(y\) is computed by

\[
\tilde{e} = y - A\hat{\xi}.
\] (7.9)

The residual vector \(\hat{e}_0\) associated with the prior information \(z_0\) is

\[
\begin{aligned}
\hat{e}_0 &= z_0 - K\xi - \\
&= (z_0 - KN^{-1}c) - (KN^{-1}K^T + P_0^{-1} - P_0^{-1})(P_0^{-1} + KN^{-1}K^T)^{-1}. \\
&= (z_0 - KN^{-1}c) - \left[ (KN^{-1}K^T + P_0^{-1})(P_0^{-1} + KN^{-1}K^T)^{-1} - P_0^{-1}(P_0^{-1} + KN^{-1}K^T)^{-1} \right] (z_0 - KN^{-1}c) = \\
&= \{I - [I - P_0^{-1}(P_0^{-1} + KN^{-1}K^T)^{-1}]\} (z_0 - KN^{-1}c) = \\
&= P_0^{-1}(P_0^{-1} + KN^{-1}K^T)^{-1} (z_0 - KN^{-1}c) \Rightarrow \\
\hat{e}_0 &= (I + KN^{-1}K^T P_0)^{-1} (z_0 - KN^{-1}c).
\end{aligned}
\] (7.10a)

The dispersion matrix of the residual vectors is derived as follows:

\[
\begin{aligned}
D\{\begin{bmatrix} \tilde{e} \\ \hat{e}_0 \end{bmatrix}\} &= D\{\begin{bmatrix} y \\ z_0 \end{bmatrix}\} - D\{\begin{bmatrix} A \\ K \end{bmatrix}\} \hat{\xi} = \\
&= \sigma_0^2 \begin{bmatrix} P^{-1} & 0 \\ 0 & P_0^{-1} \end{bmatrix} - \sigma_0^2 \begin{bmatrix} A \\ K \end{bmatrix} [N^{-1} - N^{-1} K^T (P_0^{-1} + KN^{-1} K^T)^{-1} K N^{-1}].
\end{aligned}
\]
From (7.11), we can write the dispersion matrices for the residual vectors individually as

\[
D\{\hat{e}\} = \sigma_0^2 (P^{-1} - AN^{-1}A^T) + \sigma_0^2 AN^{-1}K^T (P_0^{-1} + KN^{-1}K^T)^{-1} KN^{-1}A^T \Rightarrow (7.12a)
\]

\[
D\{\hat{e}\} = \sigma_0^2 [P^{-1} - A(N + K^TP_0K)^{-1}A^T],
\]

and

\[
D\{\hat{e}_0\} = \sigma_0^2 P_0^{-1} - \sigma_0^2 KN^{-1}K^T + \sigma_0^2 KN^{-1}K^T (P_0^{-1} + KN^{-1}K^T)^{-1} KN^{-1}K^T =
\]

\[
= \sigma_0^2 P_0^{-1} - \sigma_0^2 KN^{-1}K^T (P_0^{-1} + KN^{-1}K^T)^{-1} - \sigma_0^2 KN^{-1}K^T.
\]

\[
\Rightarrow D\{\hat{e}_0\} = \sigma_0^2 P_0^{-1} (I + P_0KN^{-1}K^T)^{-1}.
\]

We summarize by listing a few equivalent formulas for \(D\{\hat{e}_0\}\).

\[
D\{\hat{e}_0\} = \sigma_0^2 P_0^{-1} (I + P_0KN^{-1}K^T)^{-1} = (7.14a)
\]

\[
= \sigma_0^2 (I + KN^{-1}K^TP_0)^{-1} P_0^{-1} = (7.14b)
\]

\[
= \sigma_0^2 P_0^{-1} (P_0^{-1} + KN^{-1}K^T)^{-1} P_0^{-1} = (7.14c)
\]

\[
= \sigma_0^2 (P_0 + P_0KN^{-1}K^TP_0)^{-1} = (7.14d)
\]

\[
= \sigma_0^2 [P_0^{-1} - K(N + K^TP_0K)^{-1}K^T] (7.14e)
\]

The symmetry of the matrix \(D\{\hat{e}_0\}\) has been exploited to get from (7.14a) to (7.14b), using the rule for the transpose of a matrix product (A.1) and the rule for the transpose of an inverse (A.2). Also (A.3) has been used in the above.
Now it remains to write a succinct form for the covariance matrix $C\{\hat{e}, \hat{e}_0\}$, beginning with the off-diagonal element of (7.11).

\[
C\{\hat{e}, \hat{e}_0\} = -\sigma_0^2 AN^{-1}K^T + \sigma_0^2 AN^{-1}K^T (P_0^{-1} + KN^{-1}K^T)^{-1}KN^{-1}K^T = 
\]

(7.15a)

\[
= -\sigma_0^2 AN^{-1}K^T (P_0^{-1} + KN^{-1}K^T)^{-1} (P_0^{-1} + KN^{-1}K^T - KN^{-1}K^T) = 
\]

(7.15b)

\[
= -\sigma_0^2 AN^{-1}K^T (I + P_0KN^{-1}K^T)^{-1} = 
\]

(7.15c)

\[
= -\sigma_0^2 A(N + K^TP_0K)^{-1}K^T = 
\]

(7.15d)

\[
= -\sigma_0^2 A(N + K^TP_0K)^{-1}K^T = 
\]

(7.15e)

The line following (7.15c) is based on relations shown in equations (A.9). To see how these equations are used, compare what follows the term $-\sigma_0^2A$ in (7.15c) and (7.15d), with the first and last lines in (A.9).

Further insight may be gained by minimizing a Lagrange target function, as in Snow (2002). This leads to the following system of normal equations, which includes an estimated vector of Lagrange multipliers $\hat{\lambda}$:

\[
\begin{bmatrix}
N & K^T \\
K & -P_0^{-1}
\end{bmatrix}
\begin{bmatrix}
\hat{\xi} \\
\hat{\lambda}
\end{bmatrix}
= 
\begin{bmatrix}
c \\
z_0
\end{bmatrix}.
\]

(7.16)

Using (7.1b) and (7.16), we can express the predicted residual vector $\hat{e}_0$ as a function of the vector of Lagrange multipliers $\hat{\lambda}$ as follows:

\[
z_0 = K^T \hat{\xi} + \hat{e}_0 = K^T \hat{\xi} - P_0^{-1} \hat{\lambda} \Rightarrow \hat{e}_0 = -P_0^{-1} \hat{\lambda}.
\]

(7.17)

Therefore, the dispersion of $\hat{e}_0$ is given also by

\[
D\{\hat{e}_0\} = P_0^{-1}D\{\hat{\lambda}\}P_0^{-1}.
\]

(7.18)

Assuming matrix $N$ is invertible, from (7.16) we see that the dispersion of $\hat{\lambda}$ can be found from

\[
D\{\hat{\lambda}\} = 
\begin{bmatrix}
N & K^T \\
K & -P_0^{-1}
\end{bmatrix}
D\{\hat{\lambda}\}
\begin{bmatrix}
N & K^T \\
K & -P_0^{-1}
\end{bmatrix}
= 
\sigma_0^2
\begin{bmatrix}
N & K^T \\
K & -P_0^{-1}
\end{bmatrix}
N^{-1}
\begin{bmatrix}
N & K^T \\
K & -P_0^{-1}
\end{bmatrix}^{-1}
= 
\sigma_0^2
\begin{bmatrix}
N & K^T \\
K & -P_0^{-1}
\end{bmatrix}
N^{-1}
\begin{bmatrix}
N & K^T \\
K & -P_0^{-1}
\end{bmatrix}^{-1}
= 
\sigma_0^2
\begin{bmatrix}
N & K^T \\
K & -P_0^{-1}
\end{bmatrix}
N^{-1}
\begin{bmatrix}
N & K^T \\
K & -P_0^{-1}
\end{bmatrix}^{-1}
= 
\sigma_0^2
\begin{bmatrix}
N & K^T \\
K & -P_0^{-1}
\end{bmatrix}
N^{-1}
\begin{bmatrix}
N & K^T \\
K & -P_0^{-1}
\end{bmatrix}^{-1}.
\]
\[
\sigma_0^2 \begin{bmatrix}
N + K^T P_0 K & 0 \\
0 & P_0^{-1} + KN^{-1}K^T
\end{bmatrix}^{-1}.
\] (7.19)

The last line was reached by successively applying the rule for the product of two inverses (A.3). From (7.19) we see that

\[
D\{\hat{\lambda}\} = \sigma_0^2 \left( P_0^{-1} + KN^{-1}K^T \right)^{-1} = \sigma_0^2 \left[ P_0 - P_0K(N + K^T P_0 K)^{-1} K^T P_0 \right].
\] (7.20)

Finally, applying the product-of-inverses rule to (7.18), we can write

\[
D\{\hat{e}_0\} = \sigma_0^2 P_0^{-1} \left( P_0^{-1} + KN^{-1}K^T \right)^{-1} P_0^{-1} = \sigma_0^2 \left( P_0 + P_0K(N + K^T P_0 K)^{-1} K^T P_0 \right)^{-1}.
\] (7.21)

Also, we see from (7.19) that

\[
C(\hat{\xi}, \hat{\lambda}) = 0.
\] (7.22)

We also note that in the GMM with stochastic constraints, the predicted residual vector \(\hat{e} = y - A\hat{\xi}\) by itself is no longer a projection of \(y\). However, the vector \(\begin{bmatrix} \hat{e}^T, \hat{e}_0^T \end{bmatrix}^T\) does represent a projection of \(\begin{bmatrix} y^T, z_0^T \end{bmatrix}^T\) since

\[
\begin{bmatrix}
\hat{e} \\
\hat{e}_0
\end{bmatrix} = \begin{bmatrix}
y - A\hat{\xi} \\
z_0 - K\hat{\xi}
\end{bmatrix} = \left\{ \begin{bmatrix}
I_n & 0 \\
0 & I_l
\end{bmatrix} - \begin{bmatrix}
A \\
K
\end{bmatrix} (N + K^T P_0 K)^{-1} \begin{bmatrix}
A^T P & K^T P_0
\end{bmatrix} \right\} \begin{bmatrix}
y \\
z_0
\end{bmatrix},
\] (7.23)

and the matrix in braces is idempotent, which can be verified by application of (A.11).

The adjusted observations and adjusted constraint values are easily computed by

\[
\mu_y = y - \hat{e} = A\hat{\xi},
\] (7.24)

and

\[
\mu_z = z_0 - \hat{e}_0 = K\hat{\xi}.
\] (7.25)

Their respective dispersion matrices are derived by simple application of variance propagation as follows:

\[
D(\mu_y) = D(A\hat{\xi}) = A \cdot D(\hat{\xi}) \cdot A^T = \sigma_0^2 \cdot A(N + K^T P_0 K)^{-1} A^T,
\] (7.26)

\[
D(\mu_z) = D(K\hat{\xi}) = K \cdot D(\hat{\xi}) \cdot K^T = \sigma_0^2 \cdot K(N + K^T P_0 K)^{-1} K^T.
\] (7.27)

Here, \(\mu_y\) is also interpreted as an estimate of the true, and thus unknown, vector of observables \(\mu_y\), where \(E\{y\} = \mu_y\); likewise, \(E\{z_0\} = \mu_z\).
7.1 Variance Component Estimate

The derivation of the variance component estimate is shown here in detail. The trace operator is employed analogously to what was done in Section 3.2. We also make use of the following expectation and dispersion relationships:

\[ E\{c + K^TP_0z_0\} = \left[\begin{array}{cc} A^T & K^TP_0 \end{array}\right] E\left[\begin{array}{c} y \\ z_0 \end{array}\right] = \]

\[ = \left[\begin{array}{cc} A^T & K^TP_0 \end{array}\right] \begin{bmatrix} A \\ K \end{bmatrix} \xi = (N + K^TP_0K)\xi, \quad (7.28a) \]

\[ D\{c + K^TP_0z_0\} = D\left[\begin{array}{cc} A^T & K^TP_0 \end{array}\right] \begin{bmatrix} y \\ z_0 \end{bmatrix} = \]

\[ = \sigma_0^2 \left[\begin{array}{cc} A^T & K^TP_0 \end{array}\right] \begin{bmatrix} P^{-1} & 0 \\ 0 & P_0^{-1} \end{bmatrix} \begin{bmatrix} PA \\ P_0K \end{bmatrix} = \sigma_0^2(N + K^TP_0K), \quad (7.29a) \]

as well as

\[ E\{(c + K^TP_0z_0)(c + K^TP_0z_0)^T\} = D\{c + K^TP_0z_0\} + E\{c + K^TP_0z_0\}E\{c + K^TP_0z_0\}^T, \quad (7.30a) \]

\[ E\{yy^T\} = D\{y\} + E\{y\}E\{y\}^T = \sigma_0^2P^{-1} + A\xi\xi^TA^T, \quad (7.30b) \]

\[ E\{z_0z_0^T\} = D\{z_0\} + E\{z_0\}E\{z_0\}^T = \sigma_0^2P_0^{-1} + K\xi\xi^TK. \quad (7.30c) \]

The estimated variance component is derived from the expectation of the combined quadratic forms of the residual vectors, \( \hat{e}^T\hat{P}\hat{e} + \hat{e}_0^T\hat{P}_0\hat{e}_0 \), as follows:

\[ E\{\hat{e}^T\hat{P}\hat{e} + \hat{e}_0^T\hat{P}_0\hat{e}_0\} = \]

\[ = E\left( \begin{bmatrix} y \\ z_0 \end{bmatrix} - \begin{bmatrix} A \\ K \end{bmatrix} \hat{\xi}^T \begin{bmatrix} P & 0 \\ 0 & P_0 \end{bmatrix} \begin{bmatrix} y \\ z_0 \end{bmatrix} - \begin{bmatrix} A \\ K \end{bmatrix} \hat{\xi} \right) = \]

\[ = E\{y^T\hat{P}y + z_0^TP_0z_0 - 2\hat{\xi}^T(c + K^TP_0z_0) + \hat{\xi}^T(N + K^TP_0K)\hat{\xi}\} = \]

\[ = E\{y^T\hat{P}y + z_0^TP_0z_0 - 2\hat{\xi}^T(c + K^TP_0z_0) + \hat{\xi}^T(c + K^TP_0z_0)\} = \]

\[ = E\{y^T\hat{P}y + z_0^TP_0z_0 - (c + K^TP_0z_0)^T(N + K^TP_0K)^{-1}(c + K^TP_0z_0)\} = \]

\[ = E\{\text{tr}(y^T\hat{P}y) + \text{tr}(z_0^TP_0z_0) - \text{tr}[(c + K^TP_0z_0)^T(N + K^TP_0K)^{-1} \cdot (c + K^TP_0z_0)]\} = \]
\[
\hat{\sigma}^2 = \frac{\tilde{e}^T P \tilde{e} + \tilde{e}_0^T P_0 \tilde{e}_0}{n - m + l}
\]

(7.31)

provides an estimate of the variance component \(\sigma^2\). Here, the numerator contains the sum of squared residuals

\[
\text{SSR} : \tilde{e}^T P \tilde{e} + \tilde{e}_0^T P_0 \tilde{e}_0,
\]

(7.32)

while the denominator contains the model redundancy

\[
r = n - m + l.
\]

(7.33)

### 7.2 Hypothesis Test for the Estimated Variance Component

Hypothesis testing can be used to validate that the least-squares solution is consistent with the stochastic constraints in the model (7.1). The test statistic to be computed is comprised of a ratio of two estimated, and therefore random, variances and thus has an \(F\)-distribution (see Section 10.4). The idea is to extract from the sum of the quadratic products in (7.32) the associated sum of squared residuals that would have been computed for the LESS within the unconstrained GMM solution, i.e., \(\xi_u = N^{-1}c\), had it been estimated. We label this quantity \(\Omega\). What
remains after extracting $\Omega$ from (7.32) is a quantity that depends on the weight matrix $P_0$. We denote this remaining portion as $R(P_0)$ to indicate that it is a function of $P_0$. Both $\Omega$ and $R(P_0)$ are scalars, and both have random properties. These two variables, which are used to form the test statistic, are defined as follows:

$$\Omega := (y - AN^{-1}c)^T P (y - AN^{-1}c) = y^T P y - c^T N^{-1}c,$$  \hspace{0.5cm} (7.34a)

$$R(P_0) := \hat{e}^T P \hat{e} + \hat{e}_0^T P_0 \hat{e}_0 - \Omega.$$ \hspace{0.5cm} (7.34b)

Note: If the matrix $N$ in (7.34a) is singular, then $N^{-1}$ can be replaced with any generalized inverse of $N$ as discussed on page 51.

Again we note that $\hat{\xi}_u = \lambda$ represents the least-squares solution within model (7.1) had the stochastic constraints been omitted. In the following derivations, we also make use of (7.6b), (7.10a), (7.10b), (7.17), and (7.18) to write formulas for $\hat{e}_0$ and $\hat{\xi}$ in terms of $\xi_u$ as follows:

$$\hat{e}_0 = z_0 - K \hat{\xi} = -P_0^{-1} \hat{\lambda} = (I_l + KN^{-1}KT_0)^{-1} (z_0 - K \hat{\xi}_u),$$  \hspace{0.5cm} (7.35)

$$\hat{\xi} = \hat{\xi}_u + N^{-1}KT_0 \hat{e}_0 = \hat{\xi}_u - N^{-1}K^T \hat{\lambda}. \hspace{0.5cm} (7.36)$$

We now begin with the quadratic form for the full predicted residual vector appearing in (7.32) (also called sum of squared residuals, SSR) and decompose it into $\Omega$ and $R(P_0)$. The crossed-out vector in the first line below is neglected since its contribution vanishes in the quadratic product.

$$\hat{e}^T P \hat{e} + \hat{e}_0^T P_0 \hat{e}_0 = \begin{bmatrix} y \\ z_0 \end{bmatrix}^T \begin{bmatrix} A \\ K \end{bmatrix}^T \begin{bmatrix} P & 0 \\ 0 & P_0 \end{bmatrix} \begin{bmatrix} y \\ z_0 \end{bmatrix} - \begin{bmatrix} A \\ K \end{bmatrix}^T \hat{\xi} =

= y^T P y - y^T PA \xi + \begin{bmatrix} 0 \\ z_0 \end{bmatrix}^T P_0 \begin{bmatrix} 0 \\ z_0 \end{bmatrix} - \begin{bmatrix} 0 \\ z_0 \end{bmatrix}^T P_0 K \hat{\xi} =

= y^T P y - y^T PA \xi_u + \begin{bmatrix} 0 \\ z_0 \end{bmatrix}^T P_0 z_0 - \begin{bmatrix} 0 \\ z_0 \end{bmatrix}^T P_0 K \hat{\xi}_u + \begin{bmatrix} 0 \\ z_0 \end{bmatrix}^T P_0 \hat{e}_0 =

= \begin{bmatrix} y^T P y & y^T PA \hat{\xi}_u \end{bmatrix} + \begin{bmatrix} z_0^T P_0 \\ \hat{e}_0 \end{bmatrix} \begin{bmatrix} \hat{\xi}_u^T (N^{-1} K^T P_0) & (c + \hat{\lambda} K^T P_0 z_0)^T \\ N^{-1} K^T P_0 \hat{\xi}_u \end{bmatrix} =

= \Omega + \hat{e}_0^T \begin{bmatrix} 0 \\ \hat{e}_0 \end{bmatrix} (I_l + P_0 K N^{-1} K^T P_0) \hat{e}_0 =

= \Omega + \begin{bmatrix} 0 \\ \hat{\xi}_u \end{bmatrix}^T (I_l + P_0 KN^{-1} K^T) \begin{bmatrix} 0 \\ \hat{\xi}_u \end{bmatrix} =

= \Omega + (z_0 - K \hat{\xi})^T (I_l + P_0 KN^{-1} K^T) \begin{bmatrix} 0 \\ \hat{\xi}_u \end{bmatrix} =

= \Omega + (z_0 - K \hat{\xi}_u)^T (I_l + P_0 KN^{-1} K^T) (z_0 - K \hat{\xi}_u) =

= \Omega + R(P_0)

Thus, $R(P_0)$ is defined as

$$R(P_0) := (z_0 - K \hat{\xi}_u)^T (P_0^{-1} + KN^{-1} K^T)^{-1} (z_0 - K \hat{\xi}_u). \hspace{0.5cm} (7.37)$$
Finally, the test statistic \( t \) can be expressed as a ratio of \( R(P_0) \) to \( \Omega \), viz.

\[
t = \frac{(\hat{e}^T P \hat{e} + \hat{e}_0^T P_0 \hat{e}_0 - \Omega)/(l - m + q)}{\Omega/(n - q)} = \frac{R(P_0)/(l - m + q)}{\Omega/(n - q)} \sim F(l - m + q, n - q). \tag{7.38}
\]

Recall from (7.2) that \( l := \text{rk}(K) \) and \( q := \text{rk}(A) \).

