LEAST SQUARES ADJUSTMENT
A MODERN APPROACH

by

PETER MEISSL

MITTEILUNGEN
der geodätischen Institute der Technischen Universität Graz
Folge 43

Graz, 1982

Herausgeber:
Geodätische Institute der Technischen Universität Graz

Redaktion für diese Folge:
Abteilung für Mathematische Geodäsie und Geoinformatik
des Institutes für Theoretische Geodäsie

Mit freundlicher Genehmigung der Geodätischen Institute der Technischen Universität Graz wurde diese Folge am Institut für Geodäsie und Geoinformation der Universität Bonn eingescannt.

Druck und Herstellung:
Druck- und Kopierzentrum der Technischen Universität Graz

Adresse:
Technische Universität Graz
Rechbauerstraße 12
A-8010 Graz, Österreich.

Preface

For his lectures at the Tongji University in Shanghai and at other institutions in China in November - December 1981, Peter Meissl prepared a set of lecture notes on contemporary least-squares adjustment and applications. Subsequently he worked on correcting and expanding them, but this was interrupted by his tragic death on May 22, 1982. (For Peter Meissl's life and work, the reader is referred to his biography by Franz Allmer, Mitteilungen der geodätischen Institute der Technischen Universität Graz, Folge 44, 1983.)

In view of the unique importance of this work, the Institute of Theoretical Geodesy decided to edit the manuscript posthumously and to publish the book in the series of the Geodetic Institutes of the Technical University, Graz, although Peter Meissl himself would certainly have included additional topics such as inner adjustment theory, expanded others such as the theory of large networks, and polished the manuscript much more before being satisfied with its publication.

The finishing of the book is due to Peter Meissl's closest associates: Dr. Norbert Bartelme, Dr. Helmut Fuchs, Dr. Bernhard Hofmann-Wellenhof, Dipl.-Ing. Wolf-Dieter Schuh and Dipl.-Ing. Manfred Wieser. In addition to being responsible for the careful editing of the manuscript, they also prepared the printing text using the word processing facilities of the computer WANG 2200 MVP.

A glance at the table of contents shows that this book is a thoroughly modern text on least-squares adjustment. In the contemporary spirit, the usual linear algebra is treated in the context of general linear spaces, which makes possible an easy transition to Hilbert space important for advanced topics. Also modern is the division into an algebraic and geometric approach (without statistics) and a stochastic approach, including statistical tests. Applications to Doppler observations, large networks, geodetic data bases, and splines essentially increase the practical usefulness. Although the book develops adjustment theory in a systematic and self-contained way, it will be best appreciated by readers who already have some elementary previous knowledge of adjustment computations.

The book needs no recommendation. Both students and research workers will find it indispensable. It is a fitting memorial of a great scientist.

Helmut Moritz

TABLE OF CONTENTS

A.5. Projectors

1. Decomposition of a vector space into a direct sum of subspaces
2. Orthocomplementary subspaces
3. Theorem by Pythagoras
4. Matrix representation of orthogonal projectors
5. Projections of functionals

A.6. Least squares adjustment

1. Projecting the vector of observations
2. Inhomogeneous form of least squares adjustment
3. Fundamental rectangular triangle of least squares adjustment
4. Least squares adjustment by projecting functionals

A.7. Partitioned matrices

1. Definitions
2. Computational rules
3. Block diagonality
4. Block Gauss elimination
5. Theoretical background of partitioned matrices

A.8. Isometric mappings between inner product spaces

1. Definitions
2. Preservation of inner products
3. Matrix representation
4. Examples of isometric mappings
5. Canonical transformation of an adjustment problem

A.9. Partial reduction

1. Partitioning the set of parameters
2. Partial reduction of the normal equations
3. Orthogonal decomposition of the parameter space
4. Partially reduced observation equations
5. Alternative derivation of the partially reduced observation equations

A.10. Adjustment phased with respect to observations

1. Formulation of the problem
2. Addition of normal equations
3. Updating the solution of the previous phase
4. Geometrical insight
5. Pre-elimination of group-internal unknowns
6. Helmert blocking

A.11. Complementary extremum principles in least squares adjustment

1. Basic geometric principle
2. Reformulation for linear manifolds
3. Adjustment by minimizing the norm of the residuals
4. Adjustment by minimizing variances

A.12. Generalized inverses

1. Range space and null space of a linear operator
2. g-inverse
3. Reflexive generalized inverse
4. Generalized inverse with least squares property
5. Generalized inverse with minimum norm property
6. Minimum norm least squares inverse
7. Pseudo inverse

A.13. Adjustment of rank-deficient systems

1. Formulation of the problem
2. Solution via generalized inverse of A
3. A minimum norm property of the covariance matrix of adjusted parameters
4. Solution via singular normal equations
5. Calculation of the ℓm-inverse
6. Application to free network adjustment

B. STOCHASTIC APPROACH TOWARD LEAST SQUARES ADJUSTMENT

B.1. Probabilities

1. Relative frequencies
2. Probability space
3. Examples
4. Calculus of probabilities

B.2. Random variables

1. One-dimensional random variables
2. Probability density function
3. n-dimensional random variables
4. Functions of random variables
5. Marginal distribution
6. Stochastic independence

B.3. Expectation, variances and covariances

1. Expectation of a one-dimensional random variable
2. Variance of a one-dimensional random variable
3. Various kinds of observation errors
4. Simple computational rules for expectation and variance
5. The case of higher dimensional random variables
6. Covariance matrix
7. Propagation of expectations and covariances
8. Important special cases
9. Zero correlation and stochastic independence

B.4. Gauss-Markoff model of least squares adjustment

1. Stochastic model
2. Unbiased estimates
3. Best linear unbiased estimation
4. Error calculus

B.5. Applications of the error propagation law

1. Triangle with three measured sides
2. First fundamental problem in the plane
3. Error ellipses
4. Polar survey with redundancy
5. Area calculated from polar survey
6. Conventionally adjusted regular traverse
7. Rigorously adjusted regular traverse
8. Systematic errors in a regular traverse

C. CONFIDENCE REGIONS AND TESTS OF LINEAR HYPOTHESES

C.1. Probability distributions used in statistical tests

1. One-dimensional Gauss distribution (normal distribution)
2. Multi-dimensional Gauss distribution (normal distribution)
3. Chi-squared distribution (χ^2-distribution)
4. Student's distribution (t-distribution)
5. Fisher's distribution (F-distribution)

C.2. Canonical transformation

1. Preliminaries
2. Making the functionals a part of the parameters
3. Orthogonal decomposition of the space L_A
4. Orthogonal decomposition of L into L_A and L_B
5. Orthonormalizing the bases of the subspaces

C.3. Distribution of various quantities resulting from least squares adjustment

1. Joint distribution of BLUE's and residuals
2. Distribution of the "weighted sum of residuals"
3. Expressions in $\Phi\tilde{x}$ and v having χ^2- or t-distribution
4. Expressions in $\tilde{x}$ and v having t-distribution

C.4. Confidence regions

1. Confidence intervals for one-dimensional Gauss variables
2. Application to the Gauss-Markoff model with known unit weight error
3. Studentization
4. Confidence regions for σ^2
5. Ellipsoidal confidence regions for sets of linear estimates

C.5. Tests of linear hypotheses

1. Linear hypotheses
2. Tests of variances
3. A simple example
4. A sophisticated example

D. SPECIAL TOPICS

D.1. Adjustment of Doppler observations

1. Transit system
2. Observing a difference in light travel time
3. Frequency shift
4. Technique of cycle counting
5. Parameters accounting for receiver imperfections
6. Transformations into an earth-fixed frame
7. Parameters accounting for orbit corrections
8. Linearization of the observation equations
9. Single station adjustment
10. Multi-station adjustment

D.2. Geodetic data bases

1. Storage media
2. Requirements for geodetic data bases
3. The data base of NGS

D.3. Cholesky's algorithm applied to normal equations of geodetic networks

1. Cholesky's algorithm for a general symmetric positive definite system
2. Partial reduction by Cholesky's algorithm
3. Geodetic normal equations
4. Geodetic interpretation of the partially Cholesky-reduced system
5. Problem of station ordering

D.4. One-dimensional cubic spline interpolation

1. Introduction
2. Parameterizing a cubic polynomial
3. Conditions at the inner nodes
4. Boundary conditions
5. Tridiagonal linear system
6. Modification of the periodic case
7. Interpolation of curves in the plane
8. Splines viewed as a vector space
9. The locality of splines

D.5. Two-dimensional spline interpolation

1. Introduction
2. Bicubic polynomials
3. Hermite bicubic interpolation
4. Bicubic splines

D.6. Geometry of exact spline interpolation

1. Formulation of the problem
2. Definition of splines
3. Existence and uniqueness of splines
4. Minimum properties of splines
5. Other examples
6. Prediction as a special case of spline interpolation
7. Noise-free collocation with trend parameters

D.7. Approximation with splines

1. Introduction
2. Approximation in one dimension
3. Basis splines with local support
4. Two dimensions

A. THE ALGEBRAIC AND GEOMETRIC APPROACH TOWARD LEAST SQUARES ADJUSTMENT

1. Vector spaces.

1.1. Definition.

A real vector space (also called real linear space) is a set of elements, called vectors, having the following properties. If $a_1,\ldots,a_m$ are vectors of the vector space V, and if $\lambda_1,\ldots,\lambda_m$ are real numbers, then the linear combination

$$\lambda_1 a_1 + \ldots + \lambda_m a_m$$

must be defined and must be an element of V.

Remark. The above definition is logically not complete. A set of familiar computational rules must be postulated: $\lambda(a+b) = \lambda a + \lambda b$, $(\lambda+\mu)a = \lambda a + \mu a$, $\lambda(\mu a) = (\lambda\mu)a$, $1a = a$. By the way, the expression λa may equally well be written as $a\lambda$.

It is seen that in a vector space essentially two mathematical operations are available, multiplication of a vector by a scalar, and addition of two vectors. The neutral element of scalar multiplication is the real number 1. The neutral element of addition is the zero vector. It is obtained either as $0a$ or as $a-a$.

Remark on notation. In the first sections we shall consistently use upper case Latin letters for vector spaces, lower case Latin letters for vectors, and lower

case Greek letters for scalars. In later chapters the rather sparse notational resources of the western world must be allocated differently.

1.2. Examples of vector spaces.

1.2.1. R, the real line is a vector space.

1.2.2. R^n, the set of n-tuples

$$a = (\alpha_1, \ldots, \alpha_n)$$

forms a vector space. The real numbers α_i are called components. Scalar multiplication and addition are defined component-wise in an obvious and familiar way.

1.2.3. The set of all linear forms

$$a = \alpha_1 \xi_1 + \ldots + \alpha_n \xi_n$$

in n variables $\xi_1, \ldots, \xi_n$, is a vector space.

1.2.4. The set of all polynomials

$$a = \alpha_0 + \alpha_1 \xi + \alpha_2 \xi^2 \ldots + \alpha_n \xi^n$$

in one variable ξ forms a vector space. The α_i are called coefficients. Note

that multiplication of two polynomials does not correspond to a characteristic structural property of a vector space. It is an additional feature of spaces of polynomials which is exploited in polynomial algebra.

1.2.5. The set of continuous functions $f(\xi)$ defined on an interval $\alpha \leq \xi \leq \beta$ is a vector space. The scalar multiple of a continuous function is continuous; so is the sum of two continuous functions.

1.2.6. The set of all solutions to a linear homogeneous system

$$\begin{aligned}
\alpha_{11}\xi_1 + \ldots + \alpha_{1m}\xi_m &= 0 \\
\alpha_{21}\xi_1 + \ldots + \alpha_{2m}\xi_m &= 0 \\
&\ldots\ldots\ldots\ldots\ldots \\
\alpha_{n1}\xi_1 + \ldots + \alpha_{nm}\xi_m &= 0
\end{aligned}$$

1.2.7. A subset U of a vector space V may be a vector space by itself. Such a subset U is called a (vector-) subspace of V. An example is the set of all polynomials of degree $\leq n$. This set is a subspace of the space of all polynomials introduced in 1.2.4. In turn, the space of all polynomials may be seen as a subspace of the space of continuous functions.

1.2.8. If V is a vector space, and if $a_1, \ldots, a_m$ are vectors in V, then the set U of all linear combinations

$$\lambda_1 a_1 + \ldots + \lambda_m a_m$$

is a vector space. It is called the linear span of $a_1,\ldots,a_m$. It is a subspace of V. In symbols

$$U = \text{span}(a_1,\ldots,a_m)$$

It will be important to investigate conditions under which U=V.

1.2.9 Translations in the plane. Consider the two-dimensional plane as the set of its points. The points are not vectors. There is no meaningful way to define e.g. the sum of two points. Thus, the plane considered as the set of its points is not a vector space. Consider now a common translation of all points in the plane. All points move the same distance, and in the same direction along parallel lines. Such a translation is represented by an arrow. An arrow has a direction and a length. If we want to know the image of a point P under the translation we place the tail of the arrow at P and its tip will show the new position P'. The translations form a vector space. A translation can be multiplied by a scalar in an obvious way; two translations can be added according to the familiar parallelogram-rule. Thus translations can be linearly combined. See figure 1.1.

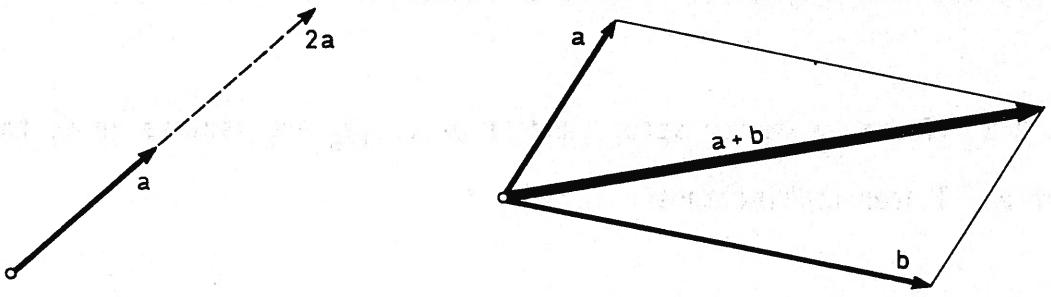


Fig. 1.1 Vectors as arrows

If one chooses an arbitrary point O as a reference point, then any point P in the plane is uniquely identified by the vector translating O into P. This vector is called position vector. In this way the plane is mapped onto the vector space (of translations). In sloppy, but convenient language, one then identifies the plane with this vector space. One should, however, bear in mind that this identification relies on the arbitrary choice of the reference point O.

1.3. Linear dependence and independence.

The vectors $a_1,\ldots,a_n$, are called linearly dependent if there exist scalars $\lambda_1,\ldots,\lambda_n$, not all equal to zero, such that

$$\lambda_1 a_1 + \ldots + \lambda_n a_n = 0$$

Linear dependence of vectors means that the zero vector can be obtained by a nontrivial linear combination.

Vectors $a_1,\ldots,a_n$ which are not linearly dependent are called linearly independent. The zero vector can only be obtained by the trivial linear combination. Such a linear combination has all λ's equal to zero.

It is seen that the vectors $a_1,\ldots,a_n$ are linearly independent if

$$\lambda_1 a_1 + \ldots + \lambda_n a_n = 0$$

implies

$$\lambda_1 = \ldots = \lambda_n = 0$$

If a vector b is a linear combination of vectors $a_1,\ldots,a_m$:

$$b = \lambda_1 a_1 + \ldots + \lambda_m a_m$$

then b is called linearly dependent of $a_1,\ldots,a_m$. The vector b is then a member of span$(a_1,\ldots,a_m)$. The m+1 vectors $a_1,\ldots,a_m,b$ are necessarily linearly dependent.

1.4. Bases.

1.4.1. Definition. A set of linearly independent vectors $e_1,\ldots,e_n$, is called a basis of V if any vector x in V can be expressed as a linear combination

$$x = \xi_1 e_1 + \ldots + \xi_n e_n$$

The numbers $\xi_1,\ldots,\xi_n$ are called coordinates of the vector x with respect to the basis $e_1,\ldots,e_n$. It follows that V is spanned by the linearly independent vectors $e_1,..,e_n$:

$$V = \text{span}(e_1,\ldots,e_n)$$

The coordinates of a vector with respect to a basis are unique. For suppose that

$$x = \xi_1 e_1 + \ldots + \xi_n e_n$$
$$x = \xi'_1 e_1 + \ldots + \xi'_n e_n$$

By subtracting the two equations a linear combination yielding the zero vector is obtained. The coefficients of this linear combination are $(\xi_i - \xi'_i)$, $i=1,\ldots,n$. From the linear independence of the basis vectors one infers that $\xi_i = \xi'_i$, $i=1,\ldots,n$.

1.4.2. Finite dimensional vector spaces. A vector space having a finite basis is called finite dimensional. The choice of a basis is not unique, however the number of vectors in a basis is unique. It is called the dimension of V. The proof of the uniqueness of the dimension is not entirely trivial. If two bases $e_1,\ldots,e_n$ and $e'_1,\ldots,e'_m$ are given in V, and if $n \leq m$ is assumed, one can successively exchange unprimed vectors against primed vectors until a basis of n primed vectors is obtained. The details of the proof are omitted.

1.4.3. Examples of bases. (Confer section 1.2 on examples of vector spaces.)

1.4.3.1. Any nonzero number of R forms a basis of R. If the vector 1 is chosen as basis, any vector has a coordinate equal to itself. Hence 1 is called the natural basis of R.

1.4.3.2. The vector space R^n of n-tuples introduced in 1.2.2 also possesses a

natural basis. It is given by

$$e_1 = (1,0,\ldots,0)$$
$$e_2 = (0,1,\ldots,0)$$
$$\ldots\ldots\ldots\ldots$$
$$e_n = (0,0,\ldots,1)$$

The coordinates α_i of an n-tuple a are then identical to the components of a.

1.4.3.3. The polynomials $1,x,x^2,\ldots,x^n$ form a natural basis of the space of polynomials of degree $\leq n$. The coordinates of a polynomial are then equal to its coefficients.

1.4.3.4. The space of all polynomials and the space of continuous functions over $\alpha \leq \xi \leq \beta$ do not have a finite bases. These spaces are infinite dimensional.

1.4.3.5. Two arrows having neither the same nor opposite directions represent a basis for the arrows (translations) in the plane.