The following hypothesis test can now be performed, where \( \mathcal{N} \) stands for the normal distribution and \( z_0 \) is an unknown quantity:

\[
H_0 : z_0 \sim \mathcal{N}(K\xi, \sigma_0^2 P_0^{-1}) \quad \text{against} \quad H_a : z_0 \sim \mathcal{N}(\kappa_0 \neq K\xi, \sigma_0^2 P_0^{-1}). \tag{7.39}
\]

The term \( H_0 \) is called the null hypothesis, and \( H_a \) is the alternative hypothesis. After choosing a level of significance \( \alpha \) and taking \( F_{\alpha, l - m + q, n - q} \) from a table of critical values for the \( F \)-distribution, the following logic can be applied:

\[
\text{If } t \leq F_{\alpha, l - m + q, n - q} \text{ accept } H_0; \text{ else reject } H_0. \tag{7.40}
\]

### 7.3 Some Comments on Reproducing Estimators

In this section we briefly discuss two estimators within the Gauss-Markov Model with stochastic constraints (7.1) that leave the constrained parameters unchanged, i.e., unchanged from the values specified in \( z_0 \). Such estimators are called reproducing estimators. For example, in a network adjustment problem the a priori coordinates of a station might need to be left unchanged by the adjustment.

For simplicity, we restrict the discussion to models of full rank, i.e., \( \text{rk} A = m \), where \( m \) is the number of columns of matrix \( A \) and also the number of parameters to estimate.

One approach that is sometimes taken to obtain a reproducing estimator is to simply adopt the estimator within the Gauss-Markov Model with fixed constraints shown in (5.7d), which is optimal for that model. Two points should be made regarding the use of that estimator within the model (7.1). First, it is not an optimal estimator within model (7.1), and, second, its dispersion matrix shown in (5.17a) and (5.18a) is not correct within model (7.1). In the following, we show the proper dispersion matrix for the reproducing estimator within model (7.1). First, we introduce different subscripts to denote various linear estimators for \( \xi \).

\( \hat{\xi}_U \) denotes the unconstrained estimator \( \hat{\xi}_U = N^{-1} c \), which is not optimal within model (7.1).

\( \hat{\xi}_K \) denotes the reproducing estimator from equation (5.7d), which is not optimal within model (7.1).

\( \hat{\xi}_S \) denotes the estimator from equation (7.6a), which is optimal within model (7.1).
First we express the estimator $\hat{\xi}_K$ as a function of the optimal estimator $\hat{\xi}_s$. Using (7.4), we can write

$$\left(N + K^T P_0 K\right)^{-1} c = \hat{\xi}_s - \left(N + K^T P_0 K\right)^{-1} K^T P_0 z_0. \quad (7.41)$$

We then repeat (5.7d) for the estimator $\hat{\xi}_K$ with $N$ replaced by $\left(N + K^T P_0 K\right)$ and $\kappa_0$ replaced by $z_0$ according to the model (7.1). This is our starting point.

$$\hat{\xi}_K = \left(N + K^T P_0 K\right)^{-1} c + \left(N + K^T P_0 K\right)^{-1} K^T \left[K\left(N + K^T P_0 K\right)^{-1} K^T\right]^{-1} \left[z_0 - K\left(N + K^T P_0 K\right)^{-1} c\right] \quad (7.42)$$

Now using (7.41) in (7.42), we can write

$$\hat{\xi}_K = \hat{\xi}_s - \left(N + K^T P_0 K\right)^{-1} K^T P_0 z_0 + \left(N + K^T P_0 K\right)^{-1} K^T \left[K\left(N + K^T P_0 K\right)^{-1} K^T\right]^{-1} \left[z_0 - K\left(N + K^T P_0 K\right)^{-1} c\right].$$

Factoring out $-\left(N + K^T P_0 K\right)^{-1} K^T \left[K\left(N + K^T P_0 K\right)^{-1} K^T\right]^{-1}$ yields

$$\hat{\xi}_K = \hat{\xi}_s - \left(N + K^T P_0 K\right)^{-1} K^T \left[K\left(N + K^T P_0 K\right)^{-1} K^T\right]^{-1} \left[\left[K\left(N + K^T P_0 K\right)^{-1} K^T\right]P_0 z_0 - z_0 + K\left(N + K^T P_0 K\right)^{-1} c\right].$$

Now, from (7.6a) we recognize $K\hat{\xi}_S$ in the above line; thus we write:

$$\hat{\xi}_K = \hat{\xi}_s + \left(N + K^T P_0 K\right)^{-1} K^T \left[K\left(N + K^T P_0 K\right)^{-1} K^T\right]^{-1} \left(z_0 - K\hat{\xi}_S\right). \quad (7.43)$$

We now have the fixed-constraint estimator $\hat{\xi}_K$ expressed as a function of the optimal estimator for model (7.1), namely $\hat{\xi}_s$. Using a familiar formula for $\left(N + K^T P_0 K\right)^{-1}$ and noting that

$$\left(N + K^T P_0 K\right)^{-1} K^T P_0 = N^{-1} K^T \left(P_0^{-1} + K N^{-1} K^T\right)^{-1},$$

we can rewrite (7.43) as:

$$\hat{\xi}_K = \hat{\xi}_s + \left[N^{-1} - N^{-1} K^T \left(P_0^{-1} + K N^{-1} K^T\right)^{-1} K N^{-1}\right]K^T \cdot \left[K N^{-1} K^T \left(P_0^{-1} + K N^{-1} K^T\right)^{-1} P_0^{-1}\right]^{-1} \left(z_0 - K\hat{\xi}_S\right). \quad (7.44)$$

Note the following useful relations:

$$\left[N^{-1} - N^{-1} K^T \left(P_0^{-1} + K N^{-1} K^T\right)^{-1} K N^{-1}\right]K^T =$$

$$= N^{-1} K^T \left(P_0^{-1} + K N^{-1} K^T\right)^{-1} P_0^{-1} \quad (7.45)$$

and
\[
[KN^{-1}K^T(P_0^{-1} + KN^{-1}K^T)^{-1}P_0^{-1}]^{-1} = P_0(P_0^{-1} + KN^{-1}K^T)(KN^{-1}K^T)^{-1}.
\] (7.46)

Equation (7.45) is derived as follows:
\[
[N^{-1} - N^{-1}K^T(P_0^{-1} + KN^{-1}K^T)^{-1}KN^{-1}]K^T = \\
N^{-1}K^T - N^{-1}K^T(P_0^{-1} + KN^{-1}K^T)^{-1}(P_0^{-1} + KN^{-1}K^T - P_0^{-1}) = \\
N^{-1}K^T - N^{-1}K^T(P_0^{-1} + KN^{-1}K^T)^{-1}(P_0^{-1} + KN^{-1}K^T) - \\
N^{-1}K^T(P_0^{-1} + KN^{-1}K^T)^{-1}(-P_0^{-1}) = \\
N^{-1}K^T - N^{-1}K^T + N^{-1}K^T(P_0^{-1} + KN^{-1}K^T)^{-1}P_0^{-1} = \\
N^{-1}K^T(P_0^{-1} + KN^{-1}K^T)^{-1}P_0^{-1}.
\]

Successive application of the rule for the product of inverted matrices was used in equation (7.46). Substituting (7.45) and (7.46) into (7.44) yields:
\[
\hat{\xi}_K = \hat{\xi}_S + N^{-1}K^T(P_0^{-1} + KN^{-1}K^T)^{-1}P_0^{-1}P_0(P_0^{-1} + KN^{-1}K^T) \\
\cdot(KN^{-1}K^T)^{-1}(z_0 - K\hat{\xi}_S) = \\
\hat{\xi}_S + N^{-1}K^T(KN^{-1}K^T)^{-1}(z_0 - K\hat{\xi}_S).
\] (7.47)

Equation (7.47) gives an elegant expression of the fixed-constraint estimator \(\hat{\xi}_K\) in terms of the optimal estimator \(\hat{\xi}_S\). Realizing that the model with stochastic constraints (7.1) becomes the model with fixed constraints (5.1) when \(P_0^{-1}\) is replaced by zero, we can replace (7.47) with (7.48) below, which is also obvious from our starting equation (7.42). This also makes the appropriate dispersion matrix \(D\{\hat{\xi}_K\}\) under model (7.1) easier to compute.
\[
\hat{\xi}_K = \hat{\xi}_U + N^{-1}K^T(KN^{-1}K^T)^{-1}(z_0 - K\hat{\xi}_U)
\] (7.48)

Note that \(C\{z_0, y\} = 0\), which allows us to apply the dispersion operator to (7.48) as follows:
\[
D\{\hat{\xi}_K\} = D\{\hat{\xi}_u - N^{-1}K^T(KN^{-1}K^T)^{-1}K\hat{\xi}_U\} + \\
+ D\{N^{-1}K^T(KN^{-1}K^T)^{-1}z_0\} \Rightarrow \\
D(\hat{\xi}_S \rightarrow \hat{\xi}_K) = \sigma_0^2N^{-1} - \sigma_0^2N^{-1}K^T(KN^{-1}K^T)^{-1}KN^{-1} + \\
+ \sigma_0^2N^{-1}K^T(KN^{-1}K^TP_0KN^{-1}K^T)^{-1}KN^{-1}.
\] (7.49)

Compare (7.49) to (5.15) to see that \(D\{\hat{\xi}_K\}\) increases by
\[
\sigma_0^2N^{-1}K^T(KN^{-1}K^TP_0KN^{-1}K^T)^{-1}KN^{-1}
\]
when the estimator \(\hat{\xi}_K\) is used for the model with stochastic constraints (7.1).
We already noted that $\hat{\xi}_K$ is a sub-optimal (reproducing) estimator within model (7.1). We now give the optimal reproducing estimator without derivation (for details see Schaffrin (1997a)).

$$\hat{\xi}_{opt-rep} = \hat{\xi}_S + K^T (KK^T)^{-1} (z_0 - K\hat{\xi}_S)$$  \hspace{1cm} (7.50)

The symbol $\hat{\xi}_S$ on the right side of (7.50) represents the optimal (“non-reproducing”) estimator. Equation (7.50) is identical to (7.47) when $N^{-1}$ is replaced by $I$.

The dispersion matrix is given by

$$D\{\hat{\xi}_{opt-rep}\} = D\{\hat{\xi}_S\} + D\{K^T (KK^T)^{-1} (z_0 - K\hat{\xi}_S)\} =$$

$$= \sigma_0^2 N^{-1} - \sigma_0^2 N^{-1} K^T (P_0^{-1} + KN^{-1}K)^{-1} KN^{-1} +$$

$$+ \sigma_0^2 K^T (KK^T)^{-1} P_0^{-1} (P_0^{-1} + KN^{-1}K)^{-1} P_0^{-1} (KK^T)^{-1} K.$$  \hspace{1cm} (7.51)

Also note that

$$E\{\hat{\xi}_{opt-rep}\} = \xi,$$  \hspace{1cm} (7.52a)

$$z_0 - K\hat{\xi}_{opt-rep} = 0,$$  \hspace{1cm} (7.52b)

$$D\{K\hat{\xi}_{opt-rep}\} = D\{z_0\} = \sigma_0^2 P_0^{-1}.$$  \hspace{1cm} (7.52c)

7.4 Summary Formulas for the Least-Squares Solution Within the Gauss-Markov Model with Stochastic Constraints

The Gauss-Markov Model with stochastic constraints is given by

$$\begin{bmatrix} y \\ z_0 \end{bmatrix} = \begin{bmatrix} A & 0 \end{bmatrix} \begin{bmatrix} \xi \\ e_0 \end{bmatrix} + \begin{bmatrix} e \\ e_0 \end{bmatrix},$$

$$\begin{bmatrix} e \\ e_0 \end{bmatrix} \sim \left( \begin{array}{cc} 0 & \sigma_0^2 \\ 0 & 0 \end{array} \right) \begin{bmatrix} P^{-1} & 0 \\ 0 & P_0^{-1} \end{bmatrix}.)$$
Table 7.1: Summary formulas for the LESS within the Gauss-Markov Model with stochastic constraints

<table>
<thead>
<tr>
<th>Quantity</th>
<th>Formula</th>
<th>Eq. No.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model redundancy</td>
<td>( r = n - m + l )</td>
<td>(7.33)</td>
</tr>
<tr>
<td>Vector of estimated parameters</td>
<td>( \hat{\xi} = \left( N + K^T P_0 K \right)^{-1} (e + K^T P_0 z_0) )</td>
<td>(7.6a)</td>
</tr>
<tr>
<td>Dispersion matrix for estimated parameters</td>
<td>( D(\hat{\xi}) = \sigma_0^2 \left( N + K^T P_0 K \right)^{-1} )</td>
<td>(7.8)</td>
</tr>
<tr>
<td>Vector of predicted residuals</td>
<td>( \hat{e} = y - A\hat{\xi} )</td>
<td>(7.9)</td>
</tr>
<tr>
<td>Dispersion matrix for residuals</td>
<td>( D(\hat{e}) = \sigma_0^2 \left[ P^{-1} - A \left( N + K^T P_0 K \right)^{-1} A^T \right] )</td>
<td>(7.12b)</td>
</tr>
<tr>
<td>Vector of residuals of prior information</td>
<td>( \hat{e}_0 = z_0 - K\hat{\xi} )</td>
<td>(7.10a)</td>
</tr>
<tr>
<td>Dispersion matrix for residuals of prior information</td>
<td>( D(\hat{e}_0) = \sigma_0^2 P_0^{-1} \left( I_l + P_0 K N^{-1} K^T \right)^{-1} )</td>
<td>(7.13)</td>
</tr>
<tr>
<td>Sum of squared residuals (SSR)</td>
<td>( \Omega = \hat{\Sigma}^T P \hat{\Sigma} + \hat{e}_0^T P_0 \hat{e}_0 )</td>
<td>(7.32)</td>
</tr>
<tr>
<td>Estimated variance component</td>
<td>( \hat{\sigma}_0^2 = (\hat{\Sigma}^T P \hat{\Sigma} + \hat{e}_0^T P_0 \hat{e}_0) / (n - m + l) )</td>
<td>(7.31)</td>
</tr>
<tr>
<td>Vector of adjusted observations</td>
<td>( \hat{\mu}_y = y - \hat{e} )</td>
<td>(7.24)</td>
</tr>
<tr>
<td>Dispersion matrix for adjusted observations</td>
<td>( D(\hat{\mu}_y) = \sigma_0^2 A \left( N + K^T P_0 K \right)^{-1} A^T )</td>
<td>(7.26)</td>
</tr>
<tr>
<td>Vector of adjusted constraints</td>
<td>( \hat{\mu}_{z_0} = z_0 - \hat{e}_0 )</td>
<td>(7.25)</td>
</tr>
<tr>
<td>Dispersion matrix for adjusted constraints</td>
<td>( D(\hat{\mu}_{z_0}) = \sigma_0^2 K \left( N + K^T P_0 K \right)^{-1} K^T )</td>
<td>(7.27)</td>
</tr>
</tbody>
</table>
Sequential Adjustments

The data model for sequential adjustments is based on two data sets, denoted by subscripts 1 and 2, respectively. The first data set is comprised of $n_1$ observations, and the second is comprised of $n_2$. It is assumed that the observations from the first data set, $y_1$, are uncorrelated with those from the second, $y_2$, i.e., $C\{y_1, y_2\} = 0$. Moreover, all parameters associated with the second data set are also associated with the first data set. Thus, the data model is written as

\[
y_1 \in \mathbb{R}^{n_1 \times 1} = A_1 \xi + e_1, \quad \text{(8.1a)}
\]
\[
y_2 \in \mathbb{R}^{n_2 \times 1} = A_2 \xi + e_2, \quad \text{(8.1b)}
\]
\[
\begin{bmatrix}
e_1 \\
e_2
\end{bmatrix} \sim (0, \sigma^2_0 \begin{bmatrix}
P_1^{-1} & 0 \\
0 & P_2^{-1}
\end{bmatrix}). \quad \text{(8.1c)}
\]

The ranks of the coefficient (design) matrices $A_1$ and $A_2$ are such that

\[
\text{rk } A_1 = \text{rk } \begin{bmatrix} A_1 \\ A_2 \end{bmatrix} = m. \quad \text{(8.2)}
\]

Note that the coefficient matrix $A_1$ has full column rank, that there is no correlation between the random error vectors $e_1$ and $e_2$, and that both data sets share a common variance component $\sigma^2_0$. Also, the total number of observations from both data sets is defined as $n := n_1 + n_2$.

The following notation is adopted for variables used in the subsequent least-squares normal equations:

\[
\begin{bmatrix}
N_{ii}, c_i
\end{bmatrix} = A_i^T P_i \begin{bmatrix} A_i \\ y_i \end{bmatrix}, \quad i \in \{1, 2\}, \quad \text{(8.3a)}
\]
\[
N_{ij} = A_i^T P_i A_j + A_j^T P_j A_i, \quad i, j \in \{1, 2\}, \quad \text{for } i \neq j. \quad \text{(8.3b)}
\]
We use a single hat to denote estimates that are based only on the first data set and a double hat to denote estimates that are based on both data sets. For example, the estimate \( \hat{\xi} \) is based only on the first data set, whereas the estimate \( \hat{\hat{\xi}} \) is based on both data sets. This makes it convenient to show estimates based on both data sets as an update to estimates based on only the first data set.

We recognize a structural similarity between the data model shown in (8.1) and the Gauss-Markov Model with stochastic constraints shown in (7.1). Given this similarity, we may immediately write down a least-squares solution for \( \xi \), and its dispersion matrix, in the form of (7.6b) and (7.8), respectively, viewing the second data set as analogous to stochastic constraints.

\[
\hat{\hat{\xi}} = \hat{\xi} + N_{11}^{-1}A^T_2(P_2^{-1} + A_2N_{11}^{-1}A^T_2)^{-1}(y_2 - A_2\hat{\xi}) = (8.4a)
\]

\[
\hat{\xi} = \hat{\xi} + (N_{11} + A^T_2P_2A_2)^{-1}A^T_2P_2(y_2 - A_2\hat{\xi}) = (8.4b)
\]

\[
D\{\hat{\xi}\} = D\{\hat{\xi}\} - \sigma_0^2N_{11}^{-1}A^T_2(P_2^{-1} + A_2N_{11}^{-1}A^T_2)^{-1}A_2N_{11}^{-1} = (8.4c)
\]

Equation (A.9a) was used in going from (8.4a) to (8.4b). It is important to note that the matrix \( (P_2^{-1} + A_2N_{11}^{-1}A^T_2) \) is of size \( n_2 \times n_2 \); whereas the size of matrix \( (N_{11} + A^T_2P_2A_2) \) is \( m \times m \). Therefore, if the second data set has only one observation, then \( n_2 = 1 \), and the update via (8.4a) is very fast! This may be the case, for example, in a real-time application where one new observation is added at each epoch in time.

It is also noted that the matrix subtracted in (8.4c) is positive-definite, which means that regardless of the precision of the second data set reflected in \( P_2 \), the dispersion of the parameters estimated from both data sets will be smaller than that estimated from only the first data set.

### 8.1 Verification of the Sequential Adjustment

In this section we discuss verification of the sequential adjustment, the aim of which is to confirm that the adjustment based on both data sets is consistent with an adjustment based only on the first data set. By consistent we mean that both the first data set only and the combined data sets fit the model well, implying that the residuals from an adjustment of the first data set would not change much in a sequential adjustment of both data sets.

We can make use of the work done in Chapter 7 to write the estimated variance component \( \hat{\sigma}_0^2 \) in a form composed of the sum of squared residuals (SSR) \( \Omega \) based on an adjustment of the first data set only and an update \( R(P_2) \) for the contribution to the SSR from the second data set, analogous to the derivation of (7.37). This facilitates hypothesis testing for the purpose of determining if the combined adjustment is consistent with an adjustment based only on the first data set. The decomposition of \( \hat{\sigma}_0^2 \) into \( \Omega \) and \( R(P_0) \) is expressed as follows:

\[
\hat{\sigma}_0^2(n - m) = \Omega + R(P_2) \text{ with } \Omega = \sigma_0^2(n_1 - m) \quad (8.5a)
\]
\begin{align*}
R(P_2) &= -(y_2 - A_2\hat{\xi})^T \hat{\lambda} \quad \text{with} \quad \hat{\lambda} = -(P_2^{-1} + A_2N_{11}^{-1}A_2^T)^{-1}(y_2 - A_2\hat{\xi}) \Rightarrow (8.5b) \\
\hat{\sigma}_0^2(n-m) &= \Omega + (y_2 - A_2\hat{\xi})^T (P_2^{-1} + A_2N_{11}^{-1}A_2^T)^{-1}(y_2 - A_2\hat{\xi}) \Rightarrow (8.5c)
\end{align*}

Then, the test statistic
\[
t = \frac{R/n_2}{\Omega/(n_1-m)} \sim F(n_2, n_1-m) \quad (8.6)
\]
can be computed to verify the sequential adjustment, i.e., that both the first data set and the combined first and second data sets fit the model well. The test statistic has an \(F\)-distribution with \(n_2\) and \(n_1-m\) degrees of freedom. For some specified significance level \(\alpha\), we may claim that the observations from the second data set are consistent with those from the first if \(t \leq F_{\alpha,n_2,n_1-m}\). See Chapter 10 for more on hypothesis testing.

### 8.2 Alternative Solution for the Normal Equations

Using the addition theory of normal equations, we may find a matrix representation of the normal equations as follows, where again the double hats refer to a solution based on both data sets:
\[
(N_{11} + N_{22})\hat{\xi} = (c_1 + c_2) \quad (8.7)
\]

These normal equations lead to
\[
N_{11}\hat{\xi} + N_{22}\hat{\xi} - c_2 = c_1 \Rightarrow \quad (8.8a)
\]
\[
N_{11}\hat{\xi} + A_2^T \hat{\lambda}_2 = c_1, \quad \text{with} \quad \hat{\lambda} = P_2(A_2\hat{\xi} - y_2) \Rightarrow \quad (8.8b)
\]
\[
y_2 = A_2\hat{\xi} - P_2^{-1}\hat{\lambda}. \quad (8.8c)
\]

Then, from (8.8b) and (8.8c), we can write the following system of least-squares normal equations:
\[
\begin{bmatrix}
N_{11} & A_2^T \\
A_2 & -P_2^{-1}
\end{bmatrix}
\begin{bmatrix}
\hat{\xi} \\
\hat{\lambda}
\end{bmatrix}
= \begin{bmatrix}
c_1 \\
y_2
\end{bmatrix} \quad (8.9)
\]

From the first row of (8.9) we get
\[
\hat{\xi} = N_{11}^{-1}c_1 - N_{11}^{-1}A_2^T\hat{\lambda} = \quad (8.10a)
\]
\[
= \hat{\xi} - N_{11}^{-1}A_2^T\hat{\lambda}. \quad (8.10b)
\]

Equation (8.10b) is an update formula as a function of the vector of estimated Lagrange multipliers \(\hat{\lambda}\). Without further derivation, we can compare (8.10b) to (8.4a) to get an expression for the estimated vector of Lagrange-multiplier as
\[
\hat{\lambda} = -(P_2^{-1} + A_2N_{11}^{-1}A_2^T)^{-1}(y_2 - A_2\hat{\xi}). \quad (8.11)
\]
Applying covariance propagation to (8.10b), we find the dispersion matrix of \( \hat{\xi} \) to be
\[
D(\hat{\xi}) = D(\hat{\xi}) - \sigma_0^2 N_{11}^{-1} A_2^T (P_2^{-1} + A_2 N_{11}^{-1} A_2^T)^{-1} A_2 N_{11}^{-1},
\]
where we used the fact that \( C\{y_2, \hat{\xi} \} = 0 \), which indicates that the observations from the second data set are uncorrelated with the estimated parameters based on the first data set only.

### 8.3 Sequential Adjustment, Rank-Deficient Case

Suppose matrix \( A_1 \) does not have full column-rank, i.e. \( \text{rk} A_1 =: q_1 < m \). Then we may introduce a datum by further splitting the system of equations as was done in Section 6.1. Let us split \( A_1 \) into an \( n_1 \times q_1 \) part denoted \( A_{11} \) and an \( n_1 \times (m - q_1) \) part denoted \( A_{12} \). Accordingly, we also split the parameter vector \( \xi \) into a \( q_1 \times 1 \) part \( \xi_1 \) and a \( (m - q_1) \times 1 \) part \( \xi_2 \). Thus, we have
\[
A_1 = [A_{11}, A_{12}], \quad \text{rk} A_{11} = q_1, \quad \text{and} \quad \xi = [\xi_1^T, \xi_2^T]^T.
\]

Next we introduce datum information \( \xi_2^0 \), such that \( \xi_2 \rightarrow \xi_2^0 \), where the subscript 2 now obviously refers to the datum, rather than a second data set. The formulas for the estimated parameters and their dispersion matrix based on the first data set only can be copied from (6.4b) and (6.5), respectively.

\[
\hat{\xi}_1 = N_{11}^{-1} (c_1 - N_{12} \xi_2^0)
\]

\[
D(\hat{\xi}_1) = \sigma_0^2 N_{11}^{-1}
\]

The estimated variance component \( \hat{\sigma}_0^2 \) is slightly different from that of (6.10) and (6.13) and is given by the formula
\[
\hat{\sigma}_0^2 = \frac{y_1^T P (y_1 - A_{11} \hat{\xi}_1 - A_{12} \xi_2^0)}{(n_1 - q_1)}
\]

or, equivalently,
\[
\hat{\sigma}_0^2 = \frac{(y_1^T P y_1 - c_1^T N_{11}^{-1} c_1)}{(n_1 - q_1)}.
\]

Note that the steps taken from (6.10) to (6.13) can be used to go from (8.14c) to (8.14d).

Now we introduce the second data set with a splitting analogous to the first, viz.
\[
y_2 = A_{21} \xi_1 + A_{22} \xi_2 + e_2, \quad e_2 \sim (0, \sigma_0^2 P_2^{-1}).
\]

The matrix \( A_{21} \) is of size \( n_2 \times q_1 \), and \( A_{22} \) is of size \( n_2 \times (m - q_1) \). No information in the second data set refers to the datum choice; it only adds to the redundancy.
provided by the first data set. The rank of the normal equation matrix is unchanged
and is expressed as
\[
\operatorname{rk} \begin{bmatrix}
A_{11} & A_{12} \\
A_{21} & A_{22}
\end{bmatrix} =: q = q_1
\]
(8.16)
The full least-squares normal equations are then written as
\[
\begin{bmatrix}
A_{11}^T P_1 A_{11} + A_{12}^T P_2 A_{21} & A_{11}^T P_1 A_{12} + A_{12}^T P_2 A_{22} \\
A_{21}^T P_1 A_{11} + A_{22}^T P_2 A_{21} & A_{21}^T P_1 A_{12} + A_{22}^T P_2 A_{22}
\end{bmatrix}
\begin{bmatrix}
\hat{\xi}_1 \\
\xi_0
\end{bmatrix}
= 
\begin{bmatrix}
A_{11}^T P_1 y_1 + A_{21}^T P_2 y_2 \\
A_{11}^T P_1 y_1 + A_{22}^T P_2 y_2
\end{bmatrix}.
\]
(8.17)
From the first row of (8.17), we may write the least-squares solution for \(\hat{\xi}_1\) directly,
followed by its dispersion matrix, as
\[
\hat{\xi}_1 = (A_{11}^T P_1 A_{11} + A_{12}^T P_2 A_{21})^{-1} \cdot \left[ (A_{11}^T P_1 y_1 + A_{21}^T P_2 y_2) - (A_{11}^T P_1 A_{12} + A_{12}^T P_2 A_{22}) \xi_0 \right],
\]
(8.18)
\[
D(\hat{\xi}_1) = \sigma^2_0 (A_{11}^T P_1 A_{11} + A_{21}^T P_2 A_{21})^{-1}.
\]
(8.19)
In order to derive update formulas, it is helpful to introduce an alternative
expression for the normal equations analogous to what was done in (8.8a) through
(8.9).
\[
(A_{11}^T P_1 A_{11}) \hat{\xi}_1 = (A_{11}^T P_1 y_1) - (A_{11}^T P_1 A_{12}) \xi_0 \Rightarrow
\]
(8.20a)
\[
N_{11} \hat{\xi}_1 = c_1 - N_{12} \xi_0
\]
(8.20b)
\[
(A_{21}^T P_2 A_{21}) \hat{\xi}_1 = (A_{21}^T P_2 y_2) - (A_{21}^T P_2 A_{22}) \xi_0 \Rightarrow
\]
(8.20c)
\[
N_{21} \hat{\xi}_1 = c_2 - N_{22} \xi_0
\]
(8.20d)
Here we have used the symbols \(N_{12}\) and \(N_{21}\) differently than defined in (8.3). Together, (8.20b) and (8.20d) comprise the first row of (8.17). Recombining (8.20b)
and (8.20d) gives
\[
(N_{11} + N_{21}) \hat{\xi}_1 = c_1 + c_2 - (N_{12} + N_{22}) \xi_0 \Rightarrow
\]
(8.21a)
\[
N_{11} \hat{\xi}_1 + A_{21}^T \hat{\lambda} = c_1 - N_{12} \xi_0, \quad \text{with} \quad \hat{\lambda} = P_2 (A_{21} \hat{\xi}_1 - y_2 + A_{22} \xi_0).
\]
(8.21b)
Note that in (8.20a)–(8.20d) a single hat was used for the estimate of \(\xi_1\) since
each respective equation represents only one set of data. The double hat in (8.21a)
denotes the estimate of $\hat{\xi}_1$ based on both data sets. From (8.21b) we can write the system of normal equations in matrix form as follows:

$$
\begin{bmatrix}
N_{11} & A_{21}^T \\
A_{21} & -P_2^{-1}
\end{bmatrix}
\begin{bmatrix}
\hat{\xi}_1 \\
\hat{\lambda}
\end{bmatrix}
=
\begin{bmatrix}
c_1 - N_{12}\xi_2^0 \\
y_2 - A_{22}\xi_2^0
\end{bmatrix}.
$$

(8.22)

The solution of (8.22) can be obtained by applying the inversion formula for a partitioned matrix as shown in (A.14).

$$
\begin{bmatrix}
\hat{\xi}_1 \\
\hat{\lambda}
\end{bmatrix}
=
\begin{bmatrix}
N_{11} & A_{21}^T \\
A_{21} & -P_2^{-1}
\end{bmatrix}^{-1}
\begin{bmatrix}
c_1 - N_{12}\xi_2^0 \\
y_2 - A_{22}\xi_2^0
\end{bmatrix}
= 
\begin{bmatrix}
N_{11}^{-1} - N_{11}^{-1}A_{21}^T S_2 A_{21} N_{11}^{-1} & N_{11}^{-1}A_{21}^T S_2 \\
S_2 A_{21} N_{11}^{-1} & -S_2
\end{bmatrix}
\begin{bmatrix}
c_1 - N_{12}\xi_2^0 \\
y_2 - A_{22}\xi_2^0
\end{bmatrix},
$$

(8.23)

with

$$
S_2 := (P_2^{-1} + A_{21} N_{11}^{-1} A_{21}^T)^{-1}.
$$

(8.24)

Finally, the estimated parameters and Lagrange multipliers are expressed as

$$
\begin{align}
\hat{\xi}_1 &= N_{11}^{-1}(c_1 - N_{12}\xi_2^0) + \\
&+ N_{11}^{-1}A_{21}^T(P_2^{-1} + A_{21} N_{11}^{-1} A_{21}^T)^{-1}[A_{21} N_{11}^{-1}(-c_1 + N_{12}\xi_2^0) + y_2 - A_{22}\xi_2^0] \\
\hat{\lambda} &= -(P_2^{-1} + A_{21} N_{11}^{-1} A_{21}^T)^{-1}(y_2 - A_{21}\hat{\xi}_1 - A_{22}\xi_2^0).
\end{align}

(8.25a-8.25b)

The dispersion matrix of the estimated vector of Lagrange multipliers is

$$
D(\hat{\lambda}) = (P_2^{-1} + A_{21} N_{11}^{-1} A_{21}^T)^{-1} D(y - A_{21}\hat{\xi}_1)(P_2^{-1} + A_{21} N_{11}^{-1} A_{21}^T)^{-1},
$$

(8.26)

since $D(\xi_2^0) = 0$. The following relations also hold:

$$
\begin{align}
C(y_2, \hat{\xi}_1) &= 0, \\
D(y - A_{21}\hat{\xi}_1) &= \sigma_0^2(P_2^{-1} + A_{21} N_{11}^{-1} A_{21}^T), \\
D(\hat{\lambda}) &= \sigma_0^2(P_2^{-1} + A_{21} N_{11}^{-1} A_{21}^T)^{-1}, \\
D(\hat{\xi}_1) &= D(\hat{\xi}_1) - \sigma_0^2 N_{11}^{-1} A_{21}^T(P_2^{-1} + A_{21} N_{11}^{-1} A_{21}^T)^{-1} A_{21} N_{11}^{-1}.
\end{align}

(8.27a-8.27d)

The estimated variance component is expressed as follows:
Now, using the addition theory of normal equations, we can write
\[ \hat{\sigma}_0^2(n-q) = \hat{\sigma}_0^2(n_1-q_1) + (y_2 - A_{21}\hat{\xi}_1 - A_{22}\xi_2^0)^T \cdot (P_2^{-1} + A_{21}N_{11}^{-1}A_{21}^T)^{-1}(y_2 - A_{21}\hat{\xi}_1 - A_{22}\xi_2^0) \Rightarrow (8.28a) \]
\[ \hat{\sigma}_0^2(n-q) = \hat{\sigma}_0^2(n_1-q_1) - \hat{\lambda}^T(y_2 - A_{21}\hat{\xi}_1 - A_{22}\xi_2^0). \] (8.28b)

Once again, from (8.20b) to (8.28b), we have used the definition \( N_{11} := A_{11}^TP_1A_{11}. \)

### 8.4 Sequential Adjustment with New Parameters

In this section we consider the case where the second data set refers to all the parameters of the first data set plus some additional new parameters. Thus we speak of \( m_1 \) parameters associated with the first data set and \( m_2 \) with the second, with \( m_2 > m_1 \) (i.e., the \( m_1 \) parameters associated with the first data set are a subset of the \( m_2 \) parameters associated with the second data set). In the double subscripts used below, the first one refers to the data set, and the second subscript refers to the matrix splitting. For example, \( A_{21} \) is that part of the design matrix from the second data set that refers to the original parameters, whereas \( A_{22} \) is associated with the new parameters observed in the second data set. We could have adopted a new symbol, e.g. \( \hat{y} \), to denote a “preprocessed observation” vector that includes the datum information. However, we have elected to continue using \( y \) and simply note that it could include datum information in addition to the observations. The data model that follows implies that we have assumed there are no correlations between the observations of data-set one and those of data-set two; it also implies that both sets of observations share a common variance component \( \sigma_0^2 \).

\[
\begin{bmatrix}
    y_1 \\
    y_2
\end{bmatrix} = \begin{bmatrix}
    A_{11} & 0 \\
    A_{21} & A_{22}
\end{bmatrix} \begin{bmatrix}
    \xi_1 \\
    \xi_2
\end{bmatrix} + \begin{bmatrix}
    e_1 \\
    e_2
\end{bmatrix},
\end{bmatrix} \sim \begin{pmatrix}
    0 \\
    0
\end{pmatrix}, \sigma_0^2 \begin{bmatrix}
    P_1^{-1} & 0 \\
    0 & P_2^{-1}
\end{bmatrix} \Rightarrow (8.29)
\]

The size of the system of equations is implied by the following:
\[
y_1 \in \mathbb{R}^{n_1}, \quad \xi_1 \in \mathbb{R}^{m_1}, \quad y_2 \in \mathbb{R}^{n_2}, \quad \xi_2 \in \mathbb{R}^{m_2-m_1}, \quad n = n_1 + n_2, \quad m_2 > m_1. \quad (8.30)
\]

Now, using the addition theory of normal equations, we can write
\[
\begin{bmatrix}
    A_{11}^T & A_{21}^T \\
    0 & A_{22}^T
\end{bmatrix} \begin{bmatrix}
    P_1 & 0 \\
    0 & P_2
\end{bmatrix} \begin{bmatrix}
    A_{11} & 0 \\
    A_{21} & A_{22}
\end{bmatrix} \begin{bmatrix}
    \xi_1 \\
    \xi_2
\end{bmatrix} \sim \begin{pmatrix}
    A_{11}^TP_1 & A_{21}^TP_2 \\
    0 & A_{22}^TP_2
\end{pmatrix} \begin{bmatrix}
    y_1 \\
    y_2
\end{bmatrix} \Rightarrow (8.31a)
\]
\[
\begin{bmatrix}
    A_{11}^TP_1A_{11} + A_{21}^TP_2A_{21} & A_{21}^TP_2A_{22} \\
    A_{22}^TP_2A_{21} & A_{22}^TP_2A_{22}
\end{bmatrix} \begin{bmatrix}
    \xi_1 \\
    \xi_2
\end{bmatrix} = \begin{pmatrix}
    A_{11}^TP_1y_1 + A_{21}^TP_2y_2 \\
    A_{22}^TP_2y_2
\end{pmatrix}. \quad (8.31b)
\]

Here again, the double-hats refer to estimates based on both data sets.

79
Now, the first data set may no longer be available, rather we may have only the estimates from the first adjustment. In this case we can use the bottom row of (8.31b) to solve for the estimates of the new parameters in terms of only the second data set, leading to

\[
\hat{\xi}_2 = (A_{22}^T P_2 A_{22})^{-1} A_{22}^T P_2 (y_2 - A_{21} \hat{\xi}_1).
\]  

(8.32)

Then, from the normal equations based solely on the first data set, we may substitute

\[
A_{11}^T P_1 y_1 = (A_{11}^T P_1 A_{11}) \hat{\xi}_1
\]

(8.33)

into the top row of the right side of (8.31b) and invert the normal-equation matrix on the left to solve for the parameter estimates. For convenience, we introduce the following symbols to use in the inverted matrix:

\[
S_1 := A_{11}^T P_1 A_{11} + A_{21}^T P_2 A_{21} - A_{21}^T P_2 A_{22} (A_{22}^T P_2 A_{22})^{-1} A_{22}^T P_2 A_{21} = (8.34a)
\]

\[
= A_{11}^T P_1 A_{11} + A_{21}^T P_2 A_{21},
\]

(8.34b)

\[
P_2 := P_2 - P_2 A_{22} (A_{22}^T P_2 A_{22})^{-1} A_{22}^T P_2,
\]

(8.34c)

\[
N_{22} = A_{22}^T P_2 A_{22}.
\]

(8.34d)

We refer to \( \tilde{P}_2 \) as a reduced weight matrix. Upon inverting the normal-equations matrix (see (A.14) for the inverse of a partitioned matrix), we find the following solution for \( \hat{\xi}_1 \) and \( \hat{\xi}_2 \):

\[
\begin{bmatrix}
\hat{\xi}_1 \\
\hat{\xi}_2
\end{bmatrix} = \begin{bmatrix}
S_1^{-1} \\
-N_{22}^{-1} (A_{22}^T P_2 A_{21}) S_1^{-1}
\end{bmatrix} \begin{bmatrix}
-S_1^{-1} (A_{21}^T P_2 A_{22}) N_{22}^{-1} \\
N_{22}^{-1} + N_{22}^{-1} (A_{22}^T P_2 A_{21}) S_1^{-1} (A_{21}^T P_2 A_{22}) N_{22}^{-1}
\end{bmatrix} \\
\left((A_{11}^T P_1 A_{11}) \hat{\xi}_1 + A_{21}^T P_2 y_2\right) \\
A_{22}^T P_2 y_2
\end{bmatrix}. 
\]

(8.35)

We can continue by using (8.33), (8.34b) and (8.34c) with the first row of (8.35) to arrive at

\[
\hat{\xi}_1 = S_1^{-1} \left\{ (A_{11}^T P_1 A_{11}) \hat{\xi}_1 + A_{21}^T P_2 y_2 - (A_{21}^T P_2 A_{22}) N_{22}^{-1} A_{22}^T P_2 y_2 \right\} = \]  

(8.36a)

\[
= S_1^{-1} \left\{ \left[ (A_{11}^T P_1 A_{11}) \hat{\xi}_1 + A_{21}^T P_2 y_2 \right] + \left[ (A_{21}^T P_2 A_{21}) - (A_{21}^T P_2 A_{21}) \right] \hat{\xi}_1 \right\} = \]  

(8.36b)

\[
= S_1^{-1} A_{21}^T P_2 (y_2 - A_{21} \hat{\xi}_1) + S_1^{-1} (A_{11}^T P_1 A_{11} + A_{21}^T P_2 A_{21}) \hat{\xi}_1 = \]  

(8.36c)

\[
= S_1^{-1} A_{21}^T P_2 (y_2 - A_{21} \hat{\xi}_1) + \hat{\xi}_1 \Rightarrow \]  

(8.36d)

\[
\hat{\xi}_1 - \hat{\xi}_1 = S_1^{-1} A_{21}^T P_2 (y_2 - A_{21} \hat{\xi}_1),
\]

(8.36e)

where (8.36e) is in the form of an update formula.
Using (8.38c), we may rewrite (8.36e) as

Moreover, we find that matrix \( \hat{\bar{P}} \). Thus we reduce the rank by modifying the original weight matrix \( P \). However, in some applications, the number of observations \( n \) in (8.36e) so that only a matrix of size \( m \times m \) set may be significantly less than \( n \). We assume that \( P \) is invertible, as implied in the given model (8.29). We now wish to check the rank of the reduced weight matrix \( \hat{P} \). It is easy to check that the product \( P^{-1} \) is idempotent. Then using (A.4) and (A.12) we find

\[
\text{rk } P = \text{tr}(P^{-1} P) = \text{tr}(I_{n_2} - A_{22}(T^T P_2 A_{22})^{-1} T P_2) = (8.37a)
\]

\[
= n - \text{tr}[A_{22}(T^T P_2 A_{22})^{-1} T P_2] = n - \text{tr}[(A^T P_2 A_{22})^{-1} A_{22} P_2 A_{22}] = (8.37b)
\]

\[
= n_2 - m_2 < n_2.
\] (8.37c)

Thus we reduce the rank by modifying the original weight matrix \( P \) to obtain \( \hat{P} \). Moreover, we find that matrix \( \hat{P} \) is singular. The parameter dispersion matrices, \( D\{\hat{\xi}_1\} \) and \( D\{\hat{\xi}_2\} \), are shown at the end of the next section.

### 8.5 Sequential Adjustment with New Parameters and Small Second Data Set

In (8.36e) we must invert the matrix \( S_1 \) to solve the system of equations. However, in some applications, the number of observations \( n_2 \) in the second data set may be significantly less than \( m_1 \). In this case we would like to reformulate the solution in (8.36e) so that only a matrix of size \( n_2 \times n_2 \) needs to be inverted.