1.4.4. Isomorphism between all vector spaces of dimension n. If V_n is a general n-dimensional vector space, and if a basis $e_1,\ldots,e_n$ is chosen, a correspondence between V_n and R^n is established. Remember that coordinates are unique. Hence any vector in V_n is uniquely mapped onto an n-tuple in R^n. The converse is trivially also true. The basis vectors of V_n are mapped onto the natural basis vectors of R^n. The mapping between V_n and R^n preserves the linear structure:

Linear combinations are mapped onto linear combinations with identical scalar coefficients. If $x,y \in V$ have coordinates ξ_i, η_i, $i=1,\ldots,n$, then $\lambda x+\mu y$ has coordinates $\lambda\xi_i+\mu\eta_i$. In view of the preservation of the linear structure, the mapping is called an isomorphism.

It is seen that all n-dimensional vector spaces are isomorphic to R^n. It suffices to study the structure of R^n in order to learn everything about finite dimensional vector spaces.

Remark. The correspondence between V_n and R^n depends on the choice of a basis $e_1,\ldots,e_n$ in V_n. A different basis leads to a different mapping. There are as many different isomorphic mappings between V_n and R^n as there are bases in V_n!

1.5. Linear equations.

The question whether a vector $b \in R^n$ is a linear combination of vectors $a_1,\ldots,a_m$ out of R^n leads to a system of n linear equations in m unknowns. The question is whether there are scalars $\xi_1,\ldots,\xi_m$ such that

$$a_1\xi_1 + \ldots + a_m\xi_m = b$$

(We prefer now to write the scalar factors ξ_i to the right of the vectors a_i.)

Let a_j, $j=1,\ldots,m$, and b be represented in terms of coordinates with respect to the natural basis:

$$a_j = \begin{bmatrix} \alpha_{1j} \\ \alpha_{2j} \\ \dots \\ \dots \\ \alpha_{nj} \end{bmatrix} \qquad b = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \dots \\ \dots \\ \beta_n \end{bmatrix}$$

Then the following set of equations must hold:

$$\begin{aligned} \alpha_{11}\xi_1 + \dots + \alpha_{1m}\xi_m &= \beta_1 \\ \alpha_{21}\xi_1 + \dots + \alpha_{2m}\xi_m &= \beta_2 \\ \dots\dots\dots\dots\dots\dots & \\ \dots\dots\dots\dots\dots\dots & \\ \alpha_{n1}\xi_1 + \dots + \alpha_{nm}\xi_m &= \beta_n \end{aligned}$$

If b equals the zero vector, the system is called homogeneous. Otherwise it is called inhomogeneous. If a homogeneous system has only the zero solution $\xi_j=0$, $j=1,\dots,m$, then the vectors a_j are linearly independent.

A linear system may also be viewed as a system of equations for forms: Find values for the unknowns ξ_j such that the forms

$$\alpha_{i1}\xi_1 + \dots + \alpha_{im}\xi_m, \quad i=1,\dots,n$$

evaluated for these ξ_j give the numbers β_i. Forms can be viewed as vectors. The above forms are represented by vectors

$$a^i = (\alpha_{i1},\ldots,\alpha_{im})$$

Also an equation

$$\alpha_{i1}\xi_1 + \ldots + \alpha_{im}\xi_m = \beta_i$$

can be put in correspondence with a vector, namely with the m+1 dimensional vector

$$(a^i,\beta_i) = (\alpha_{i1},\ldots,\alpha_{im},\beta_i)$$

One may start to form linear combinations of these vectors which result in very simple vectors (equations). This is the idea behind the familiar elimination procedures. The final stage of the Gauss-Jordan elimination procedure looks as follows.

$$
\begin{aligned}
\xi_1 \qquad & + \alpha^r_{1,r+1}\xi_{r+1} + \ldots + \alpha^r_{1,m}\xi_m = \beta^r_1 \\
\xi_2 \qquad & + \alpha^r_{2,r+1}\xi_{r+1} + \ldots + \alpha^r_{2,m}\xi_m = \beta^r_2 \\
& \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\
& \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\
\xi_r \; & + \alpha^r_{r,r+1}\xi_{r+1} + \ldots + \alpha^r_{r,m}\xi_m = \beta^r_r \\
& 0 = \beta^r_{r+1} \\
& 0 = 0 \\
& \ldots\ldots \\
& \ldots\ldots \\
& 0 = 0
\end{aligned}
$$

The vectors (equations) of the final stage are linear combinations of those of the initial system. However the converse is also true because any step during the Gauss-Jordan algorithm is reversible. Hence the equations of the initial and the last stage span the same space. The systems of equations are equivalent.

Remark. It may not always be possible to obtain a final stage of equations in which the first r unknowns $\xi_1,\ldots,\xi_r$ are isolated as shown above. A reordering of equations and/or unknowns may be necessary in order to ensure the validity of the above final system.

We introduce the matrix of the homogeneous system. It is a rectangular array A of elements α_{ij}

$$A = \begin{bmatrix} \alpha_{11} & \alpha_{12} & \dots & \dots & \alpha_{1m} \\ \alpha_{21} & \alpha_{22} & \dots & \dots & \alpha_{2m} \\ \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots \\ \alpha_{n1} & \alpha_{n2} & \dots & \dots & \alpha_{nm} \end{bmatrix}$$

We also introduce the "augmented" matrix of the inhomogeneous system

$$(A,b) = \begin{bmatrix} \alpha_{11} & \alpha_{12} & \dots & \alpha_{1m} & \beta_1 \\ \alpha_{21} & \alpha_{22} & \dots & \alpha_{2m} & \beta_2 \\ \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots \\ \alpha_{n1} & \alpha_{n2} & \dots & \alpha_{nm} & \beta_n \end{bmatrix}$$

A matrix can be seen as a collection of "row vectors", and, alternatively, as a collection of "column vectors". We shall shortly talk of "rows" and "columns" of a matrix. The following facts are easily deduced from the structure of the GaussJordan reduced system.

(1) The matrix A has r linearly independent rows.

(2) The matrix A has r linearly independent columns. It is seen that the number of linearly independent rows and that of linearly independent columns coincide. This number is called the rank of A. It is denoted

$$r = \text{rank}(A)$$

(3) The solutions $x=(\xi_1,\ldots,\xi_m)$ of the homogeneous system form an $m - r$ dimensional vector subspace of V_m. A basis is provided by the columns of the following matrix.

$$\begin{bmatrix} \alpha^r_{1,r+1} & \alpha^r_{1,r+2} & \cdots & \alpha^r_{1,m} \\ \alpha^r_{2,r+1} & \alpha^r_{2,r+2} & \cdots & \alpha^r_{2,m} \\ \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ \alpha^r_{r,r+1} & \alpha^r_{r,r+2} & \cdots & \alpha^r_{r,m} \\ -1 & 0 & \cdots & 0 \\ 0 & -1 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & -1 \end{bmatrix}$$

(4) If

$$\beta^r_{r+1} = 0$$

then the rank of the augmented matrix equals

$$\text{rank}(A,b) = \text{rank}(A) = r$$

In this case the linear system is consistent. It has a solution. A particular solution is provided by

$$\begin{aligned}
\xi_1 &= \beta_1^r \\
\xi_2 &= \beta_2^r \\
&\cdots\cdots\cdots \\
&\cdots\cdots\cdots \\
\xi_r &= \beta_r^r \\
\xi_{r+1} &= 0 \\
\xi_{r+2} &= 0 \\
&\cdots\cdots\cdots \\
&\cdots\cdots\cdots \\
\xi_n &= 0
\end{aligned}$$

(5) If $\beta_{r+1}^r \neq 0$, then

$$\text{rank}(A,b) = \text{rank}(A) + 1 = r+1$$

In this case the (inhomogeneous) system is inconsistent. It has no solution.

(6) The general solution of a consistent inhomogeneous system is obtained as the

sum of the general solution of the homogeneous system and a particular solution of the inhomogeneous system.

(7) If $m = r$, the solution is unique if it exists.

(8) If $m = n = r$ the solution always exists and is unique. A is an n by n matrix of rank n. Such a matrix is called regular.

2. Linear operators.

2.1. Definitions.

A linear operator Λ is a mapping between vector spaces V and W. It maps any $x \in V$ uniquely onto a $y \in W$. In symbols:

$$y = \Lambda(x)$$

Not every vector $y \in W$ must be the image of a vector $x \in V$. Therefore one says that Λ maps V into W. Should the images of all vectors x in V really cover the whole space W, and should one wish to emphasize this fact, one says that Λ maps V onto W.

The images of two different vectors $x_1, x_2 \in V$ may coincide in W. Thus the "pre-images" of a vector $y \in W$ need not be unique in V. An important subclass of linear operators will have unique pre-images.

The fundamental property of linearity of the operator Λ is expressed by the following equation:

$$\Lambda(\lambda_1 x_1 + \lambda_2 x_2) = \lambda_1 \Lambda(x_1) + \lambda_2 \Lambda(x_2)$$

Thus a linear operator maps a linear combination of vectors onto the linear combination of the individual image vectors in a way that the scalars are preserved.

2.2. Examples of linear operators.

2.2.1. The linear equations

$$\eta_1 = \alpha_{11}\xi_1 + \ldots + \alpha_{1m}\xi_m$$
$$\eta_2 = \alpha_{21}\xi_1 + \ldots + \alpha_{2m}\xi_m$$
$$\ldots\ldots\ldots\ldots\ldots$$
$$\eta_n = \alpha_{n1}\xi_1 + \ldots + \alpha_{nm}\xi_m$$

define a linear operator mapping R^m into R^n. Thus we have obtained another important interpretation of a linear system of equations.

2.2.2. Taking the derivative of a polynomial defines a linear operator mapping the space of polynomials onto itself. The subspace of polynomials of degree $\leq n$ is mapped into itself. The space of images is that one of polynomials having degree $\leq n-1$.

2.2.3. Interpolating a continuous function at $n+1$ distinct locations $\xi_0, \xi_1, \ldots, \xi_n$ by a polynomial of degree $\leq n$, defines a linear operator from the space of continuous functions onto the $n+1$ dimensional space of polynomials of degree $\leq n$.

2.2.4. A linear operator mapping a vector space V into R, the set of real numbers, is called a linear functional. The zero functional assigns zero to any vector out of V. All other linear functionals map V onto R. Examples of linear functionals follow.

2.2.5. A linear form

$$\alpha(x) = \alpha_1\xi_1 + \ldots + \alpha_m\xi_m$$

is a linear functional defined on R^m.

2.2.6. Evaluating a continuous function at a fixed location ξ defines a linear functional on the space of continuous functions.

2.3. Matrix representation of linear operators.

Let Λ be a linear operator mapping the m-dimensional vector space V_m into the n-dimensional space V_n. Choose a basis $e_1,\ldots,e_m$ in V_m and a basis $f_1,\ldots,f_n$ in V_n. Represent

$$x = \sum_{j=1}^{m} \xi_j e_j$$

$$y = \sum_{j=1}^{n} \eta_j f_j$$

The image $\Lambda(e_j)$ of the basis vector e_j is a vector in V_n. Let its representation in terms of the basis $f_1,\ldots,f_n$ be

$$\Lambda(e_j) = \sum_{i=1}^{n} \alpha_{ij} f_i, \quad j=1,\ldots,m$$

We now expand

$$y = \Lambda(x) = \Lambda\left\{ \sum_{j=1}^{m} \xi_j e_j \right\} = \sum_{j=1}^{m} \xi_j \Lambda(e_j) =$$

$$= \sum_{j=1}^{m} \xi_j \sum_{i=1}^{n} \alpha_{ij} f_i = \sum_{i=1}^{n} \left\{ \sum_{j=1}^{m} \alpha_{ij} \xi_j \right\} f_i$$

Comparing with

$$y = \sum_{i=1}^{n} \eta_i f_i$$

and recalling the uniqueness of coordinates, we obtain

$$\eta_i = \sum_{j=1}^{m} \alpha_{ij} \xi_j$$

We see that the linear operator Λ is represented by a matrix

$$A = \begin{bmatrix} \alpha_{11} & \alpha_{12} & \dots & \dots & \alpha_{1m} \\ \alpha_{21} & \alpha_{22} & \dots & \dots & \alpha_{2m} \\ \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots \\ \alpha_{n1} & \alpha_{n2} & \dots & \dots & \alpha_{nm} \end{bmatrix}$$

The matrix representation relies on the chosen basis. A different basis leads to a different matrix representation.

2.4. Composition of mappings, matrix product.

Let

$$y = M(x)$$

$$z = \Lambda(y)$$

be two mappings. M maps V_m into V_n, and Λ maps V_n into V_p. The composite mapping $N = \Lambda \circ M$ is defined as

$$z = N(x) = \Lambda \circ M\,(x) = \Lambda(M(x))$$

Let

$$A = \begin{bmatrix} \alpha_{11} & \cdots & \alpha_{1n} \\ \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots \\ \alpha_{p1} & \cdots & \alpha_{pn} \end{bmatrix}$$

and

$$B = \begin{bmatrix} \beta_{11} & \cdots & \beta_{1m} \\ \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots \\ \beta_{n1} & \cdots & \beta_{nm} \end{bmatrix}$$

be the matrix representation of Λ and M, respectively. We are going to find the matrix representation of $N = \Lambda \circ M$, denoted

$$C = \begin{bmatrix} \gamma_{11} & \cdots & \cdots & \gamma_{1m} \\ \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ \gamma_{p1} & \cdots & \cdots & \gamma_{pm} \end{bmatrix}$$

Substituting

$$\eta_k = \sum_{j=1}^{m} \beta_{kj}\xi_j$$

into

$$\varsigma_i = \sum_{k=1}^{n} \alpha_{ik}\eta_k$$

one obtains

$$\varsigma_i = \sum_{k=1}^{n} \alpha_{ik} \sum_{j=1}^{m} \beta_{kj}\xi_j = \sum_{j=1}^{m} \left\{ \sum_{k=1}^{n} \alpha_{ik}\beta_{kj} \right\} \xi_j$$

It follows that

$$\gamma_{ij} = \sum_{k=1}^{n} \alpha_{ik}\beta_{kj}$$

This leads to the definition of the matrix product

$$C = AB$$

The matrix product is associative, i.e.

$$A(BC) = (AB)C = ABC$$

Associativity follows immediately from the associativity of mappings. It may also be proved directly.

2.5. Inverse operator, inverse matrix.

Let Λ be a linear operator mapping V_n onto W_n. Let $A=(\alpha_{ij})$ be the matrix representation of Λ. It follows that the linear system

$$\eta_i = \sum_{j=1}^{n} \alpha_{ij}\xi_j, \quad i=1,\ldots,n$$

has a solution for any choice of η_i, $i=1,\ldots,n$. From the theory of linear equations it follows that rank(A)=n, and that the solution is necessarily unique. Hence we obtain a mapping Λ^{-1} mapping W_n back onto V_n. The mapping is necessarily linear. Let A^{-1} be its matrix representation. We call Λ^{-1} the inverse operator of Λ, and A^{-1} the inverse matrix of A. It follows that

$$\Lambda^{-1}\circ\Lambda = I$$
$$A^{-1}A = I$$

Here I denotes in the first case the identity operator mapping V_n identically onto itself: $x = I(x)$. In the second case I denotes the matrix representation of the identity operator. We have

$$I = \begin{bmatrix} 1 & 0 & \dots & \dots & 0 & 0 \\ 0 & 1 & \dots & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \\ 0 & 0 & \dots & \dots & 1 & 0 \\ 0 & 0 & \dots & \dots & 0 & 1 \end{bmatrix}$$

The inverse operator of the inverse Λ^{-1} is Λ again:

$$\Lambda \circ \Lambda^{-1} = I$$
$$A A^{-1} = I$$

The inverse matrix A^{-1} may be calculated by the Gauss-Jordan procedure. The procedure must be carried out for a general right hand side η_i, i=1,...,n. (Equivalently, one may apply Gauss-Jordan for n right hand sides represented by the columns of the identity matrix I.)

2.6. Linear functionals.

A linear operator from V_n into R was called a linear functional. Confer example 2.2.4. We write

$$\rho = \lambda(x), \; x \epsilon V_n, \; \rho \epsilon R$$

to indicate that λ evaluated at the vector x gives the real number ρ.

Example: In two dimensions vectors may be represented by arrows (see section 1.2.9). Linear functionals may then be visualized as systems of equally spaced parallel lines with an orientation. In order to evaluate the functional for a vector, i.e. an arrow, one counts the line spacings between tail and top of the vector. Loosely speaking, one counts how many lines are intersected by the arrow. See fig. 2.1. The sign is taken in agreement with the orientation. The idea generalizes to higher dimensions if systems of parallel hyperplanes are taken instead of systems of lines.

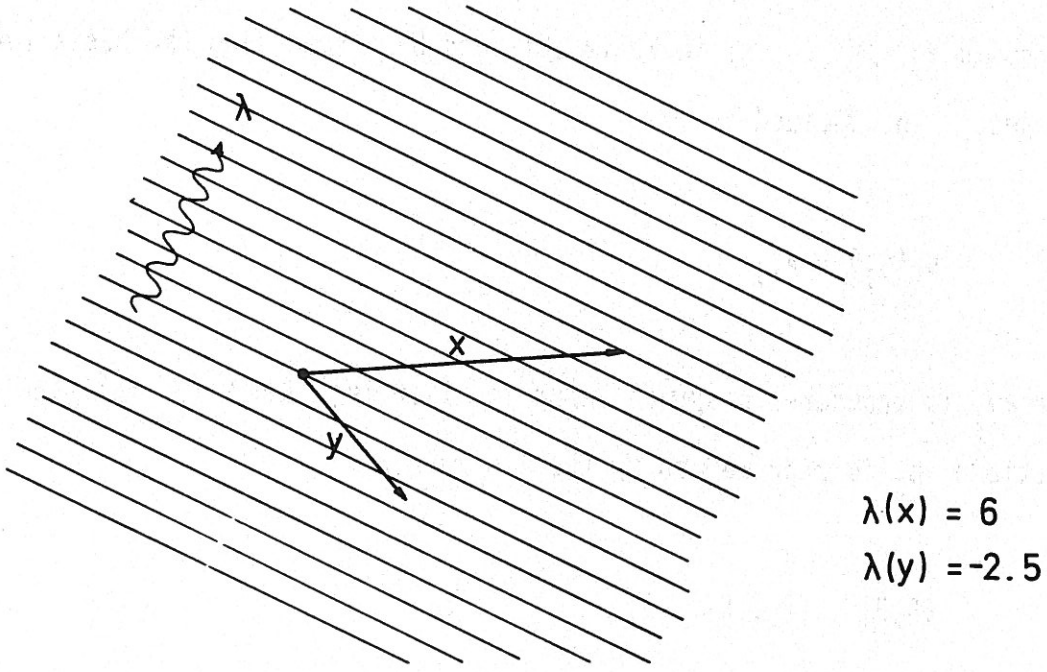


Fig. 2.1. Linear functionals represented by systems of lines

After choosing a basis e_j, $j=1,\ldots,n$, in V_n, a representation of the functional λ by a 1 by n matrix $\underline{\lambda}$ is obtained:

$$\underline{\lambda} = (\lambda_1, \ldots, \lambda_n)$$

If a vector x has coordinates ξ_j, then the functional λ evaluated at x gives the number

$$\lambda(x) = \sum_{i=1}^{n} \lambda_i \xi_i$$

Linear functionals form a vector space. Any linear combination of linear functionals is a linear functional. The vector space of functionals defined on V_n is called the dual vector space. It is frequently denoted V_n'. A basis dual to the basis e_j, j=1,...,n, in V_n is obtained by introducing the basis functionals ε_j, j=1,...,n, defined by

$$\varepsilon_i(e_j) = \delta_{ij}$$

Here δ_{ij} is Kronecker's symbol (equaling 1 if i=j, and 0 if i≠j). The basis functional ε_j is represented by the 1xn matrix

$$(0,0,\ldots,0,1,0,\ldots 0)$$

where the 1 appears at the j-position.