We have an alternative expression for matrix \( S_1 \) in (8.34b), the inverse of which can also be derived as follows:

\[
S_1^{-1} = [(A_{11}^T P_1 A_{11}) + (A_{21}^T \hat{P}_2 A_{21})]^{-1} = (8.38a)
\]

\[
= \{[I_{m_1} + (A_{21}^T \hat{P}_2 A_{21}) (A_{11}^T P_1 A_{11})^{-1}] (A_{11}^T P_1 A_{11}) \}^{-1} = (8.38b)
\]

\[
= (A_{11}^T P_1 A_{11})^{-1} [I_{m_1} + (A_{21}^T \hat{P}_2 A_{21}) (A_{11}^T P_1 A_{11})^{-1}]^{-1}.
\] (8.38c)

Using (8.38c), we may rewrite (8.36e) as

\[
\hat{\xi}_1 - \xi_1 = (A_{11}^T P_1 A_{11})^{-1} [I_{m_1} + (A_{21}^T \hat{P}_2 A_{21}) (A_{11}^T P_1 A_{11})^{-1}]^{-1} A_{21}^T \hat{P}_2 (y_2 - A_{21} \hat{\xi}_1) = (8.39a)
\]

\[
= (A_{11}^T P_1 A_{11})^{-1} A_{21}^T \hat{P}_2 [I_{m_2} + A_{21} (A_{11}^T P_1 A_{11})^{-1} A_{21}^T \hat{P}_2]^{-1} (y_2 - A_{21} \hat{\xi}_1).
\] (8.39b)

Here, we have made use of (A.9) in the step from (8.39a) to (8.39b), with two of the matrices in (A.9) set to identity. Note that the matrix to invert inside the square brackets is of size \( m_1 \times m_1 \) in (8.39a) but is size \( n_2 \times n_2 \) in (8.39b). The choice of which equation to use will usually be determined by the smaller of \( m_1 \) and \( n_2 \). Also, we have the relation

\[
- \hat{\lambda} = [I_{n_2} + A_{21} (A_{11}^T P_1 A_{11})^{-1} A_{21}^T \hat{P}_2]^{-1} (y_2 - A_{21} \hat{\xi}_1),
\] (8.40)
which means that the solution for the first subset of parameters may also be expressed as

\[
\hat{\xi}_1 - \xi_1 = -(A_{11}^T P_1 A_{11})^{-1} A_{21}^T \hat{P}_2 \hat{\lambda}.
\]  
(8.41)

Now we begin with (8.32) to find a solution for the parameters \(\hat{\xi}_2\) in terms of the Lagrange multipliers \(\hat{\lambda}\):

\[
\hat{\xi}_2 = (A_{22}^T P_2 A_{22})^{-1} A_{22}^T P_2 (y_2 - A_{21} \hat{\xi}_1) = (A_{22}^T P_2 A_{22})^{-1} A_{22}^T P_2 \left\{ (y_2 - A_{21} \hat{\xi}_1) - A_{21} (A_{11}^T P_1 A_{11})^{-1} A_{21}^T P_2 \right\} = (A_{22}^T P_2 A_{22})^{-1} A_{22}^T P_2 \left\{ I_{n_2} + A_{21} (A_{11}^T P_1 A_{11})^{-1} A_{21}^T P_2 \right\}^{-1} (y_2 - A_{21} \hat{\xi}_1) = -(A_{22}^T P_2 A_{22})^{-1} A_{22}^T P_2 \hat{\lambda}.
\]  
(8.42)

The inverse formula of (A.7) was used in the last step to reach (8.42c), with matrices \(T\), \(W\), and \(V\) in (A.7) set to identity.

To facilitate computing the parameter dispersion matrix we write the following system of normal equations, noting that (8.43b) is in the form of an update solution:

\[
\begin{bmatrix}
A_{11}^T P_1 A_{11} + A_{21}^T P_2 A_{21} \\
A_{22}^T P_2 A_{22}
\end{bmatrix}
\begin{bmatrix}
\hat{\xi}_1 \\
\hat{\xi}_2
\end{bmatrix}
= \begin{bmatrix}
(A_{11}^T P_1 A_{11}) \hat{\xi}_1 + A_{21}^T P_2 y_2 \\
A_{22}^T P_2 y_2
\end{bmatrix}
\Rightarrow
\begin{bmatrix}
A_{11}^T P_1 A_{11} + A_{21}^T P_2 A_{21} \\
A_{22}^T P_2 A_{22}
\end{bmatrix}
\begin{bmatrix}
\hat{\xi}_1 - \xi_1 \\
\hat{\xi}_2
\end{bmatrix}
= \begin{bmatrix}
A_{21}^T P_2 (y_2 - A_{21} \hat{\xi}_1) \\
A_{22}^T P_2 (y_2 - A_{21} \hat{\xi}_1)
\end{bmatrix}
\]  
(8.43)

Note that (8.43a) is equivalent to (8.31b) shown earlier.

We have already inverted the normal-equation matrix in (8.35). Taking elements from (8.35), we may write the parameter dispersion and covariance matrices as follows:

\[
D\{\hat{\xi}_1\} = \sigma_0^2 s_1^{-1} = \sigma_0^2 (A_{11}^T P_1 A_{11} + A_{21}^T \hat{P}_2 A_{21})^{-1},
\]  
(8.44a)

\[
C\{\hat{\xi}_1, \hat{\xi}_2\} = -D\{\hat{\xi}_1\} (A_{21}^T P_2 A_{22}) (A_{22}^T P_2 A_{22})^{-1},
\]  
(8.44b)

\[
D\{\hat{\xi}_2\} = \sigma_0^2 (A_{22}^T P_2 A_{22})^{-1} - (A_{22}^T P_2 A_{22})^{-1} (A_{22}^T P_2 A_{21}) C\{\hat{\xi}_1, \hat{\xi}_2\}.
\]  
(8.44c)

Each of the above covariance matrices (8.44a) through (8.44c) include the matrix \(s_1^{-1}\), which implies that a matrix of size \(n_1 \times m_1\) must be inverted. However,
with the insertion of $I_{n_2}$ into (8.44a), and with appropriate matrix groupings, we may apply the inversion formula (A.7) to find an inverse of smaller dimension as shown in the following:

$$D\{\hat{\xi}_1\} = \sigma_0^2 \left[ (A_{11}^T P_1 A_{11}) + (A_{21}^T P_2) I_{n_2} A_{21} \right]^{-1} =$$

$$= \sigma_0^2 N_{11}^{-1} - \sigma_0^2 N_{11}^{-1} A_{21}^T P_2 (I_{n_2} + A_{21} N_{11}^{-1} A_{21}^T P_2)^{-1} A_{21} N_{11}^{-1} =$$

$$= \sigma_0^2 N_{11}^{-1} - \sigma_0^2 N_{11}^{-1} A_{21}^T \bar{P}_2 \hat{\lambda}. \quad (8.45c)$$

Here, we have used $N_{11} := A_{11}^T P_1 A_{11}$ for compactness. The parenthetical term that must be inverted in equation (8.45b) is an $n_2 \times n_2$ matrix, which, again, may be much smaller than an $m_1 \times m_1$ matrix, depending on the application. Of course, the matrix $(A_{11}^T P_1 A_{11})^{-1}$ is also size $m_1 \times m_1$, but it is assumed that this inverse had already been performed in the adjustment of the first data set and was saved for subsequent use.

The estimated variance component is expressed as

$$\hat{\sigma}_0^2 (n - m_2) = \hat{\sigma}_0^2 (n_1 - m_1) - (y_2 - A_{21} \hat{\xi}_1)^T \bar{P}_2 \hat{\lambda}. \quad (8.46a)$$

Then, substituting (8.40) leads to

$$\hat{\sigma}_0^2 (n - m_2) = \hat{\sigma}_0^2 (n_1 - m_1) +$$

$$+ (y_2 - A_{21} \hat{\xi}_1)^T \bar{P}_2 [I_{n_2} + A_{21} (A_{11}^T P_1 A_{11})^{-1} A_{21}^T P_2]^{-1} (y_2 - A_{21} \hat{\xi}_1) \Rightarrow \quad (8.46b)$$

$$\hat{\sigma}_0^2 (n - m_2) = \hat{\sigma}_0^2 (n_1 - m_1) + (y_2 - A_{21} \hat{\xi}_1)^T \bar{P}_2 (y_2 - A_{21} \hat{\xi}_1). \quad (8.46c)$$
Data models introduced prior to this chapter have either admitted observation equations with parameters or condition equations without parameters, but not both. In contrast, the Gauss-Helmert Model (GHM) allows conditions and parameters to be combined in the same equations. Thus, the GHM can be viewed as being more flexible (or more general) than either the Gauss-Markov Model (GMM) (Chapter 3) or the Model of Condition Equations (Chapter 4), since it combines aspects of both. In some cases, the GHM might be useful for dealing with complicated observation equations, for example when multiple observations are related to one or more parameters via specified (possibly nonlinear) equations.

We begin our discussion of the GHM with a leveling-network example in order to contrast the GMM with the Model of Condition Equations and to show how the GHM combines the information used in those two models. The diagram in Figure 9.1 shows a leveling network with four points \( (P_1, P_2, P_3, P_4) \) that has been observed in two closed loops comprised of a total of five observations \( (y_1, y_2, y_3, y_4, y_5) \). First we present a (rank deficient) partitioned GMM as

\[
\begin{align*}
\mathbf{y} &= \mathbf{A}_1 \mathbf{\xi}_1 + \mathbf{A}_2 \mathbf{\xi}_2 + \mathbf{e}, \\
\mathbf{e} &\sim (\mathbf{0}, \sigma_0^2 \mathbf{P}^{-1}), \\
\text{rk} \mathbf{A}_1 &= \text{rk} \left[ \mathbf{A}_1 \mid \mathbf{A}_2 \right] =: q < m,
\end{align*}
\]

where the coefficient matrix \( \mathbf{A} \) and the vector of unknown parameters \( \mathbf{\xi} \) have been partitioned, respectively, as

\[
\mathbf{A} = \begin{bmatrix} \mathbf{A}_1 & \mathbf{A}_2 \\ n \times q & n \times (m-q) \end{bmatrix} \quad \text{and} \quad \mathbf{\xi} = \begin{bmatrix} \mathbf{\xi}_1^T \\ 1 \times q \\ \mathbf{\xi}_2^T \\ 1 \times (m-q) \end{bmatrix}^T.
\]

85
In this example, the number of unknown parameters is $m = 4$ (heights of four points). Since leveled height-differences supply no information about the height datum, we can only estimate the heights of three of the points with respect to the remaining fourth one. That explains why $\text{rk} \, A = q = 3 < m$, implying a datum deficiency of $m - q = 1$. Thus, the model has been partitioned so that $\xi_1$ contains three estimable heights, and $\xi_2$ is a single non-estimable height, which could be assigned a datum value. In this example, we arbitrarily chose point $P_4$ for the non-estimable height. As was stated in Chapter 6, we have the relationship $A_2 = A_1 L$ for some $q \times (m-q)$ matrix $L$, which means that matrix $A_2$ is a linear combination of the columns of matrix $A_1$, reflecting the rank deficiency of matrix $A = [A_1 \, | \, A_2]$.

The problem could also be solved within the Model of Condition Equations introduced in Chapter 4, which reads

$$By = Be, \ e \sim (0, \sigma^2_0 P^{-1}) \quad (9.3a)$$

with the orthogonality condition

$$i. \ B \left[ \begin{array}{l|l} A_1 & A_2 \end{array} \right] = 0, \quad (9.3b)$$

and the rank condition

$$ii. \ \text{rk} \, B = r = n - \text{rk} \, A_1. \quad (9.3c)$$

These two conditions ensure equivalent least-squares solutions within the models of (9.1) and (9.3a) as discussed in Section 4.2.

We have the following design (coefficient) matrices and parameter vectors for the example leveling network, for which it is easy to verify that both conditions i
and ii are satisfied:

\[
A_1 = \begin{bmatrix}
-1 & 1 & 0 \\
-1 & 0 & 1 \\
0 & -1 & 1 \\
0 & -1 & 0 \\
0 & 0 & -1
\end{bmatrix}, \quad A_2 = \begin{bmatrix}
0 \\
0 \\
1 \\
1
\end{bmatrix}, \quad B = \begin{bmatrix}
1 & -1 & 1 & 0 & 0 \\
0 & 0 & 1 & -1 & 1
\end{bmatrix},
\]

(9.4)

\[
\xi_1 = \begin{bmatrix}
h_1 \\
h_2 \\
h_3
\end{bmatrix}, \quad \xi_2 = [h_4].
\]

Here, \(h_i\) represents the height of point \(P_i\).

Now we wish to introduce a new coefficient matrix \(B\) that does not contain matrix \(A\) in its nullspace, so that we are left with a Model of Condition Equations with parameters. For now we use the symbol \(\bar{B}\) in order to distinguish it from the coefficient matrix used in the Model of Condition Equations, which does contain matrix \(A\) in its nullspace (i.e., \(BA = 0\), but \(\bar{BA} \neq 0\)). Similarly, we introduce other bar-terms to form the following GHM:

\[
\bar{y} = \bar{B}y = \bar{w} = \bar{B}A_1 \xi_1 + \bar{B}A_2 \xi_2 + \bar{Be},
\]

(9.5a)

\[
\bar{Be} \sim (0, \sigma^2 BP^{-1} \bar{B}^T),
\]

(9.5b)

\[
\text{rk}(\bar{B}) =: \bar{r}.
\]

(9.5c)

The size of \(\bar{B}\) is \(\bar{r} \times n\), implying that \(\bar{B}\) has full row-rank. The GHM in (9.5) is equivalent to the GMM in (9.1) if, and only if,

iii. \(\text{rk}(\bar{BA}_1) + r = \bar{r} \iff n = \bar{r} + q - \text{rk}(\bar{BA}_1) = \text{rk} \bar{B} + \text{rk} A - \text{rk}(\bar{BA}_1)\)

iv. \(\text{rk}(\bar{BA}_1) + r = \bar{r} \iff n = \bar{r} + q - \text{rk}(\bar{BA}_1) = \text{rk} \bar{B} + \text{rk} A - \text{rk}(\bar{BA}_1)\)

Note that, through the matrix \(\bar{B}\), one observation is eliminated for each eliminated parameter. Referring to the level network example, we may wish to eliminate the height of point \(P_3\) from the parameter list (perhaps it is a temporary benchmark of no particular interest). This can be done by introducing the following example matrix \(\bar{B}\):

\[
\bar{B} = \begin{bmatrix}
1 & 0 & 0 & 0 & 0 \\
0 & 1 & -1 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 \\
0 & 0 & 1 & 0 & 1
\end{bmatrix}, \quad \bar{BA}_2 = \begin{bmatrix}
0 \\
0 \\
1 \\
1
\end{bmatrix}, \quad \bar{BA}_1 = \begin{bmatrix}
-1 & 1 & 0 \\
-1 & 1 & 0 \\
0 & -1 & 0 \\
0 & -1 & 0
\end{bmatrix}.
\]

87
With these example matrices we have \( n = 5, \bar{r} = \text{rk} \bar{B} = 4, q = \text{rk} A_1 = 3, \) and \( \text{rk}(\bar{B} A_1) = 2. \) Since \( n - \bar{r} = 1, \) the single column of zeros in \( \bar{B} A_1 \) satisfies condition iii. Also, condition iv is satisfied since \( 5 = 4 + 3 - 2. \)

As an aside, we note that it is also possible to remove \( l \) estimable parameters via the splitting of the constraint equation introduced in (5.1).

\[
\kappa_0 = K \begin{bmatrix} \xi_1 \\ \xi_2 \end{bmatrix} \Rightarrow \begin{bmatrix} \xi_1 \\ \xi_2 \end{bmatrix} = K^{-1} K_1 \kappa_0 - K_1^{-1} K_2 \xi_2 \tag{9.6a}
\]

Here, \( K_1 \) is a \( l \times l \) invertible matrix, and \( K_2 \) is of size \( l \times (m-l). \) Upon substitution for \( \xi_1 \) of (9.6b) into (9.1), we find the following modified system of observation equations with \( l \) parameters eliminated:

\[
y = A_1 \xi_1 + A_2 \xi_2 + e = A_1 K_1^{-1} \kappa_0 + (A_2 - A_1 K_1^{-1} K_2) \xi_2 + e. \tag{9.7}
\]

The \( l \times 1 \) vector \( \xi_1 \) has vanished on the right side of (9.7). While this technique is possible, it might not be used frequently in practice.

We could derive the solution for \( \xi \) within the GHM (9.5) from statistical principles via BLUUE (Best Linear Uniformly Unbiased Estimate), but here we use the equivalent principle of LESS (LEast-Squares Solution) shown in Section 4.2. In the following, we recombine coefficient matrices \( A_1 \) and \( A_2 \) back into the single matrix \( A \) and recombine the partitioned parameter vector back into a single vector \( \xi = \begin{bmatrix} \xi_1^T \\ \xi_2^T \end{bmatrix}^T. \) Therefore, we can rewrite (9.5) as

\[
\bar{w} = \bar{B} A_1 \xi_1 + \bar{B} A_2 \xi_2 + \bar{B} e = \bar{A} \xi + \bar{B} e, \tag{9.8}
\]

where another bar-symbol was introduced for convenience, viz. \( \bar{A} := \bar{B} A. \)

Our target function should minimize a quadratic form in the random error vector \( e. \) That is, we minimize \( e^T P e \) rather than \( (\bar{B} e)^T (\bar{B} P^{-1} \bar{B}^T)^{-1} (\bar{B} e). \) Then, the Lagrange target function is written as

\[
\Phi(e, \xi, \lambda) =: e^T P e + 2\lambda^T (\bar{B} e + \bar{A} \xi - \bar{w}), \tag{9.9}
\]

which must be made stationary with respect to the unknown vectors \( e, \xi, \) and \( \lambda. \) This is done by imposing the Euler-Lagrange necessary conditions, which results in the following system of equations:

\[
\frac{1}{2} \frac{\partial \Phi}{\partial e} = P \bar{e} + \bar{B}^T \hat{\lambda} \doteq 0, \tag{9.10a}
\]

\[
\frac{1}{2} \frac{\partial \Phi}{\partial \xi} = \bar{A}^T \hat{\lambda} \doteq 0, \tag{9.10b}
\]

\[
\frac{1}{2} \frac{\partial \Phi}{\partial \lambda} = \bar{B} \bar{e} + \bar{A} \hat{\xi} - \bar{w} \doteq 0. \tag{9.10c}
\]
The vectors of predicted random errors (residuals) and estimated parameters are then solved for as follows:

\[ \hat{e} = -(P^{-1}B^T)\hat{\lambda} \Rightarrow \]
\[ -(B^TP^{-1}B^T)\hat{\lambda} = \bar{w} - \bar{A}\bar{\xi} \Rightarrow \]
\[ \hat{\lambda} = (B^P^{-1}B^T)^{-1}(\bar{w} - \bar{A}\bar{\xi}) \Rightarrow \]
\[ (B^P^{-1}B^T) \text{ is invertible} \]
\[ -\bar{A}^T\hat{\lambda} = \bar{A}^T(B^P^{-1}B^T)^{-1}(\bar{w} - \bar{A}\bar{\xi}) = 0 \Rightarrow \text{mult. by } \bar{A}^T \text{ and using (9.10b)} \]
\[ \bar{A}^T(B^P^{-1}B^T)^{-1}\bar{A}\bar{\xi} = \bar{A}^T(B^P^{-1}B^T)^{-1}\bar{w} \]

Finally, we arrive at

\[ \hat{\xi} = [\bar{A}^T(B^P^{-1}B^T)^{-1}\bar{A}]^{-1}\bar{A}^T(B^P^{-1}B^T)^{-1}\bar{w} \] (9.11a)

and

\[ \hat{e} = (P^{-1}B^T)(B^P^{-1}B^T)^{-1}(\bar{w} - \bar{A}\bar{\xi}) \] (9.11b)

for the estimated parameters and predicted residuals, respectively. Equation (9.11a) has the same form as the normal equations derived within the GMM. The dispersion matrix for the estimated parameter vector \( \hat{\xi} \) is expressed by

\[ D\{\hat{\xi}\} = \sigma_0^2[\bar{A}^T(B^P^{-1}B^T)^{-1}\bar{A}]^{-1}. \] (9.12)

And the dispersion matrix for the residual vector reads

\[
D\{\hat{e}\} = P^{-1}B^T(B^P^{-1}B^T)^{-1}\{\bar{B} \cdot D\{\hat{e}\} \cdot B^T - \bar{A} \cdot D\{\hat{\xi}\} \cdot \bar{A}^T\}(B^P^{-1}B^T)^{-1}BP^{-1}, \] (9.13)

with \( D\{\hat{e}\} = \sigma_0^2P^{-1} \) as stated in the model (9.1).

Notation change: For the remainder of the discussion we drop the bars from the symbols as a matter of convenience. Recall that the bars were introduced in the first place to distinguish between the matrix \( B \) introduced in (9.5) and that used in Chapter 4 for the Model of Condition Equations. Dropping the bars means that \( \bar{B} \rightarrow B, \bar{w} \rightarrow w, \bar{A} \rightarrow BA \). With this simplified notation, we rewrite the solution (9.11a) as follows:

\[ \hat{\xi} = [(BA)^T(B^P^{-1}B^T)^{-1}BA]^{-1}(BA)^T(B^P^{-1}B^T)^{-1}w. \] (9.14)

The dispersion of \( \hat{\xi} \) is derived in parts as follows:

\[
D\{(BA)^T(B^P^{-1}B^T)^{-1}w\} = (BA)^T(B^P^{-1}B^T)^{-1}D\{w\}(B^P^{-1}B^T)^{-1}BA = \\
= \sigma_0^2(BA)^T(B^P^{-1}B^T)^{-1}(B^P^{-1}B^T)(B^P^{-1}B^T)^{-1}BA = \\
= \sigma_0^2(BA)^T(B^P^{-1}B^T)^{-1}BA,
\]
therefore
\[
D(\hat{\xi}) = [(BA)^T(BP^{-1}B)^{-1}BA]^{-1}D((BA)^T(BP^{-1}B)^{-1}w) \\
= [(BA)^T(BP^{-1}B)^{-1}BA]^{-1}\sigma_0^2(BA)^T(BP^{-1}B)^{-1}BA \\
\cdot [(BA)^T(BP^{-1}B)^{-1}BA]^{-1} \\
D(\hat{\xi}) = \sigma_0^2[(BA)^T(BP^{-1}B)^{-1}BA]^{-1}.
\] (9.15)

### 9.1 Iteratively Linearized Gauss-Helmert Model

In this section we present the Gauss-Helmert Model (GHM) as an iteratively linearized model, showing how to form both the model and the least-squares solution within the model at each step of an iteration scheme. The reader should be clear that the coefficient matrices \(A\) and \(B\) used in this section are not the same as the coefficient matrices \(A\) and \(B\) used in the GMM and Model of Condition Equations, respectively. For a more detailed discussion, see Schaffrin and Snow (2010).

Suppose we are given a non-linear functional model that relates \(n\) observations \(y\) to \(m\) unknown parameters \(\Xi\) among \(m + r\) non-linear condition equations \(b\) such that
\[
b(y - e_{\hat{\mu}}, \Xi) = 0, \quad e \sim (0, \sigma_0^2P^{-1}),
\] (9.16)
representing also a mapping \(b : \mathbb{R}^{m+n} \rightarrow \mathbb{R}^{m+r}\). Equation (9.16) is a non-linear Gauss-Helmert Model.

By introducing the “true” \(n \times 1\) vector of observables \(\mu\) as
\[
\mu := y - e = E\{y\},
\] (9.17)
the least-squares objective for model (9.16) is then defined by
\[
e^TPe = \min \text{ subject to } b(\mu, \Xi) = 0.
\] (9.18)

An iterative linearization of (9.16), together with the least-squares estimation of the unknown parameters \(\Xi\) and prediction of the unknown random errors \(e\), can be formed as follows.

Begin by assigning initial values to the unknowns \(\mu_0\) and \(\Xi_0\), e.g., \(\mu_0 = y - 0\) and \(\Xi_0 = (A^TPA)^{-1}A^TPy\). Then execute the following conditional loop.

While
\[
\delta < \|\hat{\xi}_j\| \quad \text{or} \quad \epsilon < \|\hat{e}^{(j)} - \hat{e}^{(j-1)}\|
\] (9.19)
for chosen thresholds \(\delta\) and \(\epsilon\), and \(j \in \mathbb{N}\), perform the following steps:

(i) Use the truncated Taylor series about expansion point \((\mu_j, \Xi_j)\):
\[
\begin{bmatrix}
\frac{\partial b}{\partial \mu} |_{\mu_j, \Xi_j} \quad \frac{\partial b}{\partial \Xi} |_{\mu_j, \Xi_j}
\end{bmatrix} \cdot 
\begin{bmatrix}
\mu - \mu_j \\
\Xi - \Xi_j
\end{bmatrix}
+ b(\mu_j, \Xi_j) = 0,
\] (9.20a)
and replace \( \mu \) with \( y - e \) in accordance with (9.17), to introduce

\[
\xi_{j+1} := \Xi - \Xi_j, \quad A^{(j)}_{(m+r) \times m} := \left. \frac{\partial b}{\partial \Xi^T} \right|_{\mu_j, \Xi_j}, \quad B^{(j)}_{(m+r) \times n} := \left. \frac{\partial b}{\partial \mu^T} \right|_{\mu_j, \Xi_j},
\]

(9.20b)

\[
w_j := b(\mu, \Xi_j) + B^{(j)} \cdot (y - \mu_j) \approx b(y, \Xi),
\]

(9.20c)

and to form the linearized Gauss-Helmert Model

\[
w_j = A^{(j)} \hat{\xi}_{j+1} + B^{(j)} e, \quad e \sim (0, \sigma_0^2 P^{-1}).
\]

(9.20d)

(ii) Produce the \((j + 1)\)th LEast-Squares Solution (LESS) for (9.20d), viz.

\[
\hat{\xi}_{j+1} = \left( (A^{(j)})^T \left[ (B^{(j)}) P^{-1} (B^{(j)})^T \right]^{-1} (A^{(j)}) \right)^{-1} \cdot (A^{(j)})^T \left[ (B^{(j)}) P^{-1} (B^{(j)})^T \right]^{-1} w_j,
\]

(9.20e)

\[
\hat{e}^{(j+1)} = P^{-1} (B^{(j)})^T \left[ (B^{(j)}) P^{-1} (B^{(j)})^T \right]^{-1} (w_j - (A^{(j)}) \hat{\xi}_{j+1}).
\]

(9.20f)

(iii) Obtain new approximate values (non-random) through

\[
\Xi_{j+1} := \hat{\Xi}^{(j+1)} - 0 = \Xi_j + \hat{\xi}_{j+1} - 0,
\]

(9.20g)

\[
\mu_{j+1} := \hat{\mu}^{(j+1)} - 0 = y - \hat{e}^{(j+1)} - 0,
\]

(9.20h)

where 0 denotes a “random zero vector” of suitable size (following Harville 1987). This means that the \( j \)th (approximate) estimates are stripped of the randomness while keeping their numerical values. While the use of 0 has some appeal to the theorist, it can be ignored in practical computations, since it does not affect numerical results.

Repeat the cycle until convergence is reached.

A already suggested, the initial approximate values for \( \mu \) might be taken from the observation vector \( y \) via \( \mu_0 := y - 0 \). Unfortunately, this has occasionally led to the misunderstanding that the so-called “misclosure vector” \( w_i \), in the \( i \)th iteration cycle, ought to be updated by \( b(\mu_i, \Xi_i) \) when, in fact, the correct update is described by (9.20c). An excellent treatment of this, and other potential pitfalls for solving non-linear least-squares problems, can be found in Pope (1972), which the reader is encouraged to read.

### 9.2 Estimated Variance Component

The \( P \)-weighted norm of the residual vector \( \hat{e} \) is defined as

\[
\Omega := \hat{e}^T P \hat{e} = (\hat{\lambda}^T B P^{-1}) P (P^{-1} B^T \hat{\lambda}) =
\]

(9.21b)
\[ ((w - BA\hat{\xi})^T(BP^{-1}B^T)^{-1})(BP^{-1}B^T)\hat{\lambda} = (9.21c) \]
\[ = (w - BA\hat{\xi})^T(BP^{-1}B^T)^{-1}(w - BA\hat{\xi}) = (9.21d) \]
\[ = (B\hat{e})^T(BP^{-1}B^T)^{-1}(B\hat{e}). \quad (9.21e) \]

Thus it follows that, the uniformly unbiased estimate of the variance component \( \sigma^2_0 \) is given by
\[ \hat{\sigma}^2_0 = (B\hat{e})^T(BP^{-1}B^T)^{-1}(B\hat{e}) \quad (9.22) \]
where, using the recombined matrix \( A \), the redundancy \( r \) is defined as
\[ r := rk B - rk(BA). \quad (9.23) \]

9.3 Equivalent Normal Equations

From (9.10b) and the second equation following (9.10c), we can recognize the following system of normal equations:
\[
\begin{bmatrix}
BP^{-1}B^T - BA \\
-(BA)^T 0
\end{bmatrix}
\begin{bmatrix}
\hat{\lambda} \\
\hat{\xi}
\end{bmatrix}
= \begin{bmatrix}
-w \\
0
\end{bmatrix}
\Rightarrow
\begin{bmatrix}
\hat{\lambda} \\
\hat{\xi}
\end{bmatrix}
= \begin{bmatrix}
BP^{-1}B^T & -BA \\
-(BA)^T & 0
\end{bmatrix}^{-1}
\begin{bmatrix}
-w \\
0
\end{bmatrix}.
\]

We want to show that the solution to this system yields the same \( \hat{\xi} \) as that of (9.14). The formula for the inverse of a partitioned matrix (see (A.14)) leads to the following solution:
\[
\begin{bmatrix}
\hat{\lambda} \\
\hat{\xi}
\end{bmatrix}
= \frac{1}{W}
\begin{bmatrix}
X_1 & X_2 \\
-W^{-1}(BA)^T(BP^{-1}B^T)^{-1} & (0-W)^{-1}
\end{bmatrix}
\begin{bmatrix}
-w \\
0
\end{bmatrix},
\]
with \( W := (BA)^T(BP^{-1}B^T)^{-1}BA \), and finally to
\[
\begin{bmatrix}
\hat{\lambda} \\
\hat{\xi}
\end{bmatrix}
= \frac{-X_1 w}{
\begin{bmatrix}
X_1 & X_2 \\
[(BA)^T(BP^{-1}B^T)^{-1}BA]^{-1}(BA)^T(BP^{-1}B^T)^{-1}w
\end{bmatrix}
\]
(9.25)

Here the symbols \( X_1 \) and \( X_2 \) represent quantities of no interest. We see that the solution for the parameters \( \xi \) is the same in (9.14).

9.4 Numerical Examples and Recommended Reading

The following papers include numerical examples for the GHM that may be of interest to the reader:

2. Line fitting in 3D: Snow and Schaffrin (2016).

3. 2D similarity transformations: Neitzel and Petrovic (2008).

In addition to these sources, Pope’s 1972 paper is highly recommended to understand how to best avoid potential pitfalls in adjusting data from iteratively linearized models.

### 9.5 Summary Formulas for the Least-Squares Solution Within the Gauss-Helmert Model

The linearized Gauss-Helmert Model (GHM) is given by

\[
\mathbf{w}_{(r+m) \times 1} = A_{(r+m) \times m} \mathbf{\xi} + B_{(r+m) \times n} \mathbf{e}, \quad \mathbf{e} \sim (0, \sigma_0^2 \mathbf{P}^{-1}).
\]

In the equations that follow, the matrix product \( BA \) that was used in the derivations above has been replaced by the symbol \( A \), which represents a more general form of the formulas. Recall that the matrix \( BA \) above included the coefficient matrix \( A \) from a Gauss-Markov Model (GMM). However, the more general formulation of the least-squares adjustment within the GHM would make no reference to quantities used in a GMM (see Section 9.1, for example).

<table>
<thead>
<tr>
<th>Quantity</th>
<th>Formula</th>
<th>Eq.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model redundancy</td>
<td>( r = \text{rk} B - \text{rk} A )</td>
<td>(9.23)</td>
</tr>
<tr>
<td>Vector of estimated parameters</td>
<td>( \hat{\mathbf{\xi}} = \left[ A^T (BP^{-1}B^T)^{-1} A \right]^{-1} A^T (BP^{-1}B^T)^{-1} \mathbf{w} )</td>
<td>(9.11a)</td>
</tr>
<tr>
<td>Dispersion matrix for estimated parameters</td>
<td>( D(\hat{\mathbf{\xi}}) = \sigma_0^2 \left[ A^T (BP^{-1}B^T)^{-1} A \right]^{-1} )</td>
<td>(9.12)</td>
</tr>
<tr>
<td>Vector of predicted residuals</td>
<td>( \hat{\mathbf{e}} = P^{-1}B^T (BP^{-1}B^T)^{-1} (\mathbf{w} - A\hat{\mathbf{\xi}}) )</td>
<td>(9.11b)</td>
</tr>
</tbody>
</table>

*Continued on next page*
<table>
<thead>
<tr>
<th>Quantity</th>
<th>Formula</th>
<th>Eq.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dispersion matrix for residuals</td>
<td>$D{\tilde{e}} = P^{-1}B^T (BP^{-1}B^T)^{-1} [B \cdot D{e} \cdot B^T - A \cdot D{\hat{\xi}} \cdot A^T] (BP^{-1}B^T)^{-1} BP^{-1}$</td>
<td>(9.13)</td>
</tr>
<tr>
<td>Sum of squared residuals (SSR)</td>
<td>$\Omega = \tilde{e}^T P \tilde{e}$</td>
<td>(9.21a)</td>
</tr>
<tr>
<td>Estimated variance component</td>
<td>$\hat{\sigma}_0^2 = \Omega / r$</td>
<td>(9.22)</td>
</tr>
</tbody>
</table>

Continued from previous page
Statistical Analysis

It is assumed that the reader of these notes has had at least a first course in statistical methods or probability theory and thus has some familiarity with hypothesis testing in statistical analysis. Therefore, key terms and concepts will be described only briefly, and the main focus will be placed on the application of hypothesis testing to parameters estimated from least-squares adjustments as described in the preceding chapters. For a broader treatment of statistical methods, and an excellent refresher on hypothesis testing in particular, see Snedecor and Cochran (1980).

Consider a normally distributed random (scalar) variable $y$ with the following first through fourth moments:

\[
\begin{align*}
E\{y\} &= \mu, & (10.1a) \\
E\{(y - \mu)^2\} &= D\{y\} = \sigma^2, & (10.1b) \\
E\{(y - \mu)^3\} &= 0, & (10.1c) \\
E\{(y - \mu)^4\} &= 3(\sigma^2)^2. & (10.1d)
\end{align*}
\]

The third moment being zero in (10.1c) means there is no skewness in the distribution of the random variable. The right side of (10.1d) indicates that there is no kurtosis (peak) in the distribution.

If (10.1c) or (10.1d) are not satisfied, the variable is not normally distributed and can be characterized as follows:

\[
\begin{align*}
E\{(y - \mu)^3\} > 0 & \iff \text{the distribution is skewed to the positive side.} & (10.2a) \\
E\{(y - \mu)^3\} < 0 & \iff \text{the distribution is skewed to the negative side.} & (10.2b) \\
E\{(y - \mu)^4\} - 3(\sigma^2)^2 > 0 & \iff \text{the distribution has positive kurtosis.} & (10.2c) \\
E\{(y - \mu)^4\} - 3(\sigma^2)^2 < 0 & \iff \text{the distribution has negative kurtosis.} & (10.2d)
\end{align*}
\]

Skewness appears in a graph of a sample of the random variable (e.g., a histogram) as a shift in the peak value from center. Positive kurtosis shows higher probability near the expected value $\mu$, which results in a taller, narrower graph. Negative
kurtosis shows higher probability in the tails of the graph; thus the graph appears flatter than that of a normally distributed variable.

The pdf (probability density function, or density function) of a normally distributed random (scalar) variable $y$ is

$$f(y) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(y-\mu)^2/2\sigma^2}, \quad (10.3)$$

where $\mu$ is the expectation of the distribution (population mean), $\sigma$ is standard deviation, $\sigma^2$ is variance, and $e$ is Euler’s number (i.e., the base of the natural logarithm). Note that the term $1/\sqrt{2\pi\sigma^2} \approx 0.4/\sigma$ denotes the amplitude of the graph of the curve, $\mu$ shows the offset of the peak from center, and $\sigma$ is the distance from the center to the inflection points of the curve.

The cdf (cumulative distribution function, or distribution function) of a normally distributed random variable is expressed as

$$F(y) = \int_{-\infty}^{y} f(t) \, dt = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{y} e^{-(t-\mu)^2/2\sigma^2} \, dt. \quad (10.4)$$

Figure 10.1 shows pdf and cdf plots for the normal distribution using various values for $\mu$ and $\sigma^2$. Line colors and types match between the pdf and cdf plots. The solid, green line represents the respective standard normal pdf and cdf curves.

Note that, in geodetic-science applications, the random variable $y$ might be an observation, an adjusted observation, a predicted residual, etc. We can standardize the random variable $y$ with the following transformation, which subtracts out the mean and divides by the standard deviation:

$$z = \frac{y - \mu}{\sigma}. \quad (10.5)$$

The standardized random variable $z$ has the following moments and probability functions:

$$E\{z\} = 0, \quad (10.6a)$$
$$D\{z\} = 1, \quad (10.6b)$$
$$pdf: \quad f(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}, \quad (10.6c)$$
$$cdf: \quad F(z) = \int_{-\infty}^{z} f(t) \, dt = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-t^2/2} \, dt. \quad (10.6d)$$

A plot of the pdf of $z$ is shown in Figure 10.2, along with example Student’s $t$-distribution curves (discussed below).

In the multivariate case, the random variable $y$ is an $n \times 1$ vector, with an $n \times n$ dispersion (covariance) matrix $\Sigma = D\{y\}$ and expectation vector $\mu = E\{y\}$, which is also size $n \times 1$. The pdf is then written as

$$f(y) = \frac{1}{(2\pi)^{n/2}\sqrt{\det \Sigma}} e^{-(y-\mu)^T \Sigma^{-1} (y-\mu)/2}. \quad (10.7)$$
Figure 10.1: pdf curve (top) and cdf curve (bottom) for the normal distribution with matching line types and colors so that the legend pertains to both graphs

And the cdf is written as

\[ F(y_1, \ldots, y_n) = \int_{-\infty}^{y_n} \cdots \int_{-\infty}^{y_1} f(t_1, \ldots, t_n) \, dt_1 \cdots dt_n. \]  

(10.8)
The elements of \( y \), i.e. \( y_1, \ldots, y_n \), are statistically independent if, and only if,
\[
f(t_1, \ldots, t_n) = f(t_1) \cdot f(t_2) \cdots f(t_n), \tag{10.9a}
\]
which implies
\[
C\{y_i, y_j\} = 0 \text{ for } i \neq j. \tag{10.9b}
\]
Equation (10.9b) states that there is no covariance between the elements of random vector \( y \).

The third and fourth moments for the multivariate case are given in (10.10a) and (10.10b), respectively.

\[
E\{(y_i - \mu_i)(y_j - \mu_j)(y_k - \mu_k)\} = 0 \text{ for } i, j, k = \{1, \ldots, n\} \tag{10.10a}
\]
\[
E\{(y_i - \mu_i)(y_j - \mu_j)(y_k - \mu_k)(y_l - \mu_l)\} = 3(\sigma_i^2)\delta_{ijkl} \text{ for } i, j, k, l = \{1, \ldots, n\} \tag{10.10b}
\]

In the following, we discuss studentized residuals, which have a \( t \)-distribution (or Student’s \( t \)-distribution). The pdf for a (scalar) variable having a \( t \)-distribution and \( \nu = n - 1 \) degrees of freedom is defined as follows:
\[
f(t) = \frac{1}{\sqrt{(n - 1)\pi}} \frac{\Gamma(n/2)}{\Gamma((n-1)/2)} \frac{1}{(1 + \frac{t^2}{n-1})^{n/2}}, \tag{10.11}
\]
where the gamma function is defined by
\[
\Gamma(n) := (n - 1)\Gamma(n - 1) = \int_0^\infty e^{-t}t^{n-1} dt = (n - 1)! \text{ for } n \in \mathbb{N}. \tag{10.12}
\]
As is known from introductory statistics, the pdf for the Student’s \( t \)-distribution resembles the pdf of the normal distribution when \( n \) is around 30. A plot of the pdf
for the Student’s t-distribution, with $\nu = 2, 4$, together with the pdf for the normal distribution, is shown in Figure 10.2.

## 10.1 Standardized and Studentized Residuals

We begin this section by restating the (full-rank) Gauss-Markov Model and writing the predicted vector of random errors within the model.

$$y = A\xi + e, \quad e \sim (0, \sigma_0^2 P^{-1}), \quad \text{rk} A = m \quad (10.13a)$$

$$\tilde{e} = (I_n - AN^{-1}A^T)e = (I_n - AN^{-1}A^T)P^{-1/2}(y - \tilde{y}) \quad (10.13b)$$

As usual, the observation vector $y$ is of size $n \times 1$, and the coefficient matrix $A$ is of size $n \times m$. Obviously, the far-right side of (10.13b) cannot be computed since $e$ is an unknown variable. However, the expression is useful for analytical purposes.

In the following, we assume that the random error vector $e$ has a normal distribution expressed by $e \sim N(0, \sigma_0^2 P^{-1})$ (where the symbol $N$ denotes normal distribution). This assumption is made for the sake of hypothesis testing in statistical analysis, which requires that test statistics be computed as a function of a specified pdf. The justification of the assumption owes to the central limit theorem as stated by Bjerhammar (1973, p. 35) (see also the footnote on page 110). However, since $e$ and $\sigma_0^2$ are unknown their respective prediction $\tilde{e}$ and estimate $\hat{\sigma}_0^2$ are used instead; consequently, the student $t$-distribution is used in place of the normal distribution for formulating hypothesis tests.

The so-called standardized residual is a function of the residual vector $\tilde{e}$ and its dispersion matrix $D\{\tilde{e}\}$ as shown in the following:

$$D\{\tilde{e}\} = \sigma_0^2(P^{-1} - AN^{-1}A^T), \quad (10.14a)$$

$$\sigma_{\tilde{e}_j}^2 = \eta_j^T D\{\tilde{e}\} \eta_j = E\{\tilde{e}_j^2\}, \quad (10.14b)$$

with

$$\eta_j := [0, \ldots, 0, 1 , 0, \ldots, 0]^T. \quad (10.14c)$$

Then, the $j$th standardized residual is defined as

$$\tilde{e}_j / \sigma_{\tilde{e}_j}. \quad (10.15)$$

Since the variance component $\sigma_0^2$ is considered unknown in the model (10.13a), we replace it with its estimate $\hat{\sigma}_0^2$, leading to the following analogous set of equations for the studentized residual:

$$\hat{\sigma}_0^2 = \frac{\tilde{e}^TP\tilde{e}}{n - \text{rk}(A)} = \frac{y^TPy - c^TN^{-1}c}{n - m}, \quad (10.16a)$$

$$\hat{D}\{\tilde{e}\} = \hat{\sigma}_0^2(P^{-1} - AN^{-1}A^T), \quad (10.16b)$$

$$\hat{\sigma}_{\tilde{e}_j}^2 = \eta_j^T \hat{D}\{\tilde{e}\} \eta_j = E\{\tilde{e}_j^2\}. \quad (10.16c)$$

---

1 The term test statistic is called test criterion by Snedecor and Cochran (1980, p. 65).
Then the studentized residual is defined as
\[
\hat{e}_j / \hat{\sigma}_0, \quad (10.17)
\]
Note that the denominator in (10.15) is constant (due to the unknown but constant variance component \( \sigma_0^2 \)), whereas the denominator of (10.17) is random due to the introduction of the estimate \( \hat{\sigma}_0^2 \), which is random. Of course the numerator is random in both cases.