The coordinates of a functional λ with respect to the dual basis are precisely the components of its matrix representation.

Example: Fig. 2.2 shows the two arrows representing the chosen basis vectors in

the two-dimensional plane. The dual basis is represented by two systems of lines also shown in the figure.

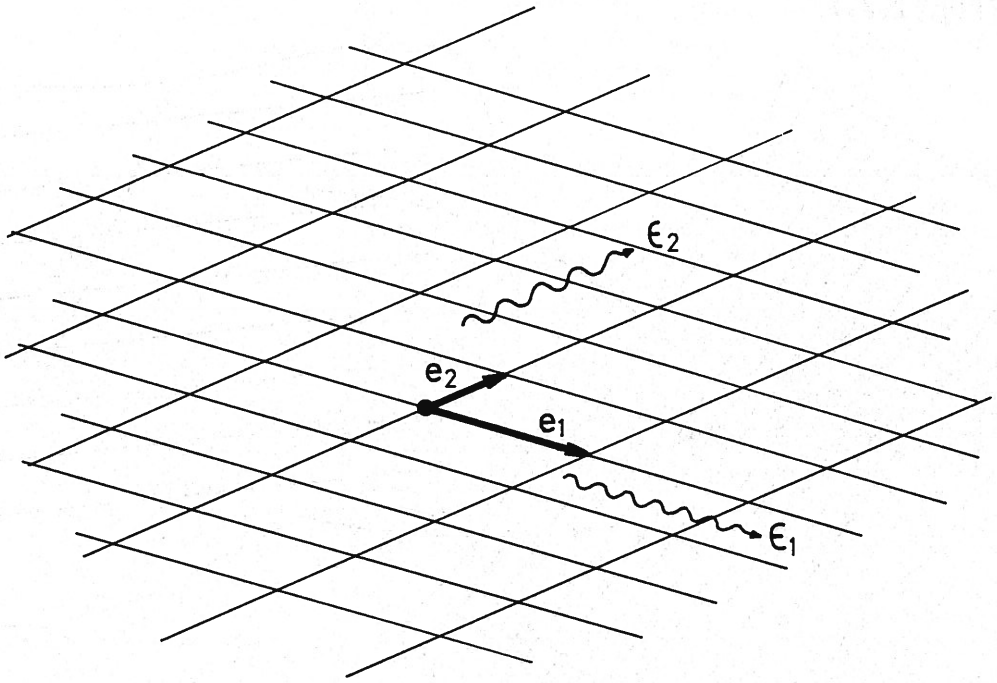


Fig. 2.2. Basis and dual basis.

2.7. Coordinates viewed as functionals.

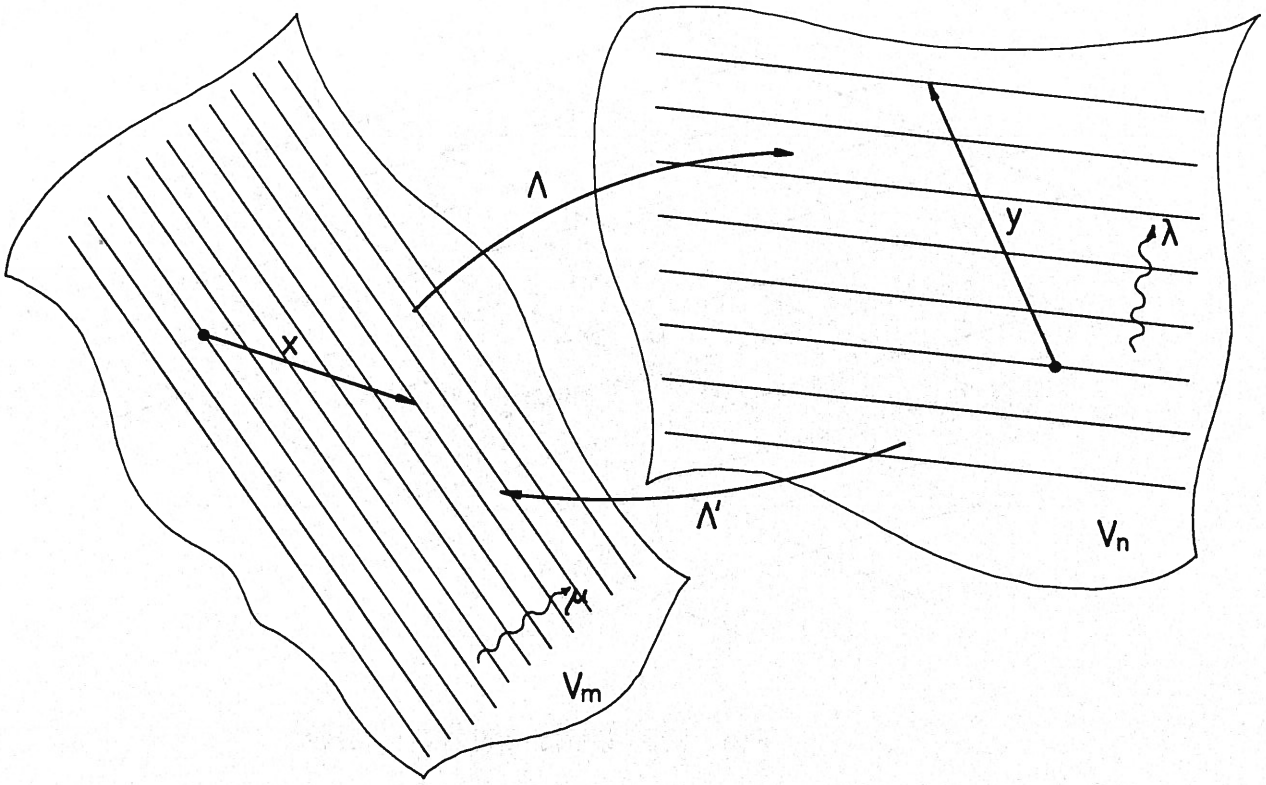
After choosing a basis e_j, $j=1,..,n$, in V_n, any vector x is represented by its coordinate n-tuple $(\xi_1,...,\xi_n)$. The mapping of x onto its i-th coordinate ξ_i is a linear functional, namely the basis functional ϵ_i.

2.8. The dual operator.

Let Λ be a linear operator from V_m into V_n. Let λ be a functional on V_n. The equation

$$\mu(x) = \lambda(\Lambda(x))$$

assigns a functional μ out of V_m' to any λ out of V_n'. A mapping Λ' from V_n' into V_m' is thus obtained. The mapping is linear. Λ' is called the dual operator of Λ (see fig. 2.3.).



$$\mu(x) = \lambda(\Lambda(x)) = \lambda(y) = 5$$

Fig. 2.3. Illustrating the definition of the dual operator.

Assume that bases in V_m, V_n, and corresponding dual bases in V_m', V_n' are chosen. Let $A=(\alpha_{ij})$ be the matrix representation of Λ. We are going to find the matrix representation of Λ'. We have

$$\mu(x) = \sum_{j=1}^{m} \mu_j \xi_j$$

On the other hand

$$\mu(x) = \lambda(\Lambda(x)) = \sum_{i=1}^{n} \lambda_i \sum_{j=1}^{m} \alpha_{ij}\xi_j = \sum_{j=1}^{m} \left\{ \sum_{i=1}^{n} \alpha_{ij}\lambda_i \right\} \xi_j$$

It is seen that

$$\mu_j = \sum_{i=1}^{n} \alpha_{ij}\lambda_i$$

Hence, if Λ is represented by the matrix

$$A = \begin{bmatrix} \alpha_{11} & \cdots & \cdots & \alpha_{1m} \\ \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ \alpha_{n1} & \cdots & \cdots & \alpha_{nm} \end{bmatrix}$$

then Λ' is represented by the transposed matrix

$$A^T = \begin{bmatrix} \alpha_{11} & \cdots & \cdots & \cdots & \alpha_{n1} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \alpha_{1m} & \cdots & \cdots & \cdots & \alpha_{nm} \end{bmatrix}$$

3. Matrix calculus.

3.1. Preliminaries.

It is time to formalise the computational rules for vectors and matrices. Matrix calculus is the appropriate tool.

An nxm matrix A is a rectangular array of real numbers α_{ij}, called its elements:

$$A = \begin{bmatrix} \alpha_{11} & \alpha_{12} & \cdots & \cdots & \alpha_{1m} \\ \alpha_{21} & \alpha_{22} & \cdots & \cdots & \alpha_{2m} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \alpha_{n1} & \alpha_{n2} & \cdots & \cdots & \alpha_{nm} \end{bmatrix}$$

It will be convenient to identify vectors with their coordinate n-tuples. This is legal in view of the isomorphism between any n-dimensional vector space V_n and R^n, the space of n-tuples. It will further be convenient to identify coordinate n-tuples with nx1 matrices, calling them "column vectors". Alternatively, a coordinate n-tuple may be identified with a 1xn matrix and called "row-vector".

As already pointed out in section 1.5, a matrix may be thought of being composed of row vectors, or, alternatively, of column vectors.

3.2 Interpretation of a matrix-vector product.

The matrix product of an nxm matrix A and an mx1 column vector gives as result

an nx1 column vector:

$$y = Ax$$

In conventional notation this means

$$\eta_i = \sum_{j=1}^{m} \alpha_{ij}\xi_j$$

The following different interpretations can be given to this system of equations.

(1) A system of linear equations. In section 1.5 the quantities η_i were denoted β_i, $i=1,\ldots,n$.

(2) The vector y is a linear combination of the columns of A. The scalar factors are given by ξ_j, $j=1,\ldots,m$.

(3) The n linear functionals, represented by the rows of A, evaluated for the vector x give the results η_i.

(4) Representation of a linear operator Λ from R^m into R^n. Confer section 2.3. If natural bases are chosen in R^m and R^n, then the images in R^n of the basis vectors in R^m are given by the columns of A:

$$\Lambda(e_j) = \sum_{i=1}^{n} a_{ij} f_i \ , \ j=1,\ldots,m$$

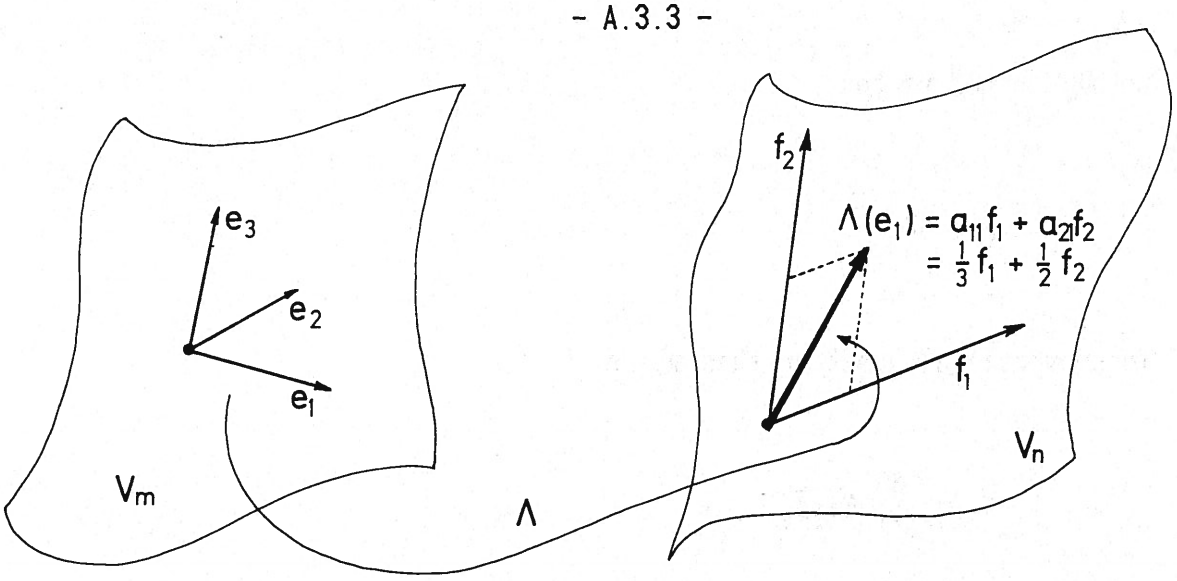


Fig. 3.1. Linear mapping of basis vectors: $\Lambda(e_j) = \sum_{i=1}^{n} a_{ij} f_i$

(5) If A is nxn, the spaces participating in the mapping may be identified. If A is nxn, the mapping is one to one and called an automorphism of R^n.

(6) Change of basis in R^n, coordinate transformation. Call the natural basis of R^n the old basis. Call the columns of a regular nxn matrix A the new basis. The relation

$$x_{old} = A\, x_{new}$$

may then be viewed as the representation of one and the same vector by coordinates with respect to the old and the new basis. (Note that x_{new} comprises the scalar factors in the linear combination of the vector x_{old} in terms of the new basis.)

The derivation of the above formula runs as follows (see also fig. 3.2!).

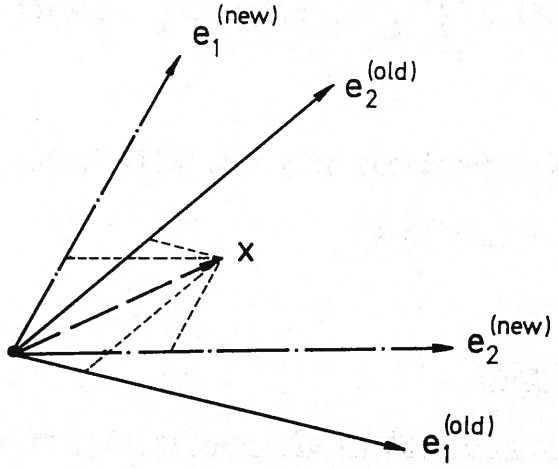


Fig. 3.2. Representing one and the same vector in terms of an old and a new basis.

$$x = \sum_{i=1}^{n} x_i^{(old)} e_i^{(old)} = \sum_{j=1}^{n} x_j^{(new)} e_j^{(new)}$$

Substituting

$$e_j^{(new)} = \sum_{i=1}^{n} a_{ij} e_i^{(old)}$$

one obtains

$$\sum_{i=1}^{n} x_i^{(old)} e_i^{(old)} = \sum_{j=1}^{n} x_j^{(new)} \sum_{i=1}^{n} a_{ij} e_i^{(old)} =$$

$$\sum_{i=1}^{n} \left(\sum_{j=1}^{n} a_{ij} x_j^{(new)} \right) e_i^{(old)}$$

By the uniqueness of coordinates we deduce

$$x_i^{(old)} = \sum_{j=1}^{n} a_{ij}\, x_j^{(new)}, \quad \text{i.e.} \quad x^{(old)} = A\, x^{(new)}$$

(7) Further interpretations of $y = Ax$ will follow after the definition of an inner product in section 4.

3.3. Matrix algebra.

3.3.1. Scalar multiplication of a matrix. An nxm matrix A may be multiplied by a scalar factor λ, yielding an nxm matrix B:

$$B = \lambda A$$

The elements of B are given by

$$\beta_{ij} = \lambda \alpha_{ij}$$

3.3.2. Sum of two nxm matrices. The sum is again nxm:

$$C = A + B$$

The elements of C are given by

$$\gamma_{ij} = \alpha_{ij} + \beta_{ij}$$

Remark. It is seen that the set of nxm matrices forms a vector space. However this viewpoint is not very important for our purposes.

3.3.3. Matrix product. It was already defined in section 2.4. Let A be pxn, B be nxm, C be pxm. The equation

$$C = AB$$

means

$$\gamma_{ij} = \sum_{k=1}^{n} \alpha_{ik}\beta_{kj}$$

For the interpretation of a matrix product as representation of a composite mapping see section 2.4. Formally the matrix product arises if a set of linear expressions is substituted into another:

$y = Bx$ substituted into $z = Ay$ gives $z = A(Bx) = (AB)x$

Important computational rules are

$$A(BC) = (AB)C = ABC \quad \ldots \text{ the associative law}$$

The associative law was already introduced in section 2.4. It was tacitly applied in the above substitution rule. We further have the rule:

$$(A+B)C = AC + BC \quad \ldots \text{ first distributive law}$$
$$A(B+C) = AB + AC \quad \ldots \text{ second distributive law}$$

$$\lambda(AB) = (\lambda A)B = A(\lambda B) = \lambda AB = AB\lambda$$

(scalar factors may be "pushed through matrix products".)

Remark: Note that the matrix product is generally not commutative. If C=AB, the product BA may not even be defined. If BA is defined, as for example in the case of nxn matrices A,B, then BA is generally different from AB.