Using \( Q \) to represent cofactor matrices, we can rewrite the standardized and studentized residuals in the following alternative forms:

**standardized residual:** \( \hat{e}_j / \sqrt{\sigma_0^2 (Q\hat{e})_{jj}} \sim N(0, 1) \), \( (10.18a) \)

**studentized residual:** \( \hat{e}_j / \sqrt{\hat{\sigma}_0^2 (Q\hat{e})_{jj}} \sim t(n - 1) \), \( (10.18b) \)

Here \( D\{\hat{e}\} = \sigma_0^2 Q\hat{e} \), and \((Q\hat{e})_{jj}\) denotes the \( j \)th diagonal element of the residual cofactor matrix \( Q\hat{e} \), and we have assumed that the standardized residuals are normally distributed, implying that the studentized residuals follow the student t-distribution. Again, it is noted that (10.18a) cannot be computed unless the variance component \( \sigma_0^2 \) is known.

**Example:** Direct observations of a single parameter \( \mu \) with weight matrix \( P = I_n \).
\[
y = \tau \mu + e, \quad e \sim N(0, \sigma_0^2 I_n), \quad \text{with} \quad \tau = [1, \ldots, 1]^T
\]
\[
\hat{\mu} = \frac{\tau^T y}{\tau^T \tau} = \frac{1}{n} \sum (y_1 + \ldots + y_n) \sim N(\mu, \sigma_0^2 / n)
\]
\[
\hat{e} = y - \tau \hat{\mu} \sim N(0, \sigma_0^2 [I_n - n^{-1} \cdot \tau \tau^T])
\]
\[
Q\hat{e} = I_n - n^{-1} \cdot \tau \tau^T
\]
\[
\hat{\sigma}_0^2 = \frac{\hat{e}^T \hat{e}}{n - 1}
\]

The formula for \( Q\hat{e} \) in the above example means that \((Q\hat{e})_{jj} = (n - 1)/n\), which shows that the more observations we have (i.e., the larger \( n \) is), the more the dispersion of the predicted random error \( D\{\hat{e}\} \) approaches the dispersion of the true random error \( D\{e\} \). In this example the standardized and studentized residuals are written as follows:

**standardized:** \( \frac{\hat{e}_j}{\sqrt{\sigma_0^2 (Q\hat{e})_{jj}}} = \frac{\hat{e}_j \sqrt{n}}{\sigma_0 \sqrt{n - 1}} \sim N(0, 1) \), \( (10.19a) \)

or alternatively: \( \frac{\hat{e}_j}{\sqrt{(Q\hat{e})_{jj}}} = \frac{\hat{e}_j \sqrt{n}}{\sqrt{n - 1}} \sim N(0, \sigma_0^2) \), \( (10.19b) \)

**studentized:** \( \frac{\hat{e}_j}{\sqrt{\hat{\sigma}_0^2 (Q\hat{e})_{jj}}} = \frac{\hat{e}_j \sqrt{n}}{\sqrt{\hat{e}^T \hat{e}}} \sim t(n - 1) \), \( (10.19c) \)
We extend the example by including a hypothesis test for the parameter estimate $\hat{\mu}$ against a specified value $\mu_0$ at a significance level $\alpha$ (see Section 10.2 for a more complete discussion of hypothesis testing).

Hypothesis test:  
\[ H_0 : E\{\hat{\mu}\} = \mu_0 \text{ against } H_A : E\{\hat{\mu}\} \neq \mu_0. \]

Test statistic:  
\[ t = \frac{\hat{\mu} - \mu_0}{\sqrt{\hat{\sigma}_0^2/n}} \sim t(n-1). \]

We accept the null hypothesis $H_0$ if $t_{-\alpha/2} \leq t \leq t_{\alpha/2}$; otherwise we reject $H_0$.

We may perform a similar test $H_0 : E\{\tilde{e}_j\} = 0$ for the $j$th residual. In this case the test statistic is the studentized residual computed by (10.19c).

### 10.2 Hypothesis Testing Within the Gauss-Markov Model

The hypothesis test introduced in Section 10.1 for direct observations of a single parameter is now extended to the Gauss-Markov Model (GMM). In introducing the GMM in Chapter 3, a probability density function was not given for the random observation errors; only the first and second moments of the random errors were specified. This is indeed all that is necessary to formulate and solve the least-squares estimation problem within the GMM. However, in order to perform classical hypothesis testing after the least-squares estimate has been computed, the probability distribution must be specified. Typically, we assume that the observation errors have a normal distribution. Then, the (full rank) GMM is written succinctly as

\[ \mathbf{y}_{n \times 1} = A_{n \times m} \mathbf{\xi} + \mathbf{e}, \quad \text{rk} A = m, \quad \mathbf{e} \sim \mathcal{N}(\mathbf{0}, \sigma_0^2 \mathbf{P}^{-1}). \quad (10.20) \]

Minimization of the observation errors via a least-squares adjustment leads to the following parameter estimate and predicted random-error vectors, shown with their corresponding normal distributions (normal because the distribution of the observations were assumed to be normal for the sake of hypothesis testing):

\[ \hat{\mathbf{\xi}} = N^{-1} \mathbf{c} \sim \mathcal{N}(\mathbf{\xi}, \sigma_0^2 N^{-1}), \quad (10.21a) \]
\[ \hat{\mathbf{e}} = (I_n - AN^{-1}A^T \mathbf{P}) \mathbf{y} \sim \mathcal{N}(\mathbf{0}, \sigma_0^2 [P^{-1} - AN^{-1}A^T]), \quad (10.21b) \]

Or equivalently, we could write for the predicted residual vector

\[ \hat{\mathbf{e}} = (I_n - AN^{-1}A^T \mathbf{P}) \mathbf{e} = \mathbf{Q}_\mathbf{e} \mathbf{P} \mathbf{y} \sim \mathcal{N}(\mathbf{0}, \sigma_0^2 \mathbf{Q}_\mathbf{e}), \quad (10.22a) \]

with

\[ \mathbf{Q}_\mathbf{e} := P^{-1} - AN^{-1}A^T. \quad (10.22b) \]

The $j$th standardized and studentized residuals are then written as

\[ j\text{th standardized residual: } \tilde{e}_j \left/ \sqrt{\sigma_0^2(Q_{\mathbf{e}})_{jj}} \right. \sim \mathcal{N}(0, 1), \quad (10.23) \]
As shown in Chapter 3, we compute the estimated reference variance within this model by
\[
\hat{\sigma}^2_0 = \frac{\bar{e}^T P \bar{e}}{n-m}.
\]
(10.25)
The hypothesis test for the \( j \)th studentized residual then becomes
\[
H_0 : E\{\tilde{e}_j\} = 0 \quad \text{versus} \quad H_A : E\{\tilde{e}_j\} \neq 0.
\]
(10.26)
Likewise, we may test individual elements of the estimated parameter vector \( \hat{\xi} \). For example, we may want to compare the \( j \)th element of the estimated parameter vector, \( \hat{\xi}_j \), against some specified value \( \xi_j^{(0)} \). In this case, the null hypothesis and computed test statistic are defined as follows:
\[
H_0 : E\{\hat{\xi}_j\} = \xi_j^{(0)} \quad \text{versus} \quad H_A : E\{\hat{\xi}_j\} \neq \xi_j^{(0)},
\]
(10.27a)
\[
t_j = \frac{\hat{\xi}_j - \xi_j^{(0)}}{\sqrt{\hat{\sigma}^2_0 (N-1)_{jj}}} \sim t(n-m) \text{ or } t^2_j \sim F(1,n-m).
\]
(10.27b)
Note that from (10.27b) we see that the square of a test statistic having a Student’s \( t \)-distribution has an \( F \)-distribution. For a given significance level \( \alpha \), we accept \( H_0 \) if \( t_{-\alpha/2} \leq t_j \leq t_{\alpha/2} \); otherwise we reject \( H_0 \). We can use a cdf table for the \( t \)-distribution to find the value of \( t_{\alpha/2}(n-m) \). Note that \( \alpha \) is the probability of making a Type I error (also called the significance level of the test), and \( n-m \) is the degrees of freedom, often denoted by \( \nu \) in the statistical literature.

10.3 Confidence Intervals for Ellipses, Ellipsoids, and Hyperellipsoids

When we estimate the mean of a population, or the parameter of a data model, we would like to be able to make some statement of the accuracy of the estimate. In statistics, a probability statement gives the probability that the estimated quantity falls within a certain interval centered on the true, but unknown mean (or model parameter). Such an interval is called a confidence interval, and its upper and lower bounds are called confidence limits. Confidence ellipses, ellipsoids, and hyperellipsoids are the respective 2-D, 3-D, and \( n \)-D analogues to confidence intervals.

10.3.1 Confidence Intervals — Univariate Case

By definition, the cdf (cumulative distribution function) of a random variable \( X \) is
\[
F_X(x) = P(X \leq x), \quad -\infty < x < \infty,
\]
(10.28)
which provides the probability that the unknown quantity \( X \) is less than or equal to the sampled value \( x \). It follows, then, that the probability that \( X \) lies within the interval \((a, b]\) is

\[
P(a < X \leq b) = F_X(b) - F_X(a). \tag{10.29}
\]

Applying (10.29) to the standard normal random variable \( z \) of (10.5), we can write the following probabilities for confidence intervals bounded by \( \pm \sigma \), \( \pm 2\sigma \), \( \pm 3\sigma \), respectively, from the mean, where \( \sigma = 1 \) since \( z \sim N(0, 1) \) according to (10.6a) and (10.6b):

\[
P(-1 < z \leq 1) = P(\mu - \sigma < y \leq \mu + \sigma) = 68.3\% \tag{10.30a}
\]
\[
P(-2 < z \leq 2) = P(\mu - 2\sigma < y \leq \mu + 2\sigma) = 95.5\% \tag{10.30b}
\]
\[
P(-3 < z \leq 3) = P(\mu - 3\sigma < y \leq \mu + 3\sigma) = 99.7\% \tag{10.30c}
\]

The intervals associated with these probability statements are commonly referred to as the “1-sigma,” “2-sigma,” and “3-sigma” confidence intervals, respectively. Other commonly used intervals are the so-called 90%, 95%, and 99% confidence intervals. For a normally distributed random variable \( z \), their respective probability statements are

\[
90\% = P(-1.645 < z \leq 1.645), \tag{10.31a}
\]
\[
95\% = P(-1.960 < z \leq 1.960), \tag{10.31b}
\]
\[
99\% = P(-2.576 < z \leq 2.576). \tag{10.31c}
\]

Probability limits correspond to the area under the graph of the associated pdf. For example, the area between \( \pm \sigma \) under the graph of the standard normal distribution shown in Figure 10.2 is 0.683, and it is 0.997 for \( \pm 3\sigma \). The regions beyond these areas are called the tails of the graph. See Figure 10.3 for a graphical representation.

### 10.3.2 Confidence Ellipses — Bivariate Case

The 2-D analogue to a confidence interval is a confidence ellipse, which can be generated from

\[
(y - \mu)^T \Sigma^{-1} (y - \mu) = k^2, \tag{10.32b}
\]

where \( k \) is a constant, and \( \rho \) is called the correlation coefficient, which is defined by

\[
\rho_{12} = \frac{\sigma_{12}}{\sigma_1 \sigma_2}. \tag{10.33}
\]

By varying \( k \), we generate a family of ellipses, each having an associated constant probability. We are using vector notation again: \( y \) is a random 2-D vector and \( \mu \)
is the expected value of $y$; i.e., $\mu = E\{y\}$. Also, we have a $2 \times 2$ dispersion matrix for $y$, namely $\Sigma$. More specifically we have

$$y = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix}, \quad \mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} = E\{ \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} \}, \quad \Sigma := D\{y\} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{bmatrix}, \quad \sigma_{12} = \sigma_{21}. \tag{10.34}$$

When speaking of the elements of the vectors and matrix in (10.34), we say that $\mu_1$ is the expected value of $y_1$; $\sigma_1^2$ is the variance of $y_1$ (with $\sigma_1$ called standard deviation), and $\sigma_{12}$ is the covariance between $y_1$ and $y_2$.

Using the above definitions, together with equation (10.7), we can write the pdf of $y$ explicitly as

$$f(y) = \frac{1}{2\pi \sigma_1 \sigma_2 \sqrt{1 - \rho_{12}^2}} \cdot \exp\left\{ -\frac{1}{2\sigma_1^2} \frac{(y_1 - \mu_1)^2}{\sigma_1^2} - \left( \frac{2\sigma_{12} (y_1 - \mu_1) (y_2 - \mu_2)}{\sigma_1^2 \sigma_2} \right) + \frac{(y_2 - \mu_2)^2}{\sigma_2^2} \right\}, \tag{10.35}$$

where $\exp$ stands for the exponential function, e.g., $\exp\{x\} = e^x$.

Each element of the vector $y$ may be normalized according to (10.5), so that the $j$th element of the normalized vector $z$ is expressed in terms of the corresponding $j$th element of $y$; that is $z_j = (y_j - \mu_j)/\sigma_j$, $j = 1, 2$. Substituting $z_j$ and $\rho_{12}$ into (10.35) we can write the following pdf for the 2-D vector $z$:

$$f(z_1, z_2) = \frac{1}{2\pi \sigma_1 \sigma_2 \sqrt{1 - \rho_{12}^2}} \cdot \exp\left\{ -\frac{1}{2(1 - \rho_{12}^2)} \left( z_1^2 - 2\rho_{12} z_1 z_2 + z_2^2 \right) \right\}. \tag{10.36}$$
Setting \( k = 1 \) in (10.32a) gives the equation for the standard confidence ellipse (also called standard error ellipse by Mikhail and Gracie (1981)). From (10.36) we see that the standard confidence ellipse can also be described by

\[
z_1^2 - 2\rho_{12}z_1z_2 + z_2^2 = 1 - \rho_{12}^2.
\]

The size, shape, and orientation of the confidence ellipse are determined by the eigenvalues and eigenvectors of the dispersion matrix \( \Sigma \).

The eigenvector-eigenvalue decomposition of the \( 2 \times 2 \) matrix \( \Sigma \) is described as follows: Denote the eigenvectors of \( \Sigma \) as \( u \), and the eigenvalues as \( \lambda_j \), \( j = 1, 2 \). Then we have the relation

\[
\Sigma u_j = \lambda_j u_j,
\]

for which we write the following characteristic equation:

\[
\det(\Sigma - \lambda I_2) = (\sigma_1^2 - \lambda)(\sigma_2^2 - \lambda) - \sigma_{12}^2 = \lambda^2 - (\sigma_1^2 + \sigma_2^2)\lambda + (\sigma_1^2\sigma_2^2 - \sigma_{12}^2) = 0. \tag{10.39}
\]

In (10.39), \( \lambda \) has been used in general to represent either eigenvalue \( \lambda_1 \) or \( \lambda_2 \). By convention, we require \( \lambda_1 \geq \lambda_2 > 0 \) and write the following solution for the roots of the characteristic equation (10.39):

\[
\lambda_{1 \text{ or } 2} = \frac{\sigma_1^2 + \sigma_2^2}{2} \pm \frac{1}{2} \sqrt{\left(\frac{\sigma_1^2 + \sigma_2^2}{2}\right)^2 - 4\sigma_1^2\sigma_2^2 + 4\sigma_{12}^2} \Rightarrow \tag{10.40a}
\]

\[
\lambda_{1 \text{ or } 2} = \frac{\sigma_1^2 + \sigma_2^2}{2} \pm \frac{1}{2} \sqrt{(\sigma_1^2 - \sigma_2^2)^2 + 4\sigma_{12}^2} > 0, \tag{10.40b}
\]

which shows that the eigenvalues must be greater than zero, since \( \Sigma \) is positive definite.

Now we must find the two corresponding eigenvectors. Let the matrix \( U \) be comprised of the two eigenvectors \( u_1 \) and \( u_2 \) such that \( U := [u_1, u_2] \). Also define a diagonal matrix comprised of the corresponding eigenvalues \( \Lambda := \text{diag}(\lambda_1, \lambda_2) \). Then according to (10.38) we have

\[
\Sigma U = U\Lambda = \tag{10.41a}
\]

\[
= \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix} \begin{bmatrix} u_{11} & u_{12} \\ u_{21} & u_{22} \end{bmatrix} = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix} \Rightarrow \tag{10.41b}
\]

\[
= \begin{bmatrix} \sigma_1^2 u_{11} + \sigma_{12} u_{21} & \sigma_1^2 u_{12} + \sigma_{12} u_{22} \\ \sigma_{12} u_{11} + \sigma_2^2 u_{21} & \sigma_{12} u_{12} + \sigma_2^2 u_{22} \end{bmatrix} = \begin{bmatrix} \lambda_1 \cdot u_{11} & \lambda_2 \cdot u_{12} \\ \lambda_1 \cdot u_{21} & \lambda_2 \cdot u_{22} \end{bmatrix}. \tag{10.41c}
\]

From (10.41c) we can write the following four equations in the four unknowns \( u_{11}, u_{12}, u_{21}, \) and \( u_{22} \):

\[
\begin{align*}
&u_{21} = \frac{(\lambda_1 - \sigma_1^2)u_{11}}{\sigma_{12}}, & u_{21} &= \frac{\sigma_{12} u_{11}}{\lambda_1 - \sigma_1^2}, & u_{12} &= \frac{\sigma_{12} u_{22}}{\lambda_2 - \sigma_1^2}, & u_{12} &= \frac{(\lambda_2 - \sigma_2^2)u_{22}}{\sigma_{12}}. & & \tag{10.42}
\end{align*}
\]
The eigenvector \( u_1 \) defines the direction of the *semimajor axis* of the confidence ellipse, while the eigenvector \( u_2 \), orthogonal to \( u_1 \), defines the *semiminor axis* direction. The square root of the eigenvalue \( \lambda_1 \) gives the semimajor-axis length, and the square root of the eigenvalue \( \lambda_2 \) gives the semiminor-axis length. Also, if \( \theta \) is the angle measured counter clockwise from the positive \( z_1 \)-axis to the semimajor axis of the confidence ellipse, then we can write the matrix \( U \) as

\[
U = [u_1, u_2] = \begin{bmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{bmatrix}.
\] (10.43)

Using (10.42) and (10.43), the angle \( \theta \) is derived as follows:

\[
\tan \theta = \frac{\sin \theta}{\cos \theta} = \frac{u_{21}}{u_{11}} = \frac{\lambda_1 - \sigma_1^2}{\sigma_{12}} = \frac{\sigma_{12}}{\lambda_1 - \sigma_2^2} = -\frac{u_{12}}{u_{22}} = \frac{\sigma_2^2 - \lambda_2}{\sigma_{12}} = \frac{\sigma_{12}}{\sigma_1^2 - \lambda_2}
\]

and

\[
\tan(2\theta) = \tan(\theta + \theta) = \frac{2\tan \theta}{1 - \tan^2 \theta} = \frac{2\sigma_{12}}{\lambda_1 - \sigma_2^2} \left[ 1 - \frac{1}{\sigma_{12}} \left( \frac{\lambda_1 - \sigma_2^2}{\lambda_1 - \sigma_2^2} \right)^2 \right] = \tan(2\theta) = \frac{2\sigma_{12} (\lambda_1 - \sigma_2^2)}{(\lambda_1 - \sigma_2^2)^2 - \sigma_{12}^2} = \frac{2\sigma_{12} (\lambda_1 - \sigma_2^2)}{[2(\lambda_1 - \sigma_2^2)]^2 - 4\sigma_{12}^2}.
\] (10.44c)

By manipulating (10.40b), we have

\[
2(\lambda_1 - \sigma_2^2) = (\sigma_1^2 - \sigma_2^2) \pm \sqrt{(\sigma_1^2 - \sigma_2^2)^2 + 4\sigma_{12}^2}\Rightarrow
\]

\[
[2(\lambda_1 - \sigma_2^2)]^2 = 2(\sigma_1^2 - \sigma_2^2)^2 \pm 2(\sigma_1^2 - \sigma_2^2) \sqrt{(\sigma_1^2 - \sigma_2^2)^2 + 4\sigma_{12}^2} + 4\sigma_{12}^2 + 4\sigma_{12}^2.
\] (10.45b)

Substituting (10.45a) and (10.45b) into (10.44c) gives

\[
\tan(2\theta) = \frac{4\sigma_{12} \left[ (\sigma_1^2 - \sigma_2^2) \pm \sqrt{(\sigma_1^2 - \sigma_2^2)^2 + 4\sigma_{12}^2} \right]}{2(\sigma_1^2 - \sigma_2^2) \left[ (\sigma_1^2 - \sigma_2^2) \pm \sqrt{(\sigma_1^2 - \sigma_2^2)^2 + 4\sigma_{12}^2} \right]} = \tan(2\theta) = \frac{2\sigma_{12}}{\sigma_1^2 - \sigma_2^2}.
\] (10.46b)

The sign of the numerical value of the right side of (10.46b) tells which quadrant the positive axis falls in.

Returning now to the notion of ellipses of constant probability represented by (10.32a), probabilities for various values of \( k \) are most easily determined by using a transformed system of equations centered on \( \mu \) and rotated so that the \( y_1 \) and \( y_2 \) axes coincide with the axes formed by the eigenvectors \( u_1 \) and \( u_2 \). Then, instead of correlated coordinates \( y_1 \) and \( y_2 \), we end up with uncorrelated coordinates \( u_1 \) and \( u_2 \).
and \( u_2 \) with respective variances \( \lambda_1 \) and \( \lambda_2 \) from (10.40b). And so the probability statement for being on or within an ellipse having semimajor and semiminor axes \( k\sqrt{\lambda_1} \) and \( k\sqrt{\lambda_2} \), respectively, is

\[
P\left\{ \frac{u_1^2}{\lambda_1} + \frac{u_2^2}{\lambda_2} < k^2 \right\} = P\left\{ \chi_2^2 < k^2 \right\} = 1 - \alpha, \quad (10.47)
\]

where \( \alpha \) is a specified level of significance. Because it is assumed that \( u_1 \) and \( u_2 \) are sampled from a normal distribution, the sum of their squares has a \( \chi_2^2 \) distribution. See section 10.4.1 for a description of the \( \chi_2^2 \) distribution.

Given a value for \( P = 1 - \alpha \), the value of \( k \) (or visa verse) can be determined from a table of values for the \( \chi_2^2 \) density function. Users of MATLAB® can generate \( P \) given \( k \) by using \( P = \text{chi2cdf}(k^2, 2) \), and \( k \) given \( P \) can be generated by \( k = \sqrt{\text{chi2inv}(P, 2)} \). Commonly used values are shown in Table 10.1. Compare the probability of 39.4% associated with the 1-sigma confidence ellipse to the value of 68.3% shown in (10.30a) for the 1-sigma confidence interval in the univariate case.

Table 10.1: “k-sigma” probabilities for various confidence ellipses. \( P = 1 - \alpha \).

<table>
<thead>
<tr>
<th>( P )</th>
<th>0.394</th>
<th>0.500</th>
<th>0.900</th>
<th>0.950</th>
<th>0.990</th>
</tr>
</thead>
<tbody>
<tr>
<td>( k )</td>
<td>1.000</td>
<td>1.177</td>
<td>2.146</td>
<td>2.447</td>
<td>3.035</td>
</tr>
</tbody>
</table>

An empirical error ellipse differs from the confidence ellipse in that the matrix \( \Sigma \) is replaced by the estimated matrix \( \hat{\Sigma}, \) such that \( \hat{\Sigma}^{-1} = \hat{\sigma}_0^{-2}P, \) where \( \hat{\sigma}_0^2 \) is the estimated variance component. In this case, the empirical error ellipse is described by

\[
\frac{(y - \hat{\mu})^T P (y - \hat{\mu})}{\hat{\sigma}_0^2} = 1. \quad (10.48)
\]

If we are evaluating \( n/2 \) number of 2-D points, so that \( P \) is of size \( n \times n, \) we may simply work with each of the \( (n/2 \) number of) \( 2 \times 2 \) block diagonal matrices of \( \hat{\sigma}_0^{-2}P \) independently to form the empirical error ellipse of each point. However, we must bear in mind that these block diagonal matrices do not tell the whole story since the off-block-diagonal elements have been ignored. In any case, it may be prudent to verify that the associated correlation-coefficients of the off-block-diagonal elements are relatively small in magnitude.

The following two examples apply to the Gauss-Markov Model (GMM):

1. Consider the GMM (10.20), with an associated least-squares solution and dispersion given in (10.21a). Assume that the parameter vector \( \xi \) is comprised of successive 2-D point coordinates such that \( (\hat{\xi}_{2i-1}, \hat{\xi}_{2i}) \) represents the coordinate estimates of the \( i \)th point. Now, also assume that we wish to compare the estimates with given (fixed) values \( (\xi_{2i-1}^0, \xi_{2i}^0), \) perhaps from published results of a previous adjustment. Then we may write the following equations for the null hypothesis and the error ellipse, where, for convenience, \( k := 2i \)
and \( j := k - 1 \):

\[
H_0 : E\{[\hat{\xi}_j, \hat{\xi}_k]^T\} = [\xi_0^j, \xi_0^k]^T, \quad (10.49a)
\]

\[
\frac{1}{\hat{\sigma}_0^2} \begin{bmatrix}
\hat{\xi}_j - \xi_0^j \\
\hat{\xi}_k - \xi_0^k
\end{bmatrix}^T \begin{bmatrix}
N_{j,j} & N_{j,k} \\
N_{k,j} & N_{k,k}
\end{bmatrix} \begin{bmatrix}
\hat{\xi}_j - \xi_0^j \\
\hat{\xi}_k - \xi_0^k
\end{bmatrix} = 1. \quad (10.49b)
\]

2. Suppose that instead of comparing the solution to given, fixed values we want to compare the results (2-D coordinate estimates) of two adjustments. Using the previously defined indices, let the estimates of the \( i \)th point of the second adjustment be represented by \((\hat{\xi}_j, \hat{\xi}_k)\). We ask the question: is the outcome of the second adjustment statistically equivalent to the first? Unless there is statistically significant overlap of the respective error ellipses, the answer is no. The null hypothesis \( H_0 \) and the test statistic \( f \) are defined as follows:

\[
H_0 : E\{[\hat{\xi}_j, \hat{\xi}_k]^T\} = E\{[\hat{\xi}_j, \hat{\xi}_k]^T\} \quad (10.50a)
\]

\[
f := \frac{1}{2} \frac{1}{\hat{\sigma}_0^2/\hat{\sigma}_0^2} \begin{bmatrix}
\hat{\xi}_j - \hat{\xi}_j \\
\hat{\xi}_k - \hat{\xi}_k
\end{bmatrix}^T D\{\begin{bmatrix}
\hat{\xi}_j - \hat{\xi}_j \\
\hat{\xi}_k - \hat{\xi}_k
\end{bmatrix}\}^{-1} \begin{bmatrix}
\hat{\xi}_j - \hat{\xi}_j \\
\hat{\xi}_k - \hat{\xi}_k
\end{bmatrix} \sim F(2, n - rk A) \quad (10.50b)
\]

In computing the test statistic \( f \) shown in (10.50b), it is assumed that the estimated variance component \( \hat{\sigma}_0^2 \) is common to both adjustments. This assumption is based on the homogeneity test \( H_0 : E\{\hat{\sigma}_0^2\} = E\{\hat{\sigma}_0^2\} \), which is discussed in Section 10.4. Note that in the case that the two adjustments are uncorrelated, we could replace the differences of parameters in the inverted dispersion matrix with their sums, since the variance of the sum of two random variables is equivalent to the variance of their difference when they are uncorrelated.

### 10.3.3 Confidence Ellipsoids and Hyperellipsoids — Multivariate Case

In the 3-D case, confidence ellipses are extended to confidence ellipsoids. But, in our general formulation of the GMM we may be working with any arbitrary higher-dimensional space, and thus we speak of confidence hyperellipsoids. Since 3-D and higher dimensions are natural extensions of the 2-D case, no further discussion is necessary. However, we do list probabilities associated with confidence ellipsoids for the 3-D case in Table 10.2. The table entries can be generated using the same MATLAB® commands shown in the previous section, except that the second argument must be 3 (degrees of freedom) instead of 2.
10.4 \( \chi^2 \)-distribution, Variance Testing, and \( F \)-distribution

This section includes the statistical topics of \( \chi^2 \)- and \( F \)-distributions as well as the topic of variance testing.

10.4.1 \( \chi^2 \)-distribution

The \( \chi^2 \)-distribution is attributed to the German geodesist F.R. Helmert from 1876. If we claim that the (unknown) random error vector \( e \) from the GMM is normally distributed as \( e \sim N(0, \sigma_0^2 P^{-1}) \), then the quadratic product \( e^T P e \) has a \( \chi^2 \)-distribution with \( \nu := \text{rk} \ P = n \) degrees of freedom, expressed by

\[
\frac{e^T P e}{\sigma_0^2} \sim \chi^2(\nu). \tag{10.51}
\]

Now, define \( x := \frac{e^T P e}{\sigma_0^2} \) (which cannot actually be computed since both \( e \) and \( \sigma_0^2 \) are unknown). Therefore, the pdf of \( x \) is written as

\[
f(x) = \begin{cases} 
\frac{1}{2^{\nu/2} \Gamma(\nu/2)} x^{(\nu-2)/2} e^{-x/2} & \text{for } x > 0 \\
0 & \text{for } x \leq 0,
\end{cases} \tag{10.52}
\]

where \( e \) is Euler's number 2.71828\ldots The gamma function \( \Gamma(\cdot) \) was defined in (10.12). Figure 10.4 shows plots of the \( \chi^2 \)-distribution for \( \nu = \{1, 3, 5, 8, 10, 30\} \) with respective colors: black, magenta, cyan, red, green, blue. Note that the peaks of the curves move to the right as \( \nu \) increases and that the curves appear to approximate the normal-distribution curve as \( \nu \) grows to 10 and larger. This agrees with our expectation that the \( \chi^2 \)-distribution is asymptotically normal, due to the central limit theorem.\(^2\)

From the variance component derivations in Section 3.2, we can write

\[
E\{e^T P e / \sigma_0^2\} = \text{tr}(P \cdot E\{ee^T\}) = \text{tr} \ I_n = n, \tag{10.53a}
\]

\[
E\{\tilde{e}^T \tilde{e} / \sigma_0^2\} = \text{tr}(P \cdot E\{\tilde{e}\tilde{e}^T\}) = \text{tr}(I_n - AN^{-1}A^T P) = n - \text{rk} \ A = n - m. \tag{10.53b}
\]

\(^2\)According to Bjorhammar (1973, section 2.15), the central limit theorem says that “the sum of \( n \) independent stochastic variables having equal expectation and variance will have a distribution that converges towards the normal distribution for \( n \to \infty \).”

Table 10.2: “\( k \)-sigma” probabilities for various confidence ellipsoids. \( P = 1 - \alpha \).

<table>
<thead>
<tr>
<th>P</th>
<th>0.199</th>
<th>0.500</th>
<th>0.900</th>
<th>0.950</th>
<th>0.990</th>
</tr>
</thead>
<tbody>
<tr>
<td>k</td>
<td>1.000</td>
<td>1.538</td>
<td>2.500</td>
<td>2.796</td>
<td>3.365</td>
</tr>
</tbody>
</table>

109
Equations (10.25) and (10.53b) lead to
\[ \hat{\mathbf{e}}^T \mathbf{P} \hat{\mathbf{e}} / \sigma_0^2 = \nu \hat{\sigma}_0^2 / \sigma_0^2 \sim \chi^2(\nu), \]
with
\[ \nu := n - m \]  
(10.54)as the degrees of freedom.

Note that though we have been discussing the random error vector \( \mathbf{e} \) and the predicted residual \( \hat{\mathbf{e}} \), the relations expressed in (10.54a) apply to all quadratic forms in normally distributed variables. Thus, when we have a vector of normally distributed variables, the corresponding quadratic form will have a \( \chi^2 \)-distribution.

### 10.4.2 Variance Testing

Suppose we want to compare the estimated variance component \( \hat{\sigma}_0^2 \) to a given quantity \( \sigma^2 \) (in the latter, the 0-subscript is not used so as not to confuse the given value with the unknown “true value”). We do so by performing the following hypothesis test at a chosen significance level \( \alpha \):

\[ H_0 : E\{\hat{\sigma}_0^2\} \leq \sigma^2 \text{ vs. } H_A : E\{\hat{\sigma}_0^2\} > \sigma^2 \]  
(10.55a)

\[ t := (n - m) \cdot (\hat{\sigma}_0^2 / \sigma^2) \sim \chi^2(n - m) \]  
(10.55b)
If $t \leq \chi^2_\alpha$ accept $H_0$; else reject $H_0$. (10.55c)

The test as shown is referred to as a one-tailed test, because the null hypothesis only states that the expectation of the estimated quantity is less than or equal to a given value (the use of $\geq$ would also constitute a one-tailed test). In contrast a two-tailed test would use an equals sign in the null hypothesis. This jargon comes from the fact that $1 - \alpha$ represents the area under the pdf curve left of the right tail in the one-tailed case, and it represents the area between both the left and right tails (each of which have area $\alpha/2$) in the two-tailed case.

Under the assumption that the data model is correct, if the estimate $\hat{\sigma}_0^2$ turns out statistically to be less than the given value $\sigma^2$, we deem our measurements to be more precise than that reflected in the weight matrix $P$. On the other hand, if $\hat{\sigma}_0^2$ proves statistically to be greater than the given value, we deem our measurements to be less precise. Usually our main concern is that $\hat{\sigma}_0^2$ reflects that our measurements are at least as precise as what is reflected by the elements of the weight matrix $P$, thus the use of a single-tailed hypothesis may be more commonly used in practice.

In the case where we need to compare two estimated reference variances $\hat{\sigma}_{0,1}^2$ and $\hat{\sigma}_{0,2}^2$ from two independent adjustments, we must compute a ratio of test statistics, which has an $F$-distribution (assuming both the numerator and denominator have $\chi^2$-distributions). Let $t_1$ and $t_2$ be the test statistics from the respective adjustments; then we can write

$$\frac{t_1/(n_1 - m_1)}{t_2/(n_2 - m_2)} = \frac{\hat{\sigma}_{0,1}^2}{\hat{\sigma}_{0,2}^2} \sim F(n_1 - m_1, n_2 - m_2),$$

(10.56)

where $n_i - m_i$, $i = 1, 2$, are the respective degrees of freedom of the two independent adjustments.

### 10.4.3 $F$-distribution

The $F$-distribution was named for its discover R.A. Fisher (1925) by G.W. Snedecor (1935). It is a distribution for the ratio of two mutually independent random variables that have $\chi^2$-distributions with degrees of freedom $v_1 := m$ and $v_2 := n - m$, respectively. The pdf of such a variable is given by

$$f(w) = \frac{\Gamma\left(\frac{m}{2} + \frac{n-m}{2}\right)m^{m/2}(n-m)^{(n-m)/2}w^{(m/2)-1}}{\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{n-m}{2}\right)(n-m+m)w^{(m/2+(n-m)/2)}} = \frac{(v_1/v_2)^{v_1/2}\Gamma((v_1 + v_2)/2)w^{(v_1/2)-1}}{\Gamma(v_1/2)\Gamma(v_2/2)(1 + v_1 w/v_2)^{(v_1+v_2)/2}}.$$  

(10.57a)

(10.57b)

As $n$ becomes large compared to $m$, the curve of the $F$-distribution approaches the curve of the normal distribution.
10.5  Hypothesis Testing on the Estimated Parameters

In the GMM, we may wish to perform a global model-check by comparing a specified parameter vector \( \xi^0 \) to the estimated vector \( \hat{\xi} \). In such a case, we may use as the test statistic the ratio of weighted norms of the difference vector \( \hat{\xi} - \xi^0 \) and the predicted residual vector \( \tilde{e} \) as follows:

\[
    w := \frac{(\hat{\xi} - \xi^0)^T A^T P A (\hat{\xi} - \xi^0)}{\sigma_0^2 m} \cdot \frac{\sigma_0^2 (n - m)}{\tilde{e}^T \tilde{P} \tilde{e}} \sim F(m, n - m).
\]  

Here we have assumed that matrix \( A \) has full rank, i.e., \( \text{rk} A = m \). Since the numerator and denominator are statistically independent of one another, the test statistic \( w \) has an \( F \)-distribution with \( m \) and \( n - m \) degrees of freedom, as shown in (10.58). Therefore, our global model-check is made by the following hypothesis test:

\[
    H_0 : E\{\hat{\xi}\} = \xi^0 \text{ vs. } H_A : E\{\hat{\xi}\} \neq \xi^0
\]

If \( w \leq F_{\alpha, m, n - m} \) accept \( H_0 \); else reject \( H_0 \).

We now show that the numerator and denominator of \( w \) are indeed independent, as required for use of the \( F \)-distribution. To do so, we only need to show that

\[
    C\{\tilde{e}^T \tilde{P} \tilde{e}, (\hat{\xi} - \xi)^T (A^T P A)(\hat{\xi} - \xi)\} = 0.
\]  

Note that, without loss of generality, we have replaced \( \xi^0 \) with \( \xi \). From (4.5e) we have \( \tilde{e} = [I_n - AN^{-1} A^T]e \). Therefore,

\[
    \tilde{e}^T \tilde{P} \tilde{e} = e^T [I_n - PAN^{-1} A^T] P [I_n - AN^{-1} A^T] e = e^T [P - PAN^{-1} A^T] e =: e^T M_1 e.
\]  

Also

\[
    A(\hat{\xi} - \xi) = e - \tilde{e} = e - (I_n - AN^{-1} A^T)(A \xi + e) = (AN^{-1} A^T) e \Rightarrow
\]

\[
    (\hat{\xi} - \xi)^T (A^T P A)(\hat{\xi} - \xi) = e^T (PAN^{-1} A^T) P (AN^{-1} A^T) e = e^T (PAN^{-1} A^T) e =: e^T M_2 e.
\]

By substitution of (10.61a) and (10.61d), the condition (10.60) is equivalent to the condition that \( e^T M_1 e \) and \( e^T M_2 e \) are independent, which holds if, and only if,

\[
    e^T M_1 D\{e\} M_2 e = 0,
\]

which is true since

\[
    e^T (P - PAN^{-1} A^T) (\sigma_0^2 P^{-1}) (PAN^{-1} A^T) e = 0.
\]
10.6 Checking an Individual Element (or 2-D or 3-D Point) in the Parameter Vector

We may use an \( l \times m \) matrix \( K \) to select a subset of size \( l \) of the parameter vector for hypothesis testing as follows:

\[
H_0 : E\{K\hat{\xi}\} = K\xi^0 = \kappa_0, \quad (10.62a)
\]
\[
H_A : E\{K\hat{\xi}\} = K\xi^0 \neq \kappa_0. \quad (10.62b)
\]

If \( l = 1 \), \( K \) is a unit row vector that extracts the relevant element from the parameter vector, in which case \( \kappa_0 \) is simply a scalar quantity. The following examples show the matrix \( K \) used for extracting a single element, a 2-D point, and a 3-D point, respectively:

\[
K : = \begin{bmatrix} 0, \ldots, 0, 1, 0, \ldots, 0 \end{bmatrix}, \text{ where } 1 \text{ appears at the } j \text{th element}; \quad (10.63a)
\]
\[
K : = \begin{bmatrix} 0_2, \ldots, 0_2, I_2, 0_2, \ldots, 0_2 \end{bmatrix}, \text{ where } K \text{ is size } 2 \times m; \quad (10.63b)
\]
\[
K : = \begin{bmatrix} 0_3, \ldots, 0_3, I_3, 0_3, \ldots, 0_3 \end{bmatrix}, \text{ where } K \text{ is size } 3 \times m. \quad (10.63c)
\]

For 2-D and 3-D points, the subscripts denote the dimension of the square sub-matrices (zero matrix or identity matrix), and \( I_n \) \( (n \in \{2, 3\}) \) is the \( j \)th sub-matrix, which means it “selects” the \( j \)th point from \( \xi \).

The test statistic is then defined as

\[
w : = \frac{[K(\hat{\xi} - \xi^0)]^T D\{K(\hat{\xi} - \xi^0)\}^{-1}[K(\hat{\xi} - \xi^0)]/\text{rk } K}{1/\sigma_0^2} = \frac{[K\hat{\xi} - \kappa_0]^T [KN^{-1}K^T]^{-1}[K\hat{\xi} - \kappa_0]/l}{\sigma_0^2} =: \frac{R/l}{\Omega/(n - m)}, \quad (10.64a)
\]

where \( \sigma_0^2 \) is assumed to be 1, and thus omitted from (10.64b). Note that since \( \xi^0 \) is a chosen (and therefore non-random) quantity to test against, the dispersion is not affected by the constant shift, i.e.,

\[
D\{K(\hat{\xi} - \xi^0)\} = D\{K\hat{\xi}\} = \sigma_0^2 KN^{-1}K^T. \quad (10.65)
\]

The symbols \( R \) and \( \Omega \) are used for convenience and are analogous to the symbols introduced in Sections 5.1 and 7.2, respectively. They are statistically independent of one another and have the following distributions:

\[
R \sim \chi^2(l), \quad \Omega \sim \chi^2(n - m). \quad (10.66)
\]

Statistical independence between the random variables \( R \) and \( \Omega \) means that their joint pdf is equivalent to the product of their individual pdf’s: \( f(R, \Omega) = f(R) f(\Omega) \). Independence can be shown by following the same line of thought as that used at the end of the previous section, where \( M_1 \) remains unchanged and \( M_2 \) is now \( PAN^{-1}K^T [KN^{-1}K^T]^{-1}K^T A^T P \). Therefore, the test statistic (10.64b) has an \( F \)-distribution represented by

\[
w \sim F(l, n - m). \quad (10.67)
\]
An alternative, more compact, form for $w$ when $l = 1$ is given by

$$w = \frac{(\hat{\xi}_j - (\kappa_0)_j)^2}{\hat{\sigma}^{-2}(N^{-1})_{jj}} \sim F(1, n - m).$$

(10.68)

The decision to accept or reject the null hypothesis is made analogous to (10.59b).