3.3.4. Transposition. Let A be nxm. The transpose A^T of A was already introduced in section 2.8 . It is an mxn matrix having elements

$$\alpha_{ij}^{(T)} = \alpha_{ji},\ i=1,\ldots,m,\ j=1,\ldots,n.$$

The following computational rule applies:

$$(AB)^T = B^T A^T$$

This may either be verified directly. It may also be inferred from the fact that A^T represents the adjoint operator of that one represented by A: Let B represent a mapping from U into V. Let A represent a mapping from V into W. Then AB represents the composite mapping from U into W. Now, A^T represents the mapping from W' into V', B^T represents that one from V' into U'. Here U',V',W' are the dual spaces. Confer section 2.8. It follows without calculation that the composite mapping from W' into U' is represented by B^TA^T (confer fig. 3.3).

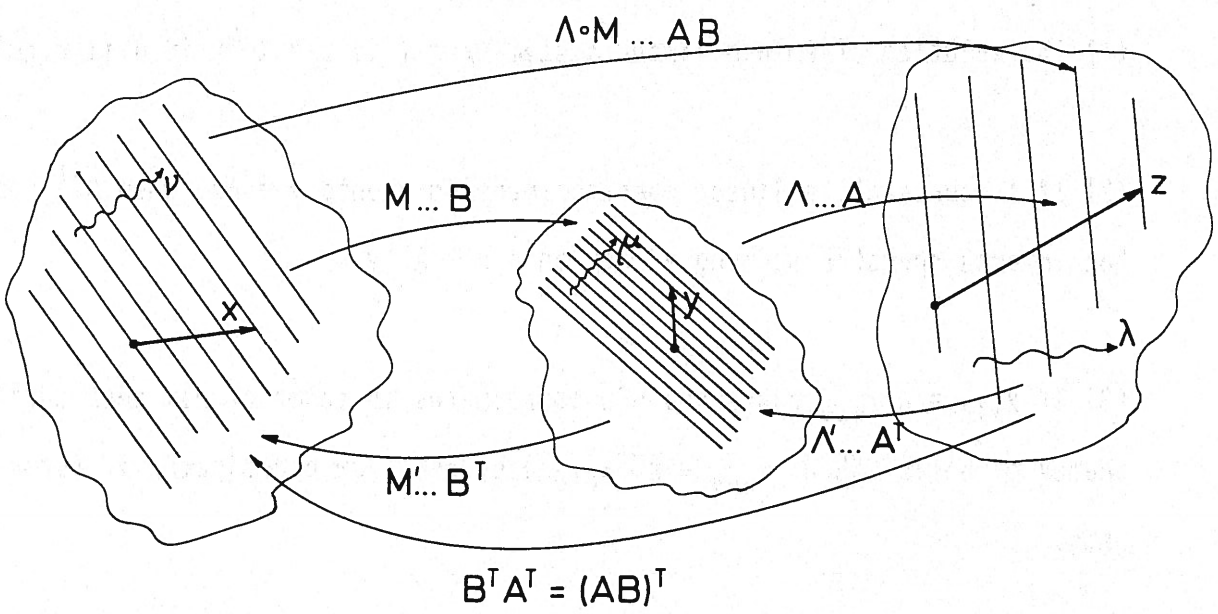


Fig. 3.3. Showing that $(\Lambda\circ M)' = M'\circ\Lambda'$, hence $(AB)^T = B^T A^T$.
(The operators Λ, M are represented by A, B, respectively.)

3.3.5. Inverse matrices. The inverse of a matrix was already introduced in section 2.5. If A is nxn and regular (i.e., rank(A)=n), then the inverse matrix A^{-1} exists and fulfills

$$A^{-1}A = AA^{-1} = I$$

If

$$y = Ax$$

then

$$x = A^{-1}y$$

These equations can be given the following interpretations:

(1) The solution of an nxn linear system $Ax = y$ is $x = A^{-1}y$ if A is regular.

(2) If A represents a linear operator mapping x onto $y = Ax$, then A^{-1} represents the inverse operator mapping y back onto $x = A^{-1}y$.

(3) If $x_{old} = A\, x_{new}$ expresses old coordinates in terms of new ones during a change of basis, then $x_{new} = A^{-1}x_{old}$ expresses new coordinates in terms of old ones.

The following computational rules apply

(1) $(AB)^{-1} = B^{-1}A^{-1}$,

(provided that A,B are nxn and invertible.). The proof relies on the associative law:

$$(B^{-1}A^{-1})(AB) = B^{-1}(A^{-1}A)B = B^{-1}B = I$$

(2) $(A^T)^{-1} = (A^{-1})^T$

The proof follows from transposing $AA^{-1} = I$.

4. Inner Products.

4.1. Definition.

Let V be a vector space. An inner product assigns a scalar number to any pair of vectors $a,b \in V$. This number is denoted (a,b). The following properties of an inner product are postulated:

$$(a,b) = (b,a) \quad \ldots\ldots\ldots\ldots\ldots\ldots \text{ symmetry}$$

$$(\lambda a,b) = \lambda(a,b) \quad \ldots\ldots\ldots\ldots\ldots \text{ homogeneity}$$

$$(a_1+a_2,b) = (a_1,b) + (a_2,b) \quad \ldots\ldots \text{ distributivity}$$

$$(a,a) > 0 \text{ if } a\neq 0 \quad \ldots\ldots\ldots\ldots \text{ positive definiteness}$$

A vector space V equipped with an inner product is called an inner product space. If V is infinite dimensional, one calls it a pre-Hilbert space.

4.2. Schwarz's inequality.

It reads

$$(a,b)^2 \leq (a,a)(b,b)$$

Proof: For any scalars λ,μ it follows from positive definiteness that

$$0 \leq (\lambda a+\mu b,\lambda a+\mu b) = \lambda^2(a,a) + 2\lambda\mu(a,b) + \mu^2(b,b)$$

Without loss of generality assume $a\neq 0$. Put $\mu=1$. Then

$$f(\lambda) = \lambda^2(a,a) + 2\lambda(a,b) + (b,b) \geq 0$$

The parabola $f(\lambda)$ must not cross the abscissa. The discriminant must be smaller or equal to zero:

$$(a,b)^2 - (a,a)(b,b) \leq 0$$

This is Schwarz's inequality.

4.3. Norms, distances.

For any vector $a \in V$ the number

$$\|a\| = \sqrt{(a,a)}$$

is meaningfully defined because $(a,a) \geq 0$. The number $\|a\|$ is called the norm or the length of the vector a.

Note that Schwarz's inequality may be rewritten as

$$|(a,b)| \leq \|a\|\|b\|$$

The following properties follow from those of the inner product:

$\|a\| \geq 0$, unless $a=0$ in which case $\|a\| = 0$ ... positivity

$\|\lambda a\| = |\lambda|\|a\|$ ... positive homogeneity

$\|a+b\| \leq \|a\| + \|b\|$ ..... triangle inequality

Proof of the triangle inequality:

$$(a+b,a+b) = (a,a) + 2(a,b) + (b,b)$$

i.e.

$$\|a+b\|^2 = \|a\|^2 + 2(a,b) + \|b\|^2$$

By Schwarz's inequality

$$\|a+b\|^2 \leq \|a\|^2 + 2\,\|a\|\|b\| + \|b\|^2$$

$$\|a+b\|^2 \leq (\|a\| + \|b\|)^2$$

Taking the square root, the triangle inequality is obtained.

The norm allows to define a distance between two vectors:

$$d(a,b) = \|a-b\|$$

The following properties follow immediately from the properties of the norm.

$d(a,b) > 0$ if $a \neq b$, $d(a,a) = 0$ ... positivity

$d(a,c) \leq d(a,b) + d(b,c)$ ... triangle inequality

The definition of a distance makes V a metric space.

Example: In the two dimensional plane, norms and inner products may be defined by a system of concentric and equally spaced circles. See fig. 4.1. If the tail of a vector is placed at the center, the circle passing through its top implies the norm.

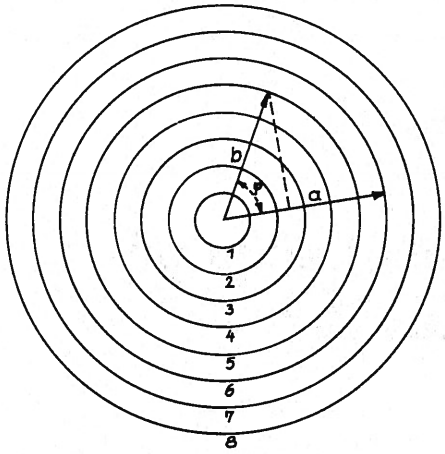


Fig. 4.1. $\|a\| = 6$, $\|b\| = 5$, $\varphi = 60°$, $(a,b) = \|a\|\,\|b\| \cos\varphi = 15$

The inner product can be defined by

$$(a,b) = \|a\|\,\|b\| \cos\varphi$$

This is $\|a\|$ times the norm of the orthogonal projection of b onto a, or likewise, $\|b\|$ times the norm of the orthogonal projection of a onto b. The four properties of the inner product should be verified.

An alternative way to define an inner product in the plane is as follows. The system of circles is changed to a system of ellipses by choosing an arbitrary axis and by shrinking the vertical distances with respect to this axis by an arbitrary factor. Norms and inner products are then defined as shown in fig.

4.2. One can say that norms and inner products of vectors in fig. 4.2 are the norms and inner products of fig. 4.1 applied to the pre-images under the affine mapping that turns circles into ellipses.

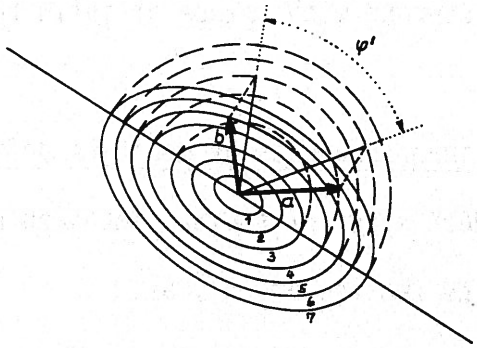


Fig. 4.2. $\|a\| = 6$, $\|b\| = 5$, $(a,b) = \|a\|\,\|b\| \cos \varphi' = 15$

4.4. Completeness, Hilbert spaces.

A sequence of vectors $a_1, a_2, \ldots$, is called a Cauchy sequence if for any positive number ε there exists an index $N(\varepsilon)$ such that

$$d(a_m, a_n) \leq \varepsilon \quad \text{for} \quad n,m \geq N(\varepsilon)$$

A metric space is called complete if any Cauchy sequence possesses a limit element in V: There must be an $a \in V$ such that for any positive ε there exists $N(\varepsilon)$ such that

$$d(a, a_n) \leq \varepsilon \quad \text{for } n \geq N(\varepsilon)$$

A complete inner product space is called a Hilbert space.

It is not difficult to show (from the completeness of R) that any finite dimensional vector space is complete. It is thus a Hilbert space, although this term is mostly used in context with spaces of infinite dimension.

4.5. Representation of inner products by positive definite matrices.

Let V_n be an inner product space of finite dimension n. Choose a basis $e_1,\ldots,e_n$. Represent the two vectors x,y as

$$x = \sum_i \xi_i e_i, \quad y = \sum_j \eta_j e_j$$

Expand

$$(x,y) = (\sum_i \xi_i e_i, \sum_j \eta_j e_j) = \sum\sum_{ij} \xi_i \eta_j (e_i, e_j)$$

Denote

$$\gamma_{ij} = (e_i, e_j)$$

then the nxn matrix

$$G = \begin{bmatrix} \gamma_{11} & \cdots & \cdots & \cdots & \gamma_{1n} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \gamma_{n1} & \cdots & \cdots & \cdots & \gamma_{nn} \end{bmatrix}$$

is symmetric. We have

$$(x,y) = \sum_{i=1}^{n}\sum_{j=1}^{n}\gamma_{ij}\xi_i\eta_j$$

Identifying x,y with their coordinate n-tuples, interpreting the coordinate n-tuples as nx1 matrices (column vectors), we write

$$(x,y) = x^TGy = y^TGx$$

We see: After choosing a basis, a given inner product is represented by a symmetric matrix. Is the converse also true? Does any symmetric matrix define an inner product? The answer is No! The matrix must fulfill one additional requirement. It must be positive definite.

Definition: A symmetric matrix G is positive definite if for any vector $x \neq 0$ the following inequality holds:

$$x^TGx > 0, \quad \text{if } x \neq 0$$

Equivalent definition: G is positive definite if for any numbers $\xi_1,\ldots,\xi_n$, not all of which are equal to zero, the inequality

$$\sum_{i=1}^{n}\sum_{j=1}^{n}\gamma_{ij}\xi_i\xi_j > 0$$

holds.

Positive definiteness is necessary for an inner product. For we must have

$$\|x\|^2 = x^T G x > 0$$

It is also sufficient because one may verify that all other properties of an inner product listed in section 4.1 are fulfilled.

Positive definite matrices are regular. G^{-1} exists. For a proof assume $Gx=0$. Multiply by x^T: $x^T G x=0$. This means $\|x\|^2 = 0$. Hence $\|x\| = 0$. Thus $x=0$. We have shown that the homogeneous system $Gx=0$ has only the zero solution. This means that G is regular.

4.6. Orthogonality.

Two vectors x,y are called orthogonal if their inner product vanishes

$$(x,y) = 0$$

Orthogonality depends on the choice of an inner product (but not on the choice of a basis!). If the basis vectors e_j, $j=1,\ldots,n$, are orthogonal, we have an orthogonal basis:

$$(e_i,e_i) = \gamma_{ii} > 0$$
$$(e_i,e_j) = 0 \quad \ldots \text{ if } i \neq j$$

If in particular $\|e_i\| = 1$, $i=1,\ldots,n$, we call the basis orthonormal. We then

have

$$(e_i, e_j) = \delta_{ij}$$

The inner product is then represented by the identity matrix:

$$G = I$$

Example: If an orthonormal basis is chosen in the plane, then the inner product of fig. 4.1, i.e.

$$(a,b) = \|a\| \, \|b\| \cos \varphi$$

is represented by the unit matrix:

$$(a,b) = a^T \, I \, b = a^T b$$

Proof: Use polar coordinates, writing

$$a = \begin{bmatrix} r_a \cos \varphi_a \\ r_a \sin \varphi_a \end{bmatrix}, \quad b = \begin{bmatrix} r_b \cos \varphi_b \\ r_b \sin \varphi_b \end{bmatrix}$$

Then

$$a^T b = r_a \, r_b \, \{ \cos \varphi_a \cos \varphi_b + \sin \varphi_a \sin \varphi_b \} =$$

$$r_a\, r_b \cos(\varphi_b - \varphi_a) = \|a\|\,\|b\| \cos\varphi = (a,b)$$

Example: Let the inner product in the plane be given as in fig. 4.2. Referring to fig. 4.2, the following choices of basis vectors imply the representation of the inner product given by the following matrices

$$e_1, e_2 \ldots \quad G = \begin{bmatrix} \gamma_{11} & \gamma_{12} \\ \gamma_{21} & \gamma_{22} \end{bmatrix}, \quad \gamma_{12} \neq 0$$

$$\bar{e}_1, \bar{e}_2 \ldots \quad G = \begin{bmatrix} 1 & 0 \\ 0 & \gamma_{22} \end{bmatrix}, \quad \gamma_{22} > 1$$

$$\bar{\bar{e}}_1, \bar{\bar{e}}_2 \ldots \quad G = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = I$$

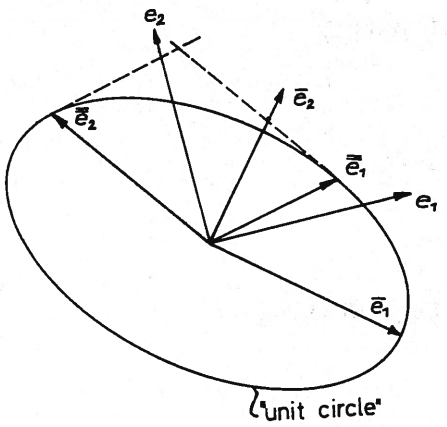


Fig. 4.3.

4.7. Gram-Schmid orthogonalization.

The question still remains whether orthogonal bases exist and how to obtain them. A complete answer is given by the Gram-Schmid orthogonalization procedure.

Let V be a vector space. Let $a_1, a_2, \ldots$, be a finite or infinite sequence of vectors. Suppose that any finite subset of these vectors is linearly independent. The orthogonalization procedure derives a sequence of orthonormal vectors $\bar{a}_1, \bar{a}_2, \ldots$, such that

$$\text{span}(a_1, a_2, \ldots, a_m) = \text{span}(\bar{a}_1, \bar{a}_2, \ldots, \bar{a}_m) \text{ for any } m=1,2,\ldots$$

The method proceeds as follows.

Put

$$\bar{a}_1 = \|a_1\|^{-1} a_1$$

Suppose now that the vectors $\bar{a}_1, \bar{a}_2, \ldots, \bar{a}_j$ have already been found fulfilling the above specified requirements. Represent the next vector $\bar{a}_{j+1}$ as

$$\bar{a}_{j+1} = \lambda_1 \bar{a}_1 + \lambda_2 \bar{a}_2 + \ldots + \lambda_j \bar{a}_j + \lambda_{j+1} a_{j+1}$$

Require orthogonality of $\bar{a}_{j+1}$ to the earlier obtained vectors $\bar{a}_1, \bar{a}_2, \ldots, \bar{a}_j$, i.e., require

$$(\bar{a}_{j+1}, \bar{a}_k) = 0, \quad k=1,\ldots,j$$

This leads to the equations

$$\lambda_k = -(\bar{a}_k, a_{j+1})\lambda_{j+1}$$

Thus we obtain

$$\bar{a}_{j+1} = \left\{ -\sum_{k=1}^{j} (\bar{a}_k, a_{j+1})\bar{a}_k + a_{j+1} \right\} \lambda_{j+1}$$

Abbreviate this as

$$\bar{a}_{j+1} = \lambda_{j+1} h_{j+1}$$

Put

$$\bar{a}_{j+1} = \|h_{j+1}\|^{-1} h_{j+1}$$

4.8. Representation of linear functionals by vectors.

Let x be a fixed vector. Then the inner product

$$(x,y) = \lambda(y)$$

assigns a number to any vector y. All requirements of λ to be a linear functional are fulfilled.