### 10.6.1 Non-central $F$-distribution

If the null hypothesis $H_0$ is false, the test statistic $w$ is said to have a non-central $F$-distribution (denoted here as $F'$), which requires a non-centrality parameter $\theta$ so that $w \sim F'(v_1, v_2, \theta)$ under $H_A$, where $v_1$ and $v_2$ have been used to denote the degrees of freedom, in general. The qualification “under $H_A$” implies that we must pose a specific alternative hypothesis $H_A$ in this case, rather than just the negation of $H_0$. For a one-tailed test, the area under the non-central $F$-distribution curve and to the right of $F_\alpha$ (from the $F$-distribution table) is denoted as $\beta$. The value of $\beta$ is also the probability of making an error of the second kind, namely to accept the null hypothesis $H_0$ when the specified alternative hypothesis $H_A$ is actually true.

The quantity $1 - \beta$ is known as the power of the test. As the value of $\theta$ increases, so does the value $1 - \beta$. Below we have rewritten (10.67) for the non-central case, with the theoretical formula for $2\theta$ following.

$$w \sim F'(l, n - m, \theta)$$

(10.69a)

$$2\theta = (K\hat{\xi} - \kappa_0)^T(K N^{-1} K^T)^{-1}(K\hat{\xi} - \kappa_0)$$

(10.69b)

Note that the non-centrality property is reflected in (10.69b) by including both the true (unknown) vector of parameters $\xi$ and its estimate $\hat{\xi}$ in bilinear form.

### 10.7 Detection of a Single Outlier in the Gauss-Markov Model

A model that expresses the $j$th observation as a potential outlier can be written as

$$y_j = a_j^T \xi^{(j)} + \xi_0 + e_j.$$  

(10.70)

The terms of the model are described as follows

- $y_j$ is the $j$th element of the $n \times 1$ observation vector $y$.
- $a_j$ is an $m \times 1$ column vector that is comprised of the $m$ elements of the $j$th row of matrix $A$ so that $[a_1, a_2, \ldots, a_n]^T := A$.
- $\xi^{(j)}$ denotes the $m \times 1$ parameter vector associated with that set of observations whose $j$th element is considered an outlier, as opposed to $\xi$, which is associated with the same set of observations except that the $j$th one is not considered as an outlier.
$e_j$ is the $j$th element of the unknown random error vector $e$.

$\xi_0^{(j)}$ is a scalar that accounts for an outlier. In other words, it accounts for a non-random error in the observation. The formula for its estimate is developed below.

The following example may be illustrative: Suppose the observation $y_j$ should have been 100 m but only a value of 10 m was recorded, then $\xi_0^{(j)}$ accounts for a 90 m blunder.

A modified GMM whose $j$th observation might be deemed an outlier is expressed as

\[
y_n \times 1 = A_{n \times m} \xi^{(j)} + \eta_j \xi_0^{(j)} + e, \quad \eta_j := \begin{bmatrix} 0, \ldots, 0, 1, 0, \ldots, 0 \end{bmatrix}^T,
\]

where $e \sim N(0, \sigma^2_P P^{-1})$.\hspace{1cm} (10.71a)

Note that the number 1 in $\eta_j$ appears at the $j$th element; all other elements are 0.

We must compare the model in (10.71) with the original GMM (3.1) that does not include an outlier. Since the model (10.71) assumes only one outlier in the data set, $n$ comparisons of the two models are necessary in order to test all $y_i$ ($i = 1, \ldots, n$) observations independently. For each comparison we introduce the constraint equation

\[
\xi_0^{(j)} = K \begin{bmatrix} \xi^{(j)} \\ \xi_0^{(j)} \end{bmatrix} = \kappa_0 = 0.
\]

(10.72)

Here $K := [0, 0, \ldots, 1]$ is of size $1 \times (m + 1)$. When we impose the constraint (10.72) upon the model (10.71), we obtain a model equivalent to the original GMM (3.1) that does not include an additional parameter to model an outlier.

Note: For the remainder of this section, we will assume that the weight matrix $P$ is diagonal: $P = \text{diag}(p_1, \ldots, p_n)$, where $p_i$ is the weight of the $i$th observation. See Schaffrin (1997b) for a treatment of outlier detection with correlated observations.

Now, we begin with the following Lagrange target function to derive a least-squares estimator in the unconstrained model (10.71):

\[
\Phi(\xi^{(j)}, \xi_0^{(j)}) = (y - A\xi^{(j)} - \eta_j \xi_0^{(j)})^T P (y - A\xi^{(j)} - \eta_j \xi_0^{(j)}),
\]

which is made stationary with respect to $\xi^{(j)}$ and $\xi_0^{(j)}$ by setting the first partial derivatives of (10.73) to zero, resulting in the following Euler-Lagrange necessary conditions:

\[
\frac{1}{2} \left[ \frac{\partial \Phi}{\partial \xi^{(j)}} \right]^T = -A^T P y + A^T P \eta_j \xi_0^{(j)} + A^T P A \xi^{(j)} \geq 0, \hspace{1cm} (10.74a)
\]

\[
\frac{1}{2} \frac{\partial \Phi}{\partial \xi_0^{(j)}} = -\eta_j^T P y + \eta_j^T P A \xi^{(j)} + \eta_j^T P \eta_j \xi_0^{(j)} \leq 0. \hspace{1cm} (10.74b)
\]
Of course the second partial derivatives are functions of $P$, which is positive-definite by definition, thereby satisfying the sufficient condition required for obtaining the minimum of (10.73). In matrix form we have

$$
\begin{bmatrix}
N & A^T P \eta_j \\
\eta_j^T P A & \eta_j^T P \eta_j
\end{bmatrix}
\begin{bmatrix}
\hat{\xi}^{(j)} \\
\hat{\xi}_0^{(j)}
\end{bmatrix}
= 
\begin{bmatrix}
c \\
\eta_j^T P y
\end{bmatrix},
$$

(10.75a)

or, because $P$ was assumed to be diagonal,

$$
\begin{bmatrix}
N & a_j p_j \\
p_j a_j^T & p_j
\end{bmatrix}
\begin{bmatrix}
\hat{\xi}^{(j)} \\
\hat{\xi}_0^{(j)}
\end{bmatrix}
= 
\begin{bmatrix}
c \\
p_j y_j
\end{bmatrix}.
$$

(10.75b)

Here, as in previous chapters, we have used the definition $[N, c] := A^T P[y, y]$. Using (A.14) for the inverse of a partitioned matrix, and decomposing the resulting inverse into a sum of two matrices, results in

$$
\begin{bmatrix}
\hat{\xi}^{(j)} \\
\hat{\xi}_0^{(j)}
\end{bmatrix}
= 
\begin{bmatrix}
N^{-1} \begin{bmatrix} c \\ p_j \end{bmatrix} + N^{-1} \begin{bmatrix} a_j p_j \\ -1 \end{bmatrix} (p_j - p_j a_j^T N^{-1} a_j p_j)^{-1} \\
0 & 0 & 0
\end{bmatrix} 
\begin{bmatrix}
\hat{\xi} \\
\hat{\xi}_0
\end{bmatrix},
$$

(10.76a)

or

$$
\begin{bmatrix}
\hat{\xi}^{(j)} \\
\hat{\xi}_0^{(j)}
\end{bmatrix}
= 
\begin{bmatrix}
N^{-1} \begin{bmatrix} c \\ p_j \end{bmatrix} - N^{-1} \begin{bmatrix} a_j p_j \\ -1 \end{bmatrix} (p_j - p_j^2 a_j^T N^{-1} a_j)^{-1} p_j (y_j - a_j^T N^{-1} c)
\end{bmatrix} 
\begin{bmatrix}
\hat{\xi} \\
\hat{\xi}_0
\end{bmatrix}.
$$

(10.76b)

From (10.76b), and recalling that $\hat{\xi} = N^{-1} c$ is based on a data set assumed to have no outliers, we can write the following difference between estimations:

$$
\hat{\xi}^{(j)} - \hat{\xi} = -N^{-1} a_j \left( \frac{y_j - a_j^T \hat{\xi}}{p_j^2 - a_j^T N^{-1} a_j} \right) = -N^{-1} a_j \left( \hat{\xi}_0 \right)_{jj},
$$

(10.77)

where $(Q e)_{jj}$ is the $j$th diagonal element of the cofactor matrix for the residual vector $\hat{e}$. For the estimated non-random error in $y_j$ we have

$$
\hat{\xi}_0^{(j)} = \frac{y_j - a_j^T \hat{\xi}}{1 - p_j a_j^T N^{-1} a_j} = \frac{\hat{\xi}_j}{(Q e)_{jj}} = \frac{\hat{\xi}_j / p_j}{(Q e)_{jj}}.
$$

(10.78)

The hypothesis test for the $j$th observation being an outlier is then written as

$$
H_0 : E\{\hat{\xi}_0^{(j)}\} = 0 \text{ versus } H_A : E\{\hat{\xi}_0^{(j)}\} \neq 0.
$$

(10.79)
The test statistic has an $F$-distribution and is computed by
\[
T_j = \frac{R_j / 1}{(\Omega - R_j) / (n - m - 1)} \sim F(1, n - m - 1). \tag{10.80}
\]

The definition of $R_j$, in terms of $\hat{\xi}_0^{(j)}$, is
\[
R_j := \frac{(\hat{\xi}_0^{(j)} - \hat{\xi}_0)^2}{KN^{-1}_A K^T} = \frac{(\hat{\xi}_0^{(j)} - \hat{\xi}_0)^2}{(p_j - p_j \hat{a}_j^T N^{-1} \hat{a}_j)^{-1} p_j (Q_\hat{\xi} P)_{jj} = \frac{\hat{\epsilon}_j^2}{(Q_\hat{\xi})_{jj}}. \tag{10.81}
\]

It is important to note that the symbols $\hat{\epsilon}$ and $Q_\hat{\xi}$ represent the residual vector and its cofactor matrix, respectively, as predicted within the GMM (3.1) — see (3.9) and (3.14c). As was already mentioned, when we impose the constraint (10.72) on model (10.71b) we reach a solution identical to the LESS within model (3.1). It is also important to understand the terms in the denominator of (10.80). As stated previously, the symbol $R$ is used to account for that portion of the $P$-weighted residual norm due to the constraints. The first parenthetical term in the denominator, $(\Omega - R_j)$, accounts for that part of the norm coming from the unconstrained solution. Here we have used $\Omega := \hat{\epsilon}^T \hat{P} \hat{\epsilon}$, with $\hat{\epsilon}$ belonging to the constrained solution (determined within the model (3.1)). Therefore, we must subtract $R$ from $\Omega$, as it is defined here, to arrive at the portion of the norm coming from the unconstrained LESS computed within model (10.71).

We note again that the equations from (10.75b) to (10.81) hold only in the case of a diagonal weight matrix $P$. Regardless of whether or not $P$ is diagonal, the quantity
\[
r_j := (Q_\hat{\xi} P)_{jj} \tag{10.82}
\]
is the $j$th so-called redundancy number, for the unconstrained solution in this case. The following properties hold for $r_j$:
\[
0 < r_j \leq 1 \text{ for } i = \{1, \ldots, n\} \quad \text{and} \quad \sum_j r_j = n - \text{rk}\ A. \tag{10.83}
\]
Note that $(Q_\hat{\xi} P)_{jj} = (Q_\hat{\xi})_{jj} p_j$ for the case that matrix $P$ is diagonal.

Finally, the matrix $N_1$ in (10.81) is defined as
\[
N_1 = \begin{bmatrix} N & a_j p_j \\ p_j a_j^T & p_j \end{bmatrix}, \tag{10.84}
\]
which appears in (10.75b). Pre- and post-multiplying $N_1^{-1}$ by $K$ extracts only its last diagonal element, which, according to the formula for inverting a partitioned matrix, turns out to be the scalar quantity $(p_j - p_j^2 a_j^T N^{-1} a_j)^{-1}$, also appearing in (10.81).

We comment that outlier detection at the 2-D and 3-D level can also be performed, for example, in testing whether observed 2-D and 3-D points are outliers. The 3-D case is also appropriate for GPS baseline adjustments, and its development is shown by Snow (2002); see also Snow and Schaffrin (2003).
Useful Matrix Relations and Identities

Product of transposes:
\[ A^T B^T = (BA)^T \]  \hspace{1cm} (A.1)

Transpose of inverse:
\[ (A^T)^{-1} = (A^{-1})^T \]  \hspace{1cm} (A.2)

Product of inverses:
\[ A^{-1} B^{-1} = (BA)^{-1} \]  \hspace{1cm} (A.3)

Rank of triple product: Given: \( A(m \times n), B(m \times m), C(n \times n) \):
\[ B, C \text{ nonsingular } \Rightarrow \text{rk}(BAC) = \text{rk}(A) \text{ or } \text{rk}(BA) = \text{rk}(A) \text{ if } C = I \]  \hspace{1cm} (A.4)

Trace invariant with respect to a cyclic permutation of factors: If the product \( ABC \) is square, then the following trace operations are equivalent:
\[ \text{tr}(ABC) = \text{tr}(BCA) = \text{tr}(CAB). \]  \hspace{1cm} (A.5)

Column space and nullspace: The column space of \( A \) is denoted by \( \mathcal{R}(A) \) and is also called the range of \( A \). Its dimension equals the rank of \( A \). The nullspace of \( A \) is denoted by \( \mathcal{N}(A) \) and is also called the kernel of \( A \). The dimension of the nullspace is \( m - \text{rk} A \), where \( m \) is the number of columns of \( A \); the dimension is also called the nullity.

The column space of \( AB \) is contained in the column space of \( A \). \hspace{1cm} (A.6)
Sherman-Morrison-Woodbury-Schur formula:

\[(T - UW^{-1}V)^{-1} = T^{-1} + T^{-1}U(W - VT^{-1}U)^{-1}VT^{-1}\]  \hspace{1cm} (A.7)

As a consequence of (A.7), we also have:

\[(I \pm W^{-1}V)^{-1} = I \mp (W \pm V)^{-1}V, \quad (A.8a)\]
\[(I \pm V)^{-1} = I \mp (I \pm V)^{-1}V, \quad (A.8b)\]
\[(I \pm W^{-1})^{-1} = I \mp (W \pm I)^{-1}. \quad (A.8c)\]

Equations (39–43) of “Useful Matrix Equalities” (handout from Prof. Schaffrin, possibly originating from Urho A. Uotila).

\[DC(A + BDC)^{-1} = (D^{-1} + CA^{-1}B)^{-1}CA^{-1} = \quad (A.9a)\]
\[= D(I + CA^{-1}BD)^{-1}CA^{-1} = \quad (A.9b)\]
\[= DC(I + A^{-1}BDC)^{-1}A^{-1} = \quad (A.9c)\]
\[= DCA^{-1}(I + BDCA^{-1})^{-1} = \quad (A.9d)\]
\[= (I + DCA^{-1}B)^{-1}DCA^{-1} \quad (A.9e)\]

Suppose the matrices \(A\) and \(B\) in (A.9) are identity matrices, then we have

\[DC(I + DC)^{-1} = (D^{-1} + C)^{-1}C = \quad (A.10a)\]
\[= D(I + CD)^{-1}C = \quad (A.10b)\]
\[= (I + DC)^{-1}DC. \quad (A.10c)\]

Definition of idempotent:

The matrix \(P\) is idempotent if \(PP = P\). Projection matrices are idempotent.

\hspace{1cm} \hspace{1cm} (A.11)

If \(P\) is idempotent, \(\text{tr}\ P = \text{rk}\ P\).

\hspace{1cm} \hspace{1cm} (A.12)

Inverse of the partitioned normal equation matrix: Assume the matrix \(N\) is of full rank and is partitioned as follows:

\[N = \begin{bmatrix} N_{11} & N_{12} \\ N_{21} & N_{22} \end{bmatrix}. \quad (A.13)\]

The following steps lead to the inverse of \(N\) expressed in terms of the partitioned
blocks:

\[
\begin{pmatrix}
N_{11} & N_{12} & I & 0 \\
N_{21} & N_{22} & 0 & I
\end{pmatrix} \to \begin{pmatrix}
I & N_{11}^{-1}N_{12} & N_{11}^{-1} & 0 \\
N_{21} & N_{22} & 0 & I
\end{pmatrix} \to
\]

\[
\begin{pmatrix}
I & N_{11}^{-1}N_{12} & N_{11}^{-1} & 0 \\
0 & N_{22} - N_{21}N_{11}^{-1}N_{12} & N_{11}^{-1} & 0
\end{pmatrix} \to
\]

\[
\begin{pmatrix}
I & N_{11}^{-1}N_{12} & 0 \\
0 & I
\end{pmatrix} \to \begin{pmatrix}
I & (N_{22} - N_{21}N_{11}^{-1}N_{12})^{-1}N_{21}N_{11}^{-1} & 0 \\
0 & I
\end{pmatrix} \to
\]

\[
\begin{pmatrix}
I & 0 & (N_{11}^{-1} + N_{11}^{-1}N_{12} \cdot W \cdot N_{21}N_{11}^{-1})^{-1}N_{21}^{-1}N_{12} \cdot W \\
0 & I
\end{pmatrix},
\]

with \(W := (N_{22} - N_{21}N_{11}^{-1}N_{12})^{-1}\). Finally we may write

\[
\begin{pmatrix}
N_{11} & N_{12} \\
N_{21} & N_{22}
\end{pmatrix}^{-1} = \begin{pmatrix}
N_{11}^{-1} + N_{11}^{-1}N_{12} \cdot W \cdot N_{21}N_{11}^{-1} & -N_{11}^{-1}N_{12} \cdot W \\
-W \cdot N_{21}N_{11}^{-1} & W
\end{pmatrix}.
\]

Note that other equivalent representations of this inverse exist. Taking directly from the Useful Matrix Equalities handout mentioned above, we write some additional expressions for the inverse.

\[
\begin{pmatrix}
N_{11} & N_{12} \\
N_{21} & N_{22}
\end{pmatrix}^{-1} = \begin{pmatrix}
Q_{11} & Q_{12} \\
Q_{21} & Q_{22}
\end{pmatrix}
\]

\[
Q_{11} = (N_{11} - N_{12}N_{22}^{-1}N_{21})^{-1} = \]

\[
= N_{11}^{-1} + N_{11}^{-1}N_{12}(N_{22} - N_{21}N_{11}^{-1}N_{12})^{-1}N_{21}N_{11}^{-1} = \]

\[
= N_{11}^{-1} + N_{11}^{-1}N_{12}Q_{22}N_{21}N_{11}^{-1}
\]

\[
Q_{22} = (N_{22} - N_{21}N_{11}^{-1}N_{12})^{-1} = \]

\[
= N_{22}^{-1} + N_{22}^{-1}N_{21}(N_{11} - N_{12}N_{22}^{-1}N_{21})^{-1}N_{12}N_{22}^{-1} =
\]

121
\[ N_{22}^{-1} + N_{21}^{-1} N_{11} N_{12} N_{22}^{-1} \]  
(A.17c)

\[ Q_{12} = -(N_{11} - N_{12} N_{22}^{-1} N_{21})^{-1} N_{12} N_{22}^{-1} = -Q_{11} N_{12} N_{22}^{-1} = \]  
(A.18a)

\[ = -N_{11}^{-1} N_{12} (N_{22} - N_{21} N_{11}^{-1} N_{12})^{-1} = -N_{11}^{-1} N_{12} Q_{22} \]  
(A.18b)

\[ Q_{21} = -N_{22}^{-1} N_{21} (N_{11} - N_{12} N_{22}^{-1} N_{21})^{-1} = -N_{22}^{-1} N_{21} Q_{11} = \]  
(A.19a)

\[ = -(N_{22} - N_{21} N_{11}^{-1} N_{12})^{-1} N_{21} N_{11}^{-1} = -Q_{22} N_{21} N_{11}^{-1} \]  
(A.19b)

In the case that \( N_{22} = 0 \), we have:

\[ Q_{22} = -(N_{21} N_{11}^{-1} N_{12})^{-1} \]  
(A.20a)

\[ Q_{11} = N_{11}^{-1} + N_{11}^{-1} N_{12} Q_{22} N_{21} N_{11}^{-1} \]  
(A.20b)

\[ Q_{12} = -N_{11}^{-1} N_{12} Q_{22} \]  
(A.20c)

\[ Q_{21} = -Q_{22} N_{21} N_{11}^{-1} \]  
(A.20d)

Schur Complement: the parenthetical term \((N_{22} - N_{21} N_{11}^{-1} N_{12})\) shown above is called the Schur Complement of \(N_{11}\).

Derivative of quadratic form:
While some authors write the derivative of a quadratic form (a scalar-valued vector function) with respect to a column vector as a row vector, we write such a derivative as a column vector. This is in agreement with the following authors: Grafarend and Schaffrin (1993); Harville (2000, pg. 295); Koch (1999, pg. 69); Lütkepohl (1996, pg. 175); Strang and Borre (1997, pg. 300). For example, given \( x \in \mathbb{R}^n \) and \( Q \in \mathbb{R}^{n \times n} \), we have

\[ \Phi(x) = x^T Q x \Rightarrow \frac{\partial \Phi}{\partial x} = 2Qx. \]  
(A.21)
Bibliography