It is important that the converse is also true: At least in finite dimensional spaces any linear functional λ can be represented by a vector x. (In Hilbert spaces of infinite dimension an additional property of functionals must be required, namely continuity). We adhere to the finite dimensional case. We identify a functional λ with the coordinate n-tuple $(\lambda_1, \ldots, \lambda_n)$ with respect to

the dual basis. We know that the λ_i are also the elements of the representation of λ by an nx1 matrix. We interpret the coordinate n-tuple as a column vector, writing

$$\lambda(y) = \lambda^T y = \sum_{i=1}^{n} \lambda_i \eta_i$$

On the other hand the inner product (x,y) is represented as

$$(x,y) = x^T G y = \sum_{i}\sum_{j} \gamma_{ij} \xi_i \eta_j$$

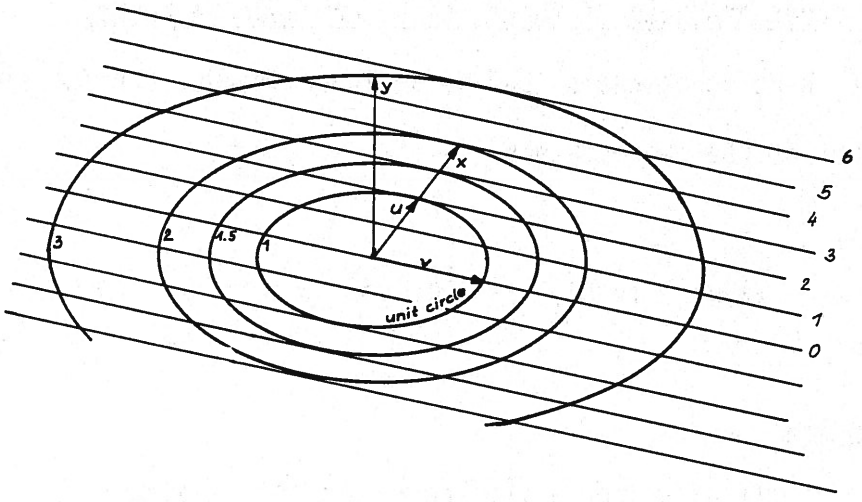


Fig. 4.4. Finding a vector x such that $\lambda(y) = (x,y)$. Choose vectors u, v with $\|u\| = \|v\| = 1$ as shown. Since $\lambda(v) = 0$, x must have the direction of u. $x = \xi u$. Since $\lambda(u) = 2$, we must have $(x,u) = (\xi u,u) = 2$, i.e. $\xi = 2$. This gives the vector x shown in the figure.

Hence the equation

$$\lambda = Gx \text{ or } \lambda_i = \sum_j \gamma_{ij}\xi_j$$

must hold in order to represent the linear functional λ by the vector x. Because G is regular, the representation is unique. We obtain

$$x = G^{-1}\lambda$$

The vector x is called the "representer" of the linear functional λ.

4.9. Inner products of functionals, reproducing kernel.

Let λ, μ be functionals. Define an inner product for functionals by their inner product of the representers:

$$(\lambda,\mu) = (x,y), \quad x=G^{-1}\lambda, \; y=G^{-1}\mu$$

We obtain:

$$(\lambda,\mu) = x^T Gy = \lambda^T G^{-1} G G^{-1}\mu = \lambda^T G^{-1}\mu = \lambda^T K\mu$$

It is seen that the inner product (λ,μ) for functionals in V' is represented by the matrix

$$K = G^{-1}$$

This matrix is called the reproducing kernel of the vector space V. The reproducing property of K is described by the equation

$$(K,x) = x$$

More precisely, if α_{ij} are the elements of K, then

$$\sum_{j=1}^{n} \alpha_{ij} \sum_{k=1}^{n} \gamma_{jk} x_k = x_i$$

This is clear from $(K,x) = KGx = Ix = x$.

4.10. The adjoint operator.

If the linear operator Λ maps V into W, then the dual operator Λ' maps functionals $\lambda \epsilon W'$ back onto functionals $\mu \epsilon V'$. The defining equation of Λ' was given in section 2.9. In slightly different notation it reads:

$$\mu(z) = \lambda(\Lambda(z)) = \Lambda'(\lambda)(z), \quad \text{for all } z \epsilon V$$

If Λ is represented by the n×m matrix A, then Λ' is represented by the m×n matrix A^T. The corresponding proof given in section 2.9 becomes very simple if matrix calculus is used. It suffices to rewrite the above equation as

$$\mu^T z = \lambda^T(Az) = (A^T\lambda)^T z, \quad \text{for all } z \epsilon V$$

If inner products are available in V and W, one can define the adjoint operator.

This is done by switching from the functionals $\mu \epsilon V'$, $\lambda \epsilon W'$ to their representers $x \epsilon V$, $y \epsilon W$. According to section 4.8 this is done by means of the relations

$$\mu(z) = (x,z), \quad \text{for all } z \epsilon V$$

and

$$\lambda(z) = (y,z), \quad \text{for all } z \epsilon W$$

The transition from λ to μ via

$$\mu = \Lambda'(\lambda)$$

corresponds to a transition from the representer y of λ to the representer x of μ:

$$x = \Lambda^*(y)$$

The operator Λ^* is linear because Λ' is linear, and because the isometries between V and V' and between W and W' are linear. Λ^* is called the adjoint operator.

The defining equation for Λ^* is obtained by rewriting the defining equation of Λ' as

$$(x,z) = (y,\Lambda(z)) = (\Lambda^*(y),z), \quad \text{for all } z \epsilon V$$

Finally we specify the matrix representation A^* of Λ^*. Let the inner products in V and W be represented by G_V, G_W, respectively. Recall that μ and λ are related to their representers x and y by

$$\mu = G_V x, \quad \lambda = G_W y$$

Substitute in

$$\mu = A^T \lambda$$

for μ and λ to obtain

$$G_V x = A^T G_W y$$

or

$$x = G_V^{-1} A^T G_W y$$

It is seen that the matrix representation of Λ^* is

$$A^* = G_V^{-1} A^T G_W$$

Remark: The matrix representation A^* of Λ^* may also be derived in the following way: Write the defining equation for Λ^* as

$$(\Lambda(x), y) = (x, \Lambda^*(y)) \quad \text{for all } x \in V,\ y \in W$$

Use the matrix representations A, A^* to obtain

$$(Ax,y) = (x,A^*y)$$

or

$$x^T A^T G_W y = x^T G_V A^* y,$$

showing once more that $A^* = G_V^{-1} A^T G_W$.

5. Projectors.

5.1. Decomposition of a vector space into a direct sum of subspaces.

Let V be a vector space, and let V_A, V_B be subspaces which have only the zero vector in common. We consider the vector space V_C of all vectors represented as

$$c = a + b, \quad a \epsilon V_A, \; b \epsilon V_B$$

We now show that the above decomposition is unique. Suppose that

$$c = a' + b', \quad a' \epsilon V_A, \; b' \epsilon V_B$$

Subtracting we obtain

$$0 = (a-a') + (b-b')$$

Now $(a-a') \epsilon V_A$, hence $(b-b') = -(a-a') \epsilon V_A$. On the other hand, $(b-b') \epsilon V_B$. It follows that $(b-b')$ is in V_A as well as in V_B. It must be the zero vector. Consequently $b=b'$. Similarly $a=a'$ is shown.

Because of the uniqueness of the decomposition, we obtain two mappings, one from V_C onto V_A, the other from V_C onto V_B:

$$a = \Pi_A(c), \quad b = \Pi_B(c), \quad c \epsilon V_C$$

It is easily seen that these mappings are linear. They are called the

projections of V_C onto V_A and V_B respectively.

Assume that V_C is of finite dimension n. Let the dimension of V_A be m. It follows necessarily that the dimension of V_B is n-m. This is easily proved by choosing a basis in V_A and a basis in V_B.

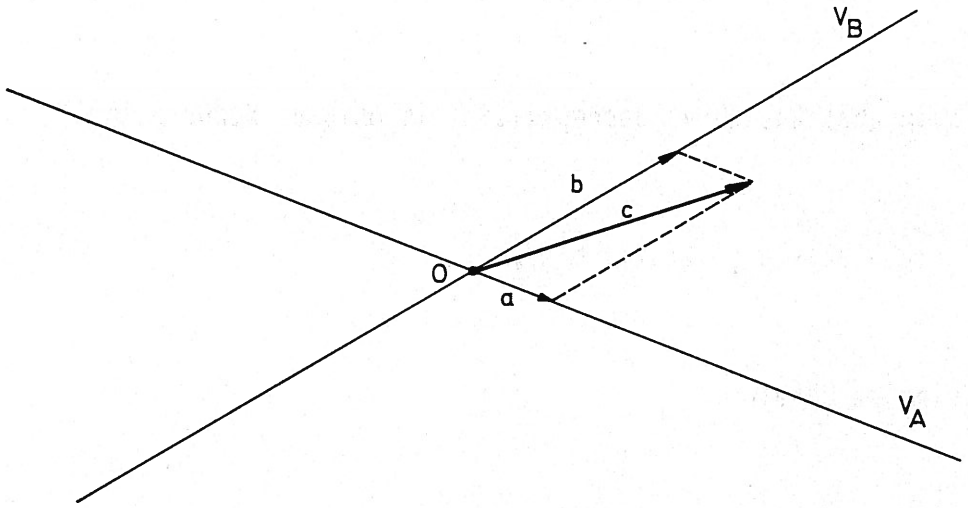


Fig. 5.1. Illustrating the uniqueness of the decomposition
$c = a+b$

5.2. Orthocomplementary subspaces.

Let V_A, V_B be subspaces of V. Assume that any vector in V_A is orthogonal to any vector in V_B:

$$(a,b) = 0, \text{ if } a \in V_A \text{ and } b \in V_B$$

A vector belonging to V_A as well as to V_B is orthogonal to itself. Its norm is zero; it must be the zero vector. Hence V_A and V_B have only the zero vector in common. We may form the direct sum V_C as we did in the previous subsection. We

call V_A and V_B orthocomplementary subspaces of V_C. We also say that V_B is the orthocomplement of V_A. Likewise, V_A is the orthocomplement of V_B. In symbols

$$V_B = V_A^{\perp}, \quad V_A = V_B^{\perp}$$

Note that the orthocomplement of the orthocomplement gives the original subspace

$$(V_A^{\perp})^{\perp} = V_A$$

The linear operators Π_A and Π_B introduced in the previous section are called orthogonal projectors.

5.3. The theorem by Pythagoras.

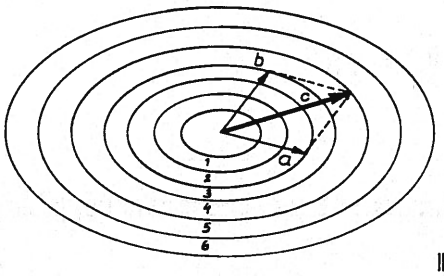
Let a vector $c \epsilon V_C$ be represented as the sum of its orthogonal projections onto orthocomplementary subspaces V_A and V_B:

$$c = a + b$$

Then

$$\|c\|^2 = \|a\|^2 + \|b\|^2$$

Proof: $\|c\|^2 = (c,c) = (a+b,a+b) = (a,a) + 2(a,b) + (b,b) = \|a\|^2+\|b\|^2$, because $(a,b)=0$.



$$\|c\|^2 = \|a\|^2 + \|b\|^2$$
$$5^2 = 3^2 + 4^2$$

Fig. 5.2. The theorem by Pythagoras.

Theorem. Consider the following extremum problem: Given $c \in V_C$, find $x \in V_A$ such that

$$\|c-x\| = \text{minimum},$$

The solution is

$$x = a = \Pi_A(c).$$

Proof: Decompose

$$c = a + b,\ a \in V_A,\ b \in V_B$$

Represent

$$c-x = (a-x) + b$$

It is seen that $(a-x) \epsilon V_A$ and $b \epsilon V_B$. The theorem by Pythagoras implies

$$\|c-x\|^2 = \|a-x\|^2 + \|b\|^2$$

This is minimal for a=x, which was to be shown.

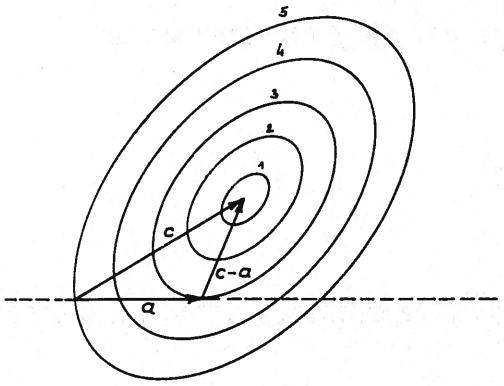


Fig. 5.3. Projection as the solution of a minimum problem

5.4. Matrix representation of orthogonal projectors.

Assume V_A, V_B and V_C of dimension m, n-m, n, respectively. Let an inner product in V_C be represented by the symmetric and positive definite matrix G. For simplicity, identify V_A, V_B, V_C with R^m, R^{n-m}, R^n. Let the columns of the nxm matrix A be a basis in V_A. Likewise, let the columns of the nx(n-m) matrix B be a basis in V_B. Then the columns of the matrix (A,B), whose columns are those of

A followed by those of B, form a basis in V_C. Moreover, the orthogonality requirement of V_A and V_B implies:

$$A^TGB = 0,\ B^TGA = 0$$

(Recall that inner products (a,b) are written a^TGb; note that the elements of A^TGB are just all inner products of any basis vector in V_A with any basis vector in V_B.)

Vectors aeV_A and beV_B are uniquely represented as

$$a = Ax,\ b = By$$

Here x and y are vectors of coordinates with respect to the bases in V_A, V_B. The decomposition

$$c = a + b,\ aeV_A,\ beV_B$$

is therefore equivalently written as

$$c = Ax + By$$

Form all inner products of c with basis vectors in A, i.e., multiply the above equation by A^TG. In view of $A^TGB = 0$ we obtain

$$(A^TGA)x = A^TGc$$

This is a set of equations called "normal equations". We prove that the mxm normal equation matrix A^TGA is symmetric and positive definite. Symmetry follows from the symmetry of G and the transposition rule for matrix products:

$$(A^TGA)^T = A^TG^T(A^T)^T = A^TGA$$

In order to prove positive definiteness, we must show that for any nonzero x we have $x^T(A^TGA)x > 0$. We note that

$$x^T(A^TGA)x = (Ax)^TG(Ax) = (Ax,Ax) = \|Ax\|^2 \geq 0$$

Assume now that $\|Ax\|^2=0$. Then $\|Ax\|=0$. By the positivity of the norm we infer $Ax=0$. From the uniqueness of coordinates with respect to the basis in V_A it follows that $x=0$. The proof of positive definiteness is complete.

The solution of the normal equations is uniquely obtained as

$$x = (A^TGA)^{-1}A^TGc$$

Inserting this into $a = Ax$, we get

$$a = \Pi_A(c) = A(A^TGA)^{-1}A^TGc = P_Ac$$

Here we have introduced the matrix

$$P_A = A(A^TGA)^{-1}A^TG$$

It is the matrix representation of the operator Π_A projecting V_C onto V_A. Similarly

$$P_B = B(B^TGB)^{-1}B^TG$$

represents Π_B, the projection operator from V_C onto V_B. The equation

$$c = a + b$$

or

$$c = \Pi_A(c) + \Pi_B(c)$$

shows that

$$\Pi_A + \Pi_B = I,$$

the identity operator in V_C. This equation implies the matrix equation

$$P_A + P_B = I$$

which may also be proved algebraically as follows. Multiply the last equation from behind by the matrix (A,B). Obtain $(P_A+P_B)(A,B) = (A,B)$. From the

regularity of (A,B) whose columns form a basis in V_C infer the desired relation $P_A+P_B=I$.

We further note the following relation

$$\Pi_A \circ \Pi_B = 0$$

The corresponding matrix relation is

$$P_A P_B = 0$$

The proof is geometrically as easy as algebraically.

5.5. Projections of functionals.

Due to the isometry between a vector space V and its dual V' (confer section 4.8), projections of functionals can be introduced in a very natural way. We consider the dual spaces V_A', V_B', V_C'. If G represents the inner product in V_C with respect to the chosen basis, then $K=G^{-1}$ represents the inner product in V_C' with respect to the dual basis. Recall that any linear functional $\lambda \epsilon V_C'$ is related to its representing vector x by $\lambda=Gx$, $x=K\lambda$. If V_A is spanned by the columns of the matrix A, then V_A' is spanned by the columns of A'=GA. Likewise V_B' is spanned by B'=GB. V_C' is decomposed into orthocomplementary subspaces V_A', V_B'. The relation

$$A'^T K B' = 0$$

expresses the orthogonality of V_A', V_B' in matrix form. The projectors Π_A', Π_B' onto V_A', V_B' are represented by

$$\lambda_A = P_A'\lambda, \quad \lambda_B = P_B'\lambda$$

with

$$P_A' = A'(A'^{T}KA')^{-1}A'^{T}K$$
$$P_B' = B'(B'^{T}KB')^{-1}B'^{T}K$$

This is all very obvious because of the isometry. It is also obvious that the representer x_A of the projected functional λ_A is the projection P_Ax of the representer x of λ. However there is the following interesting characterization of functionals $\alpha \epsilon V_A'$ and $\beta \epsilon V_B'$ in terms of the vector subspaces V_A, V_B:

$\alpha \epsilon V_A'$ is equivalent to $\alpha(y) = 0$ for any $y \epsilon V_B$

$\beta \epsilon V_B'$ is equivalent to $\beta(x) = 0$ for any $x \epsilon V_A$

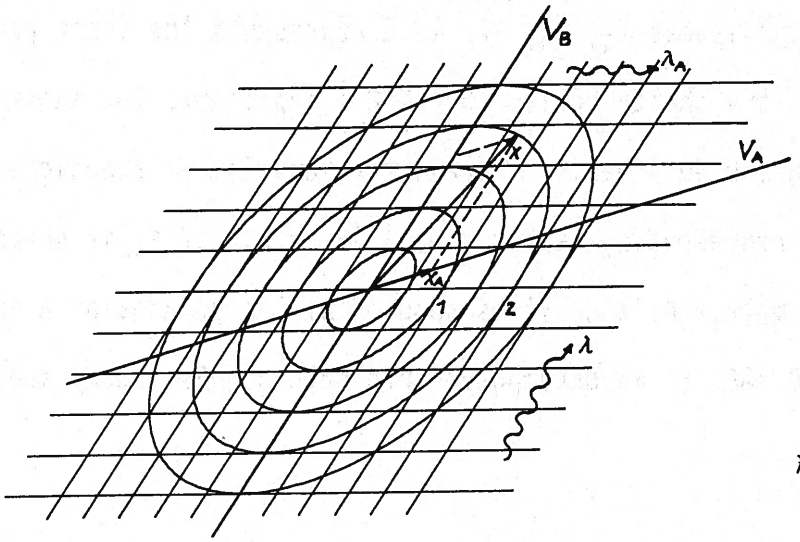


Fig. 5.4. The projection of a functional.

This is clear, because if $\alpha \epsilon V_A'$, α is represented by a vector $a \epsilon V_A$, and $\alpha(y) = (a,y)$. It follows that $\alpha(y) = 0$ for any $y \epsilon V_B$. On the other hand, if a functional α is represented by a, and if $\alpha(y) = (a,y) = 0$ for any $y \epsilon V_B$, then a must be in V_A. Hence α must be in V_A'.

The projector of a functional $\lambda_A = P_A' \lambda$ is characterized in the following way

$$\lambda_A(y) = \lambda(y) \quad \text{for } y \epsilon V_A$$
$$\lambda_A(y) = 0 \quad \text{for } y \epsilon V_B$$

The proof follows again easily from $\lambda_A(y) = (x_A, y)$.

The projection operator Π_A' is the dual of the projection operator Π_A. Cf. section 2.9.

$$\lambda_A(x) = (\Pi_A' \lambda)(x) = \lambda(\Pi_A(x))$$

For a proof one just inserts for x vectors either in V_A or in V_B.

We also have the following

Theorem. If λ is a linear functional in V_C', than the projection $\lambda_A = P_A' \lambda$ is the solution of the following extremum problem: Given λ, find a functional μ such that

$\|\mu\|$ = minimum

subject to

$\mu(a) = \lambda(a)$ for any $a \epsilon V_A$

Proof: The relation $\mu(a)=\lambda(a)$ for any $a\epsilon V_A$, is equivalent to $(\lambda-\mu)(a) = 0$ for any $a\epsilon V_A$. By our earlier characterization of V_B', this is equivalent to $(\lambda-\mu)\epsilon V_B'$. This in turn is equivalent to $(\lambda-\mu)_A=0$, or $\mu_A=\lambda_A$. We see that any μ qualifying for minimization is representable as $\mu=\lambda_A+\nu$, $\nu\epsilon V_B'$. By the theorem of Pythagoras we have $\|\mu\|^2 = \|\lambda_A\|^2 + \|\nu\|^2$. Hence $\mu=\lambda_A$ is the smallest.

Suppose that μ_A is a functional whose domain is V_A. Its values for vectors not in V_A are unspecified. We call μ an extension of μ_A to V_C if μ is defined on all of V_C and if $\mu(a) = \mu_A(a)$ for any $a\epsilon V_A$. The following theorem is rather obvious.

Theorem. The extension μ of μ_A having the smallest possible norm is given by

$\mu(a) = \mu_A(a)$ ... for $a\epsilon V_A$

$\mu(b) = 0$ ..... for $b\epsilon V_B$

Equivalently

$\mu(x) = \mu_A(\Pi_A(x))$, $x\epsilon V_C$

The minimal norm is given by

$$\|\mu\| = \|\mu_A\|$$

Of course, μ_A is the projection of μ onto V_A'.

Remark: Referring to section 4.10 on the redefinition of the adjoint operator in case of inner product spaces, note that P_A is a self-adjoint operator:

$$P_A^* = P_A, \quad P_A^* = G^{-1}P_A{}^T G$$

The adjoint equals the original operator.

6. Least squares adjustment.

6.1. Projecting the vector of observations.

It remains to change notation in order to conform with sacred traditions in least squares adjustment. Let L be the n-dimensional vector space of observations. A vector $\ell \in L$ has coordinates

$$\ell = \begin{bmatrix} \ell_1 \\ \ell_2 \\ .. \\ .. \\ \ell_n \end{bmatrix}$$

Any coordinate corresponds to one individual measurement such as a direction, a distance, an azimuth or a Doppler count. Of course, in an originally nonlinear problem the observations are replaced by small increments with respect to approximative quantities. Note that we deviate from the earlier rule to use Greek letters for coordinates.

We introduce the subspace L_A of adjusted observations. Corrections v must be added to the observations in order to force the adjusted observations into L_A:

$$\ell + v \in L_A$$

Equivalently

$$\ell + v = Ax$$

As before, the columns of the nxm matrix A are a basis of L_A. The vector v of corrections is a member of L. The size of any vector in L is measured by a norm derived from an inner product. Let the inner product be represented by the "weight matrix" P. The matrix P is symmetric and positive definite. It need not be diagonal, although in most applications it is assumed to be so. We like to have corrections as small as possible. Thus we arrive at

$$\|\ell - a\| = \text{minimum}, \quad a \in L_A, \text{ i.e., } a = Ax$$

The solution was already obtained in the previous section:

Form the normal equations

$$(A^T P A)x = A^T P \ell$$

to obtain

$$\ell + v = Ax = A(A^T P A)^{-1} A^T P \ell = P_A \ell$$

The corrections are given by

$$v = -(\ell - P_A \ell) = -(I - P_A)\ell = -P_B \ell$$

The requirement $\ell + v \in L_A$ is equivalently replaced by $\ell + v$ orthogonal to L_B, i.e.

$$B^T P(\ell + v) = 0$$

In conditioned adjustment it is customary to replace the columns of the matrix B by the corresponding functionals. Confer section 4.8. Thus one introduces

$$B' = PB, \quad B = P^{-1}B'$$

Inserting into the previous equation gives the condition equations

$$B'^{T}(\ell+v) = 0$$

Minimizing v gives, as we know, the solution

$$v = -P_B\ell$$

$$v = -B(B^{T}PB)^{-1}B^{T}P\ell = -P^{-1}B'(B'^{T}P^{-1}B')^{-1}B'^{T}\ell$$

One introduces correlates k by

$$k = -(B'^{T}P^{-1}B')^{-1}B'^{T}\ell$$

This permits us to write

$$v = P^{-1}B'k$$

The correlates are the solution of the normal equations of adjustment by condition:

$$(B'^{T}p^{-1}B')k + w = 0$$

The vector w of discrepancies is given by

$$w = B'^{T}\ell$$

6.2 Inhomogeneous form of least squares adjustment.

Frequently an adjustment problem is posed as follows.

Minimize $\|v\|^2$ subject to

$$\ell+v = a_0 + Ax \quad \text{(variation of parameters)}$$

or subject to

$$B^{T}(\ell+v) = b_0 \quad \text{(conditions)}$$

In the case of variation of parameters the requirement is

$$\|\ell-(a_0+Ax)\|^2 = \text{minimum}$$

If it is rewritten as

$$\|(\ell-a_0) - Ax\|^2 = \text{minimum}$$

we arrive at the earlier case with ℓ replaced by $\ell-a_0$.

The solution is obtained from the normal equations

$$(A^T PA)x = A^T P(\ell - a_0)$$

$$v = -(I - P_A)(\ell - a_0)$$

In case of the conditioned adjustment we introduce a particular solution a_0 of the inhomogeneous system $B'^T a_0 = b_0$. We then have

$$B'^T((\ell - a_0) + v) = 0$$

This reduces again to the earlier case if ℓ is replaced by $(\ell - a_0)$. The solution is

$$v = -P_B(\ell - a_0) = PB'k$$

with

$$k = -(B'^T P^{-1} B')w$$

and

$$w = B'^T(\ell - a_0) = B'^T \ell - b_0$$

6.3. The fundamental rectangular triangle of least squares adjustment.

The triangle is formed by

hypothenuse	$c = \ell - a_0$
1st short side	$a = Ax = \Pi_A(\ell - a_0)$
2nd short side	$b = -v = \Pi_B(\ell - a_0) = (I - P_A)(\ell - a_0)$

The vector $(\ell-a_0)$ is orthocomplementary decomposed into a=Ax and -v. The theorem by Pythagoras shows:

$$\|\ell-a_0\|^2 = \|Ax\|^2 + \|v\|^2$$

or

$$\|v\|^2 = \|\ell-a_0\|^2 - \|Ax\|^2$$

or

$$v^TPv = (\ell-a_0)^TP(\ell-a_0) - x^TA^TPAx$$

Putting a=Ax, one also recognizes $(a,a) = (a,\ell-a_0+v) = (a,\ell-a_0)$, because $(a,v) = 0$. Hence

$$v^TPv = (\ell-a_0)^TP(\ell-a_0) - (\ell-a_0)^TPAx$$

Furthermore $(v,v) = -(v,\ell-a_0)$. Using $v = P^{-1}B'k$, one gets

$$v^TPv = -k^TB'^TP^{-1}P(\ell-a_0) = -k^T(B'^T\ell-b_0) = -k^Tw$$

6.4. Least squares adjustment by projecting functionals.

Let L' be the dual space of L, the space of observations. Any $\lambda \in L'$ is a linear form in the observables. Thus if the observations are angles, distances e.t.c., then λ may refer to a coordinate of a station, to an area, or to any other quantity depending linearly on the observations. (Nonlinear adjustment problems have to be linearized, of course.) Recall that the coordinates

of any vector are functionals too. Hence λ may just refer to any particular coordinate ℓ_i of ℓ.

The size of λ is measured by its norm $\|\lambda\|$. We have

$$\|\lambda\|^2 = \lambda^T Q \lambda$$

Here $Q=P^{-1}$, the matrix representation of the reproducing kernel of L.

The subspace L_A is the space of adjusted observations. We want to replace λ by a functional λ_A such that $\lambda_A(a)$ coincides with $\lambda(a)$ for any adjusted observation $a \in L_A$, and such that λ_A is as small as possible. The minimum problem

$$\|\mu\| = \text{minimum}$$

subject to

$$\mu(a) = \lambda(a) \text{ for any } a \in L_A$$

was solved in section 5.5. The solution is the projection

$$\lambda_A = \Pi_A'(\lambda)$$

If λ is identified with its coordinate column vector λ (with respect to the dual basis), then

$$\lambda_A = P_A' \lambda$$

with

$$P'_A = A'(A'^T Q A')^{-1} A'^T Q$$

Putting A'=PA, we obtain

$$\lambda_A = PA(A^T PA)^{-1} A^T \lambda$$

Applying the adjusted functional toward the observation vector, we get

$$\lambda_A(\ell) = \lambda_A^T \ell = \lambda^T A(A^T PA)^{-1})A^T P\ell = \lambda^T P_A \ell = \lambda^T(\ell_A)$$

The familiar rule is recovered. The adjusted linear functional applied to the original observations is obtained by inserting the adjusted observations into the original functional. This demonstrates the equivalence of the two adjustment procedures.

Remark. The functional approach toward least squares adjustment has the following advantages:

(1) It lends itself to a statistical interpretation. Confer part B.

(2) It generalizes to vectors of infinitely many observations. It may not be possible to assign a finite norm to such a vector of observations. Hence least squares adjustment of stochastic processes relies on the functional approach.

7. Partitioned matrices.

7.1. Definitions.

Consider a matrix of size $(n_1+ n_2)x(m_1+ m_2)$:

$$A = \left[\begin{array}{ccc:ccc} \alpha_{1,1} & \cdots & \alpha_{1,m_1} & \alpha_{1,m_1+1} & \cdots & \alpha_{1,m_1+m_2} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \alpha_{n_1,1} & \cdots & \alpha_{n_1,m_1} & \alpha_{n_1,m_1+1} & \cdots & \alpha_{n_1,m_1+m_2} \\ \hdashline \alpha_{n_1+1,1} & \cdots & \alpha_{n_1+1,m_1} & \alpha_{n_1+1,m_1+1} & \cdots & \alpha_{n_1+1,m_1+m_2} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \alpha_{n_1+n_2,1} & \cdots & \alpha_{n_1+n_2,m_1} & \alpha_{n_1+n_2,m_1+1} & \cdots & \alpha_{n_1+n_2,m_1+m_2} \end{array}\right]$$

The matrix may be partitioned as indicated by the dashed lines. Calling the submatrices A_{11}, A_{12}, A_{21}, A_{22}, one writes

$$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}$$

The concept of partitioning may be generalized in an obvious way:

$$A = \begin{bmatrix} A_{11} & A_{12} & \cdots & A_{1m} \\ \vdots & & & \vdots \\ A_{n1} & A_{n2} & \cdots & A_{nm} \end{bmatrix}$$

The submatrices are sometimes called "blocks". Also vectors may be partitioned into subvectors.

$$x = \left[\begin{array}{c} \xi_1 \\ \xi_2 \\ \cdot \\ \cdot \\ \xi_m \\ \hline \xi_{m+1} \\ \cdot \\ \cdot \\ \xi_n \end{array}\right] = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

7.2. Computational rules.

7.2.1. Scalar multiplication. Partitioned matrices may be multiplied by a scalar. Obviously all submatrices are multiplied by the scalar.

7.2.2. Addition. Partitioned matrices may be added. Provided that the dimensions of corresponding submatrices coincide, one simply adds corresponding submatrices.

7.2.3. Transposition. The transpose of a partitioned matrix may be formed, e.g.

$$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \qquad A^T = \begin{bmatrix} A_{11}^T & A_{21}^T \\ A_{12}^T & A_{22}^T \end{bmatrix}$$

7.2.4. Matrix multiplication. Partitioned matrices may be multiplied under suitable circumstances. This is best explained by an example. Let

$$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \\ A_{31} & A_{32} \end{bmatrix} \qquad B = \begin{bmatrix} B_{11} & B_{12} & B_{13} & B_{14} \\ B_{21} & B_{22} & B_{23} & B_{24} \end{bmatrix}$$

We say that A has p=3 generalized rows and n=2 generalized columns. The number of generalized columns of A coincides with the number of generalized rows of B. B has m=4 generalized columns. Let the matrix C be composed of p×m blocks:

$$C = \begin{bmatrix} C_{11} & C_{12} & C_{13} & C_{14} \\ C_{21} & C_{22} & C_{23} & C_{24} \\ C_{31} & C_{32} & C_{33} & C_{34} \end{bmatrix}$$

Then the product

$$C = AB$$

may be formed by

$$C_{ij} = \sum_{k=1}^{n} A_{ik} B_{kj}$$

provided that all matrix products $A_{ik}B_{kj}$ are defined. This is the case if in any

block multiplication $A_{ik}B_{kj}$ the first factor has as many (ordinary) columns as the second factor has (ordinary) rows.

7.3 Block diagonality.

The n×n matrix A is called block-diagonal if A is represented as

$$A = \begin{bmatrix} A_{11} & 0 \\ 0 & A_{22} \end{bmatrix}$$

Here A_{11} is $n_1 \times n_1$ and A_{22} is $n_2 \times n_2$ and $n_1+n_2=n$. A_{11} and A_{22} are square matrices.

If the inverse matrices A_{11}^{-1}, A_{22}^{-1} exist, then

$$A^{-1} = \begin{bmatrix} A_{11}^{-1} & 0 \\ 0 & A_{22}^{-1} \end{bmatrix}$$

This is readily proved by verifying

$$AA^{-1} = \begin{bmatrix} A_{11}A_{11}^{-1} & 0 \\ 0 & A_{22}A_{22}^{-1} \end{bmatrix} = \begin{bmatrix} I & 0 \\ 0 & I \end{bmatrix} = I$$

7.4 Block-Gauss-elimination.

Consider a linear system

$$Ax = b$$

Let it be consistently partitioned as

$$\begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix}$$

Let A_{11}, A_{22} be square matrices

$$\left[\begin{array}{c|c} A_{11} & A_{12} \\ \hline A_{21} & A_{22} \end{array}\right] \left[\begin{array}{c} x_1 \\ \hline x_2 \end{array}\right] = \left[\begin{array}{c} b_1 \\ \hline b_2 \end{array}\right]$$

According to the computational rules for partition we may write

$$A_{11}x_1 + A_{12}x_2 = b_1$$
$$A_{21}x_1 + A_{22}x_2 = b_2$$

This looks very similar to 2 equations in 2 unknowns. Let us apply the familiar elimination procedure. We assume that A_{11}^{-1} exists. We premultiply the first equation by $A_{21}A_{11}^{-1}$ and subtract from the 2^{nd}. The result is

$$A_{11}x_1 + A_{12}x_2 = b_1$$
$$0 + (A_{22}-A_{21}A_{11}^{-1}A_{12})x_2 = b_2 - A_{21}A_{11}^{-1}b_1$$

We assume that we may uniquely solve the second set of equations:

$$x_2 = (A_{22}-A_{21}A_{11}^{-1}A_{12})^{-1}(b_2-A_{21}A_{11}^{-1}b_1)$$

We substitute back into the first equation obtaining:

$$A_{11}x_1 = b_1 - A_{12}(A_{22}-A_{21}A_{11}^{-1}A_{12})^{-1}(b_2-A_{21}A_{11}^{-1}b_1)$$
$$x_1 = [A_{11}^{-1} + A_{11}^{-1}A_{12}(A_{22}-A_{21}A_{11}^{-1}A_{12})^{-1}A_{21}A_{11}^{-1}]b_1$$
$$- A_{11}^{-1}A_{12}(A_{22}-A_{21}A_{11}^{-1}A_{12})^{-1}b_2$$

Abbreviating

$$A_{11}^{(-1)} = A_{11}^{-1} + A_{11}^{-1}A_{12}A_{22}^{(-1)}A_{21}A_{11}^{-1}$$
$$A_{12}^{(-1)} = -A_{11}^{-1}A_{12}A_{22}^{(-1)}$$
$$A_{21}^{(-1)} = -A_{22}^{(-1)}A_{21}A_{11}^{-1}$$
$$A_{22}^{(-1)} = (A_{22}-A_{21}A_{11}^{-1}A_{12})^{-1}$$

we get

$$x_1 = A_{11}^{(-1)}b_1 + A_{12}^{(-1)}b_2$$
$$x_2 = A_{21}^{(-1)}b_1 + A_{22}^{(-1)}b_2$$

or

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} A_{11}^{(-1)} & A_{12}^{(-1)} \\ A_{21}^{(-1)} & A_{22}^{(-1)} \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \end{bmatrix}$$

or

$$x = A^{-1}b$$

we have recovered the block-decomposition of the inverse!

7.5 Theoretical background of partitioned matrices.

Let V be a vector space of dimension n and let V_1 and V_2 be vector subspaces of dimension n_1, n_2 with $n = n_1 + n_2$. Assume that the zero vector is the only vector common to V_1 and V_2. As explained in section 5.1, V is the direct sum of V_1 and V_2. The subspaces V_1 and V_2 are not necessarily orthocomplementary.

Any vector $x \in V$ is uniquely decomposed as

$$x = x_1 + x_2, \quad x_1 \in V_1, \quad x_2 \in V_2$$

As pointed out in section 5.1, the two mappings $x \to x_1$ and $x \to x_2$ are linear. We write

$$x_1 = \Pi_1(x)$$
$$x_2 = \Pi_2(x)$$

The Π_1, Π_2 are called projection operators. They are not necessarily orthogonal projectors. Let V_m and V_n be vector spaces. Let

$V_m = V_{m_1} + V_{m_2}$, V_m is the direct sum of V_{m_1} and V_{m_2}

$V_n = V_{n_1} + V_{n_2}$, V_n is the direct sum of V_{n_1} and V_{n_2}

Let Λ be a linear operator $V_m \to V_n$. Let

$$y = \Lambda(x)$$

Decompose

$$x = x_1 + x_2, \qquad x_1 \in V_{m_1},\ x_2 \in V_{m_2}$$
$$y = y_1 + y_2, \qquad y_1 \in V_{n_1},\ y_2 \in V_{n_2}$$

We have

$$y = \Lambda(x)$$

Let Π_1, Π_2 be $x \to x_1$, $x \to x_2$

and R_1, R_2 be $y \to y_1$, $y \to y_2$

Then

$$y_1 = R_1 y = R_1 \circ \Lambda(x) = R_1 \circ \Lambda(\Pi_1(x) + \Pi_2(x))$$

i.e.

$$y_1 = \Lambda_{11}(x_1) + \Lambda_{12}(x_2)$$

similarly:

$$y_2 = \Lambda_{21}(x_1) + \Lambda_{22}(x_2)$$

Note that Λ_{ij} maps V_{m_j} into V_{n_i}. Choose bases as follows

$$
\begin{aligned}
V_{m_1} &: e_1 \quad , \ldots , e_{m_1} \\
V_{m_2} &: e_{m_1+1} , \ldots , e_m \\
V_{n_1} &: f_1 \quad , \ldots , f_{n_1} \\
V_{n_2} &: f_{n_1+1} , \ldots , f_n
\end{aligned}
$$

Let A be the matrix representation of Λ. Partition

$$
A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}
$$

It is easily verified that A_{ij} are the representations of Λ_{ij}. Thus it is seen that the calculus of partitioned matrices relies on two assumptions:

(1) Decomposition of the participating vector spaces into direct sums of subspaces. (The decomposition needs not be orthogonal. In fact, an inner product may not even be defined.)

(2) An appropriate choice of basis vectors: Any subspace must be spanned by a subset of the basis vectors.

8. Isometric mappings between inner product spaces.

8.1. Definitions.

Let V and W be vector spaces and let Λ be a linear operator from V into W:

$$y = \Lambda(x),\ x\epsilon V,\ y\epsilon W$$

The mapping is called isometric if

$$\|\Lambda(x)\| = \|x\|$$

It follows that no vector $x\neq 0$ can be mapped onto the zero vector. If Λ maps V onto W, the mapping is invertible.

8.2. Preservation of inner products.

Isometric mappings preserve the inner product. For let

$$y_1 = \Lambda(x_1),\quad y_2 = \Lambda(x_2)$$

Then $\|y_1-y_2\| = \|x_1-x_2\|$ implies

$$(y_1-y_2,y_1-y_2) = (x_1-x_2,x_1-x_2)$$

or

$$(y_1,y_1) - 2(y_1,y_2) + (y_2,y_2) = (x_1,x_1) - 2(x_1,x_2) + (x_2,x_2)$$

Due to isometry we have $(y_1,y_1) = (x_1,x_1)$, $(y_2,y_2) = (x_2,x_2)$. Hence

$$(y_1, y_2) = (x_1, x_2)$$

8.3 Matrix representation.

Let V and W be of dimension n. Choose bases in V and W. Let the positive definite matrices G_V and G_W represent the inner products in V,W. Let the nxn matrix A represent the operator Λ. Identify vectors with their coordinate n-tuples. We write as usual

$$(x_1, x_2) = x_1^T G_V x_2, \qquad x_1, x_2 \in V$$
$$(y_1, y_2) = y_1^T G_W y_2, \qquad y_1, y_2 \in W$$

The isometry requirement implies for any $x_1, x_2 \in V$:

$$x_1^T G_V x_2 = (Ax_1)^T G_W (Ax_2) = x_1^T A^T G_W A x_2$$

It follows that

$$G_V = A^T G_W A$$

8.4. Examples of isometric mappings.

8.4.1. The isometric mapping between any vector space V and its dual V', the space of linear functionals. We refer to section 2.6. The transformation matrix equals G_V. For if the vector x is mapped onto the functional ξ, we have $\xi = G_V x$. The matrix $G_{V'}$ equals $K_V = G_V^{-1}$.

8.4.2. Change of basis in V. Let $e_1,\ldots,e_n$ be the old basis in V, and G represent the inner product. Let $e'_1,\ldots,e'_n$ be the new basis, and G' be the corresponding representation of the inner product in V. As shown in section 3.2, the relation between old coordinate vectors x and new ones x' is

$$x = Ax'$$

(The j-th column of A contains the scalar factors expressing e'_j in terms of e_i, i=1,...,n.).

Since the inner product is a property of V and not of any basis, it must be preserved during transformation

$$x^TGy = (Ax')^TG(Ay') = x'^T(A^TGA)y' = x'^TG'y'$$

It follows that

$$G' = A^TGA$$

8.4.3. Isometry between a subspace and the space of its parameters. Let V_A be an m-dimensional subspace of the n-dimensional space V_n. Let V_A be spanned by the linearly independent columns of the nxm matrix A. Let the matrix G represent the inner product in V_n. V_A has an inner product inherited from V_n:

$$(a_1,a_2) = a_1^TGa_2$$

Any vector $a \in V_A$ is uniquely represented by its coordinates with respect to the columns of A. These columns may be viewed as a basis of V_A. Thus the system of equations

$$a = Ax$$

has a unique solution x. There is a one to one mapping between V_A and the space X of m-dimensional coordinate vectors x. In order to preserve the inner product, one must require

$$(x_1, x_2) = (a_1, a_2), \quad a_1 = Ax_1, \; a_2 = Ax_2$$

Letting G_X denote the matrix of the inner product in X, one finds

$$(x_1, x_2) = x_1^T G_X x_2^T = (a_1, a_2) = x_1^T A^T G A x_2$$

Thus

$$G_X = A^T G A$$

This looks the same as before; however this time A is not invertible.

8.5. Canonical transformation of an adjustment problem.

We consider adjustment by variation of parameters:

$$\ell + v = Ax \quad \text{with weight matrix } P$$

We denote by L the space of observations and by L_A that of the adjusted observations. We choose an orthonormal basis in L_A. Let the matrix $\bar{A}$ comprise the new orthonormal vectors. Confer section 4.7. The columns of $\bar{A}$ are expressible in terms of the columns of A:

$$\bar{A} = AC$$

We introduce new parameters

$$x = Cy$$

The new adjustment problem is

$$\ell+v = ACy = \bar{A}y$$

Consider the matrix $\bar{B}$ having orthonormal columns and spanning the orthocomplement of V_A. We have

$$\bar{A}^T P \bar{A} = I,\ \bar{B}^T P \bar{B} = I,\ \bar{A}^T P \bar{B} = 0$$

Consider an isometric transformation to new observations ℓ' given by

$$\begin{bmatrix} \ell_1' \\ \ell_2' \end{bmatrix} = \begin{bmatrix} \bar{A}^T P \\ \bar{B}^T P \end{bmatrix} \ell$$

The matrices

$$\begin{bmatrix} \bar{A}^T P \\ \bar{B}^T P \end{bmatrix}$$

and

$$(\bar{A}, \bar{B})$$

are inverse to each other. This is easily seen by multiplying these two partitioned matrices and minding the orthogonality relations between $\bar{A}$ and $\bar{B}$.

Premultiply the adjustment problem by

$$\begin{bmatrix} \bar{A}^T P \\ \bar{B}^T P \end{bmatrix}$$

To obtain

$$\ell_1' + v_1' = Iy$$
$$\ell_2' + v_2' = 0$$

The weight matrix of the new observations ℓ' is obtained as

$$P' = \begin{bmatrix} \bar{A}^T \\ \bar{B}^T \end{bmatrix} P \quad (\bar{A}, \bar{B}) = I$$

The solution of the canonically transformed adjustment problem is obviously

$$\ell_1' = y, \quad v_1' = 0$$

$$\ell_2' = 0, \quad v_2' = -\ell_2'$$

Remark: Deriving the canonical form requires no less computational work than solving the adjustment problem conventionally. The benefit is theoretical insight.

9. Partial reduction.

9.1 Partitioning the set of parameters.

Consider an adjustment problem by variation of parameters

$$\ell+v = Ax, \qquad \text{weight matrix } P$$

Assume A being an n×m matrix. The m-dimensional vector of parameters x is partitioned into an m_1-dimensional vector x_1 and an m_2-dimensional vector x_2:

$$x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

Of course, $m_1 + m_2 = m$. Partition the columns of A accordingly

$$A = (A_1, A_2)$$

The adjustment problem is then written as

$$\ell+v = (A_1, A_2) \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

or

$$\ell+v = A_1x_1 + A_2x_2$$

We are primarily interested in adjusted values of x_2. The parameters x_1 play an auxiliarly role, as for example orientation unknowns.

9.2 Partial reduction of the normal equations.

The normal equations are

$$(A^T PA)x = A^T P\ell$$

or briefly

$$Gx = r$$

The partitioning of A induces a partitioning of G and r

$$\begin{bmatrix} A_1^T PA_1 & A_1^T PA_2 \\ A_2^T PA_1 & A_2^T PA_2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} A_1^T P\ell \\ A_2^T P\ell \end{bmatrix}$$

We abbreviate this as

$$\begin{bmatrix} G_{11} & G_{12} \\ G_{21} & G_{22} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} r_1 \\ r_2 \end{bmatrix}$$

This may also be written

$$G_{11}x_1 + G_{12}x_2 = r_1$$
$$G_{21}x_1 + G_{22}x_2 = r_2$$

We may eliminate x_1 from these equations in the same way as this was done

in section 7.4. We obtain the partially reduced set of normal equations for x_2:

$$(G_{22}-G_{21}G_{11}^{-1}G_{12})x_2 = r_2 - G_{21}G_{11}^{-1}r_1$$

This is abbreviated as

$$\bar{G}_{22}x_2 = \bar{r}_2$$

Our intention is to understand these equations geometrically.

9.3. Orthocomplementary decomposition of the space of adjusted observations and its parameter space.

Let L denote the space of observations, and let L_A be the space of adjusted observations. L_A is spanned by the columns of A. We now decompose L_A into L_{A_1}, the space spanned by the columns of A_1, and into $L_{\bar{A}_2}$. $L_{\bar{A}_2}$ is the orthocomplement of L_{A_1} in L_A. It is spanned by the columns of a matrix $\bar{A}_2$ which is yet to be determined. The following relations must hold

$$(A_1, \bar{A}_2) \text{ span } L_A$$
$$A_1^T P \bar{A}_2 = 0$$

We use the isometry between L_A and its parameter space X. Confer section 8.4.3. The columns of A are mapped onto the natural basis of X. The inner product in X is represented by

$$G = A^T P A$$

Any vector x in X is represented as

$$x = \begin{bmatrix} I & 0 \\ 0 & I \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} I \\ 0 \end{bmatrix} x_1 + \begin{bmatrix} 0 \\ I \end{bmatrix} x_2$$

i.e.

$$x = J_1 x_1 + J_2 x_2$$

X is the direct sum of two subspaces X_1 and X_2 spanned by the columns of J_1, J_2. These subspaces are not orthogonal. We have

$$J_i^T G J_j = G_{ij}, \quad i,j=1,2$$

We replace J_2 by $\bar{J}_2$ orthogonal to J_1. We proceed formally in a similar fashion as in section 4.8 on Gram-Schmid orthogonalization. Just confer the way, the second vector $\bar{a}_2$ was derived there. We represent:

$$\bar{J}_2 = J_2 - J_1 C$$

where the matrix C is yet to be determined. Requiring $J_1^T G \bar{J}_2 = 0$ leads to

$$0 = G_{12} - G_{11} C, \quad \text{i.e.} \quad C = G_{11}^{-1} G_{12}$$

Thus

$$\begin{aligned} \bar{J}_2 &= J_2 - J_1 G_{11}^{-1} G_{12} = \\ &= J_2 - J_1 (J_1^T G J_1)^{-1} J_1^T G J_2 \\ &= J_2 - P_{J_1} J_2 = (I - P_{J_1}) J_2 \end{aligned}$$

We view the columns of $(J_1, \bar{J}_2)$ as a new basis in X. A vector x is represented as

$$x = J_1 y_1 + \bar{J}_2 y_2$$

Inserting for $\bar{J}_2$ we get

$$x = (J_1 y_1 - P_{J_1} J_2 y_2) + J_2 y_2$$

The vector in parentheses is in X_1, (spanned by J_1), the second vector on the right hand side is in X_2 (spanned by J_2). A comparision with the earlier representation $x = J_1 x_1 + J_2 x_2$ shows:

$$J_1 x_1 = J_1 y_1 - J_1 G_{11}^{-1} G_{12} y_2$$
$$J_2 x_2 = J_2 y_2$$

Because the columns of J_1 and J_2 are linearly independent, we get

$$x_1 = y_1 - G_{11}^{-1} G_{12} y_2$$
$$x_2 = y_2$$

This expresses the old coordinates of the vector x in terms of the new ones. The orthocomplementary decomposition of X into spaces spanned by J_1, $\bar{J}_2$ induces an orthocomplementary decomposition of L_A into spaces spanned by A_1, $\bar{A}_2$. A vector $a \in L_A$ was previously represented as

$$a = A_1 x_1 + A_2 x_2$$

now it is represented by

$$a = A_1 y_1 + \bar{A}_2 y_2$$

The new representation is obtained either by substituting for x_1, x_2:

$$\begin{aligned} a &= A_1(y_1 - G_{11}^{-1} G_{12} y_2) + A_2 y_2 \\ &= A_1 y_1 + (A_2 - A_1 G_{11}^{-1} G_{12}) y_2 \\ &= A_1 y_1 + (A_2 - A_1 (A_1^T P A_1)^{-1} A_1^T P A_2) y_2 \\ &= A_1 y_1 + (I - P_{A_1}) A_2 y_2 \\ &= A_1 y_1 + \bar{A}_2 y_2 \end{aligned}$$

or by noting that the dependence of $\bar{A}_2$ on A_1, A_2 must be the same as that of $\bar{J}_2$ on J_1, J_2:

$$\bar{A}_2 = A_2 - A_1 G_{11}^{-1} G_{12} = A_2 - A_1 (A_1^T P A_1)^{-1} A_1^T P A_2 = (I - P_{A_1}) A_2$$

In any case, the desired matrix $\bar{A}_2$ is obtained as

$$\bar{A}_2 = (I - P_{A_1}) A_2$$

Our adjustment problem is thus transformed into

$$\ell+v = A_1 y_1 + \bar{A}_2 y_2 = A_1 y_1 + \bar{A}_2 x_2$$

because $y_2 = x_2$.

Verify that $\bar{A}_2^T P \bar{A}_2 = G_{22} - G_{21} G_{11}^{-1} G_{12} = \bar{G}_{22}$, $\bar{A}_2^T P \ell = r_2 - G_{21} G_{11}^{-1} r_1 = \bar{r}_2$. Hence the transformed normals are found to be:

$$G_{11} y_1 = r_1$$
$$\bar{G}_{22} x_2 = \bar{r}_2$$

they decompose into two independent sets. The second one is identical to the partially reduced normals for x_2.

The partially reduced normals give x_2. The question remains how to find the residuals v without calculating y_1 from the first set of the above equations.

9.4. The partially reduced observation equations.

If we did solve the complete set of transformed normals, we would get v from

$$\ell+v = A_1 y_1 + \bar{A}_2 x_2$$

Here

$$y_1 = G_{11}^{-1} r_1 = (A_1^T P A_1)^{-1} A_1^T P \ell$$

We see that

$$\ell - A_1(A_1^T P A_1)^{-1} A_1^T P \ell + v = \bar{A}_2 x_2$$

or

$$\ell - P_{A_1}\ell + v = \bar{A}_2 x_2$$
$$(I - P_{A_1})\ell + v = \bar{A}_2 x_2$$
$$\bar{\ell} + v = \bar{A}_2 x_2$$

The last set is called partially reduced observation equations. They involve the pseudo observations

$$\bar{\ell} = (I - P_{A_1})\ell$$

It is important to note that the normals obtained from the partially reduced observation equations are just the partially reduced normals for x_2.

$$(\bar{A}_2^T P \bar{A}_2)x_2 = \bar{A}_2^T P \bar{\ell} = \bar{A}_2^T P(\ell - P_{A_1}\ell) = \bar{A}_2^T P \ell \quad \text{i.e. } \bar{G}_{22} x_2 = \bar{r}_2$$

(Mind that $\bar{A}_2$ is orthogonal to $P_{A_1}\ell$)

9.5. Alternative derivation of the partially reduced observation equations.

Consider the orthocomplementary decomposition of L into 3 subspaces L_{A_1}, $L_{\bar{A}_2}$, L_B. The spaces L_{A_1}, $L_{\bar{A}_2}$ are already familiar. They are spanned by the columns of the matrices A_1, $\bar{A}_2$. The space L_B is the orthocomplement of the direct sum of these spaces. It is also the orthocomplement of L_A, $A = (A_1, A_2)$. The space L_B is spanned by the columns of the matrix B. It holds that

$$A_1^T P \bar{A}_2 = 0, \quad A_1^T P B = 0, \quad \bar{A}_2^T P B = 0$$

Let a new basis of L be given by the union of the columns of A_1, $\bar{A}_2$, B. We transform the observations ℓ to the new basis:

$$\ell = (A_1, \bar{A}_2, B) \begin{bmatrix} \ell_1' \\ \ell_2' \\ \ell_3' \end{bmatrix}$$

Because the subspaces L_{A_1}, $L_{\bar{A}_2}$, L_B are orthogonal, it also holds that

$$A_1\ell_1 = P_{A_1}\ell, \quad \bar{A}_2\ell_2 = P_{\bar{A}_2}\ell, \quad B\ell_3 = P_B\ell$$

Inserting the matrix representations of the projectors (e.g. $P_{A_1} = A_1(A_1^T P A_1)^{-1} A_1^T P$) one finds the formula expressing ℓ' in terms of ℓ:

$$\begin{bmatrix} \ell_1' \\ \ell_2' \\ \ell_3' \end{bmatrix} = \begin{bmatrix} (A_1^T P A_1)^{-1} A_1^T P \\ (\bar{A}_2^T P \bar{A}_2)^{-1} \bar{A}_2^T P \\ (B^T P B)^{-1} B^T P \end{bmatrix} \ell$$

The weight matrix of the new observations is the representation of P with respect to the new basis. One finds

$$\begin{bmatrix} A_1^T P A_1 & 0 & 0 \\ 0 & \bar{A}_2^T P \bar{A}_2 & 0 \\ 0 & 0 & B^T P B \end{bmatrix} = \begin{bmatrix} G_{11} & 0 & 0 \\ 0 & G_{22} & 0 \\ 0 & 0 & B^T P B \end{bmatrix}$$

It is important to note the blockdiagonal structure. The observation equations transform as:

$$\ell_1' + v_1' = y_1$$
$$\ell_2' + v_2' = x_2$$
$$\ell_3' + v_3' = 0$$

The 3 subproblems for ℓ_1', ℓ_2', ℓ_3' are completely independent. Any subproblem has separate observations, corrections and unknowns. Essential is also the block diagonal structure of the transformed weight matrix. There is no coupling of the subproblems due to weights. One immediately obtains the solution

$$y_1 = \ell_1', \quad v_1' = 0$$
$$x_2 = \ell_2', \quad v_2' = 0$$
$$v_3' = -\ell_3'$$

Because the 3 subproblems are independent, the result for x_2 is unaffected if we put

$$\ell_1' = 0, \quad y_1 = 0$$

Hence the problem

$$0 + v_1' = 0$$
$$\ell_2' + v_2' = x_2$$
$$\ell_3' + v_3' = 0$$

yields correct results for x_2 and v_1', v_2', v_3'. Transforming this problem backward to the old basis, one obtains

$$A_1 \ell_1' + B\ell_3' + v = \bar{A}_2 x_2$$

or

$$P_{A_1} \ell + P_B \ell + v = \bar{A}_2 x_2$$

or

$$(I - P_{\bar{A}_2})\ell + v = \bar{A}_2 x_2$$

These are the partially reduced observation equations.

10. Adjustment phased with respect to observations.

10.1 Formulation of the problem.

Consider an adjustment problem by variation of parameters.

$$\ell + v = Ax, \quad \text{weight matrix } P$$

Decompose the vector ℓ into two subvectors. Decompose v and A, accordingly

$$\ell = \begin{bmatrix} \ell_1 \\ \ell_2 \end{bmatrix} \quad v = \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} \quad A = \begin{bmatrix} A_1 \\ A_2 \end{bmatrix}$$

Assume that the weight matrix P has the following special structure

$$P = \begin{bmatrix} P_{11} & 0 \\ 0 & P_{22} \end{bmatrix}$$

There are no weights coupling the two groups of observations.

One may imagine that the two groups of observations refer to different time periods. It could also be that they refer to different geographical regions with partial overlap. In any case, one can consider the two separate adjustment problems

$$\ell_1 + v_1 = A_1 x \quad \text{weight matrix } P_{11}$$
$$\ell_2 + v_2 = A_2 x \quad \text{weight matrix } P_{22}$$

and one can obtain their separate solutions $x_{(1)}$, $x_{(2)}$. The question is, how the solution of the entire problem is related to these partial solutions.

One can also proceed differently. Suppose that the observations ℓ_1 are available earlier. Then one calculates $x_{(1)}$ from the first set of the above relations. This is phase 1 of the adjustment. Subsequently observations ℓ_2 become available. One is then interested to calculate in the second phase the solution x of the entire problem by using $x_{(1)}$, the solution of the previous phase in combination with the observations ℓ_2 of the second phase. Of course, one can consider more than two phases. However the essential features of phased adjustment become transparent if only two phases are considered.

10.2 Addition of normal equations.

The normal equations of the two separate problems are

$$(A_1^T P_{11} A_1)x = A_1^T P_{11} \ell_1$$
$$(A_2^T P_{22} A_2)x = A_2^T P_{22} \ell_2$$

The normal equations of the entire problem are

$$\begin{bmatrix} A_1 \\ A_2 \end{bmatrix}^T \begin{bmatrix} P_{11} & 0 \\ 0 & P_{22} \end{bmatrix} \begin{bmatrix} A_1 \\ A_2 \end{bmatrix} x = \begin{bmatrix} A_1 \\ A_2 \end{bmatrix}^T \begin{bmatrix} P_{11} & 0 \\ 0 & P_{22} \end{bmatrix} \begin{bmatrix} \ell_1 \\ \ell_2 \end{bmatrix}$$

This is evaluated as

$$(A_1^T P_{11} A_1 + A_2^T P_{22} A_2)x = (A_1^T P_{11} \ell_1 + A_2^T P_{22} \ell_2)$$

It is seen that the normals of the entire problem are obtained by adding the normals of the two phases. The added normals can be solved to give the solution x.

10.3 Updating the solution of the previous phase.

Consider the problem of the first phase

$$\ell_1 + v_1 = A_1 x$$

Its solution is

$$x_{(1)} = (A_1^T P_{11} A_1)^{-1} A_1^T P_{11} \ell_1$$

Let L_1 be the space of observations ℓ_1. We consider the subspace L_{A_1} of L_1 spanned by the columns of A_1. We decompose L_1 into orthocomplementary subspaces L_{A_1} and L_{B_1}. The space L_{B_1} is spanned by columns of B_1. The residuals of phase 1 are

$$v_{(1)} = A_1 x_{(1)} - \ell_1$$

As shown in section 6.1, $v_{(1)}$ may be represented as

$$v_{(1)} = -B(B_1^T P_{11} B_1)^{-1} B_1^T P_{11} \ell_1$$

The second representation for $v_{(1)}$ will only be needed for the purpose of mathematical proofs.

Because L_{B_1} is the orthocomplement of L_{A_1}, we have

$$A_1^T P_{11} B_1 = 0$$

Consider now a transformation of the observations of phase 1.

$$\ell'_{11} = (A_1^T P_{11} A_1)^{-1} A_1^T P_{11} \ell_1 = x_{(1)}$$
$$\ell'_{12} = (B_1^T P_{11} B_1)^{-1} B_1^T P_{11} \ell_2$$

Note that the two matrices

$$\begin{bmatrix} (A_1^T P_{11} A_1)^{-1} A_1^T P_{11} \\ (B_1^T P_{11} B_1)^{-1} B_1^T P_{11} \end{bmatrix} \quad \text{and } (A_1, B_1)$$

are inverse to each other. (Multiply them to obtain the unit matrix I). Hence the transformation is equivalent to

$$\begin{bmatrix} \ell_1 \\ \ell_2 \end{bmatrix} = (A_1, B_1) \begin{bmatrix} \ell'_{11} \\ \ell'_{12} \end{bmatrix}$$

The weight matrix of the transformed observations is given by

$$P' = (A_1, B_1)^T P (A_1, B_1) = \begin{bmatrix} A_1^T P A_1 & 0 \\ 0 & B_1^T P B_1 \end{bmatrix}$$

The transformed adjustment problem is

$$\ell'_{11} + v'_{11} = x$$
$$\ell'_{12} + v'_{12} = 0$$

Equivalently

$$x_{(1)} + v'_{11} = x, \quad \text{weight matrix} \quad A_1^T P_{11} A_1$$
$$\ell'_{12} + v'_{12} = 0, \quad \text{weight matrix} \quad B_1^T P_{11} B_1$$

The two problems are independent, because the weight matrix is block diagonal. All the information on x available from phase 1 is contained in the first set

$$x_{(1)} + v'_{11} = x \quad , \quad \text{weight matrix} \quad A_1^T P_{11} A_1$$

We add the observation equations of the second phase, arriving at the problem

$$x_{(1)} + v'_{11} = x$$
$$\ell_2 + v_2 = A_2 x$$

with weight matrix

$$\begin{bmatrix} A_1^T P_{11} A_1 & 0 \\ 0 & P_{22} \end{bmatrix}$$

Forming the normals gives precisely the normals of the entire problem as they were obtained earlier:

$$(A_1^T P_{11} A_1 + A_2^T P_{22} A_2)x = (A_1^T P_{11} \ell_1 + A_2^T P_{22} \ell_2)$$

(The calculation of the normals is easy and is omitted.) The solution x is calculated. One calculates residuals of the second phase.

$$v_{11}' = x - x_{(1)}$$
$$v_2 = A_2 x - \ell_2$$

The following relationship between residuals $v_{(1)}$, v_{11}' and the residuals

$$v = \begin{bmatrix} v_1 \\ v_2 \end{bmatrix}$$

of the original problem is interesting.

Theorem:

$$v^T P v = v_1^T P_{11} v_1 + v_2^T P_{22} v_2 =$$
$$= v_{(1)}^T P_{11} v_{(1)} + v_{11}'^T A_1^T P_{11} A_1 v_{11}' + v_2^T P_{22} v_2$$

Equivalently

$$v^T P v = (\ell_1 - Ax_{(1)})^T P_{11} (\ell_1 - Ax_{(1)}) +$$
$$+ (x - x_{(1)})^T A_1^T P_{11} A_1 (x - x_{(1)}) +$$
$$+ (\ell_2 - A_2 x)^T P_{22} (\ell_2 - A_2 x)$$

The first term on the right hand side is the weighted sum of residuals obtained from phase 1. The second and third term comprise the weighted sum of residuals from phase 2. We thus obtain:

"The weighted sum of residuals of the combined phases is the sum of the weighted sums of residuals of the individual phases."

Proof: The theorem of Pythagoras applied to the entire problem and to the individual phases gives the following three relations. (Confer section 6.3.)

$$v_1^T P_{11} v_1 + v_2^T P_{22} v_2 = \ell_1^T P_{11} \ell_1 + \ell_2^T P_{22} \ell_2 - x^T A_1^T P_{11} A_1 x - x^T A_2^T P_{22} A_2 x$$
$$v_1^T P_{11} v_1 = \ell_1^T P_{11} \ell_1 - x_{(1)}^T A_1^T P_{11} A_1 x_{(1)}$$
$$v_{11}^T A_1^T P_{11} A_1 v_{11} + v_2^T P_{22} v_2 = x_{(1)}^T A_1^T P_{11} A_1 x_{(1)} +$$
$$+ \ell_2^T P_{22} \ell_2 - x^T A_1^T P_{11} A_1 x - x^T A_2^T P_{22} A_2 x$$

It is seen that the first relation is the sum of the second and third. This proves the theorem.

10.4 Geometrical insight.

The space L of observations is represented as the direct sum of two orthocomplementary subspaces L_1, L_2. Bases in L_1 and L_2 are chosen and

subsequently combined to a basis of L. This allows us to use the calculus of partitioned matrices as outlined in section 7.5. The matrix of the inner product necessarily decomposes into blockdiagonal form.

$$P = \begin{bmatrix} P_1 & 0 \\ 0 & P_2 \end{bmatrix}$$

because any vector in L_1 is orthogonal to any vector in L_2.

We consider the space L_A, spanned by the columns of A_1. It is a subspace of L_1. The subspace L_1 is viewed as the direct sum of L_{A_1} and its orthocomplement L_{B_1} in L_1. A change of basis in L_1 is performed such that the columns of A_1 and B_1 become basis vectors. This induces the transformation $\ell_1 \rightarrow \ell_1'$. The entire space of observations L is now the direct sum of three orthocomplementary subspaces L_{A_1}, L_{B_1}, L_2. The space of adjusted observations L_A is only participating in L_{A1}, L_2. Only the zero vector is common to L_A and L_{B_1}. Hence the space L_{B_1} may be ignored in the adjustment problem. This leads to the simplified setup of the second phase.

$$\begin{aligned} x_{(1)} + v_{11} &= x \\ \ell_2 + v_2 &= A_2 x \end{aligned}$$

with weight matrix

$$\begin{bmatrix} A_1^T P A_1 & 0 \\ 0 & P_{22} \end{bmatrix}$$

10.5 Pre-elimination of group-internal unknowns.

Phased adjustment as outlined above is not very effective from the viewpoint of computational efficiency. It becomes a powerful tool if it is combined with partial reduction as presented in chapter 9. A great benefit arises if there are sets of auxiliary unknowns, each one referring to only one group of observations. If auxiliary unknowns y_i are only present in group i, we call them "group-internal" unknowns. It suffices to consider the setup

$$\begin{aligned} \ell_1 + v_1 &= H_1 h_1 + A_1 x \\ \ell_2 + v_2 &= H_2 h_2 + A_2 x \end{aligned} \qquad \text{weight matrix } P = \begin{bmatrix} P_{11} & 0 \\ 0 & P_{22} \end{bmatrix}$$

Here h_1 are auxiliary unknowns which are internal to group 1. The unknowns h_2 are internal to group 2. H_1 and H_2 are the corresponding design matrices.

A remarkable simplification of the computation results if the group internal unknowns are eliminated before the groups are combined.

As outlined in chapter 9., the elimination can be accomplished in two different ways. If

$$\ell_i + v_i = H_i h_i + A_i x$$

are the observation equations, one may either form the partially reduced observation equations

$$\bar{\ell}_i + v_i = \bar{A}_i x, \quad \bar{A}_i = A_i - P_{H_i} A_i$$

which lead to the partially reduced normals

$$\bar{A}_i^T P_{ii} \bar{A}_i x = \bar{A}_i^T P_{ii} \ell$$

or one may form the normals for a group

$$\begin{bmatrix} H_i^T P_{ii} H_i & H_i^T P_{ii} A_i \\ A_i^T P_{ii} H_i & A_i^T P_{ii} A_i \end{bmatrix} \begin{bmatrix} h_i \\ x \end{bmatrix} = \begin{bmatrix} H_i^T P \ell \\ A_i^T P \ell \end{bmatrix}$$

and eliminate the auxiliary unknowns h_i. The result will be the same set of partially reduced normals, although the immediately derived expression looks differently, namely as

$$(A_i^T P_{ii} A_i - A_i^T P_{ii} H_i (H_i^T P_{ii} H_i)^{-1} H_i^T P_{ii} A_i) x =$$
$$A_i^T P \ell - A_i^T P_{ii} H_i (H_i^T P_{ii} H_i)^{-1} H_i^T P \ell, \qquad i=1,2$$

In any case, the partially reduced normals

$$\bar{G}_{ii} x = \bar{r}_i$$

of all groups may be added to give the partially reduced normals of the entire system

$$(\bar{G}_{11} + \bar{G}_{22}) x = \bar{r}_1 + \bar{r}_2$$

A proof for the validity of the procedure of group-wise elimination of group-internal unknowns can certainly be given in geometrical terms. Occasionally, however, it is preferable to use calculus. We have to show that the same result is obtained if partial reduction is done in the conventional way, i.e. by ignoring the block decomposition of our system resulting from decomposing ℓ into ℓ_1 and ℓ_2. Write the observation equations as

$$\ell + v = Hh + Ax, \quad \text{weight matrix } P$$

whereby

$$H = \begin{bmatrix} H_1 & 0 \\ 0 & H_2 \end{bmatrix}, \quad A = \begin{bmatrix} A_1 \\ A_2 \end{bmatrix}, \quad P = \begin{bmatrix} P_{11} & 0 \\ 0 & P_{22} \end{bmatrix}$$

In agreement with section 9.5 we obtain the partially reduced observation equations as

$$\bar{\ell} + v = \bar{A}x, \quad \text{weight matrix } P$$

with

$$\bar{\ell} = (I-P_H)\ell, \quad \bar{A} = (I-P_H)A$$

Forming P_H, one recognizes

$$P_H = \begin{bmatrix} P_{H_1} & 0 \\ 0 & P_{H_2} \end{bmatrix}, \qquad P_{H_i} = H_i(H_i^T P_{ii} H_i)^{-1} H_i^T P_{ii}, \qquad i=1,2$$

Hence

$$\bar{\ell} = \begin{bmatrix} \bar{\ell}_1 \\ \bar{\ell}_2 \end{bmatrix}, \quad \text{with} \quad \bar{\ell}_i = (I-P_{H_i})\ell_i,$$

$$\bar{A} = \begin{bmatrix} \bar{A}_1 \\ \bar{A}_2 \end{bmatrix}, \quad \text{with} \quad \bar{A}_i = (I-P_{H_i})A_i$$

These are precisely the quantities occurring in the partially reduced observation equations obtained by considering the two groups separately.

This completes the proof for the case of group-wise partially reduced observation equations. The proof for the validity of the group-wise partially reduced normals is even simpler. It is omitted.

Remark: A geometric proof would start from the afore mentioned decomposition of the vector space L into a direct sum of orthocomplementary subspaces L_1 and L_2. In analogy to section 9.6 each subspace L_i, i=1,2, is further decomposed into 3 orthocomplementary spaces L_{H_i}, $L_{\bar{A}_i}$ and L_{B_i}. One considers L_H as the direct sum of L_{H_1} and L_{H_2}. The projector Π_H onto L_H is represented as

$$\Pi_H = \Pi_{H_1} + \Pi_{H_2}$$

Because $L_{H_i} \subset L_i$, i=1,2, this may also be written

$$\Pi_H = \Pi_{L_1} \circ \Pi_{H_1} \circ \Pi_{L_1} + \Pi_{L_2} \circ \Pi_{H_2} \circ \Pi_{L_2}$$

Using this together with

$$\bar{\ell} = (I-\Pi_H)\ell, \quad \ell_i = \Pi_{L_i}\ell, \quad \bar{\ell}_i = \Pi_{L_i}\bar{\ell}$$

one gets

$$\bar{\ell}_i = \Pi_{L_i} \circ (I-\Pi_{H_i})\ell_i$$

In agreement with section 7.5, one recognizes that the operators $\Pi_{L_i} \circ (I-\Pi_{H_i})$ map L_i into L_i, i=1,2. They map L_j, j≠i, onto zero. With respect to the basis of L_i, those operators are represented by the matrices

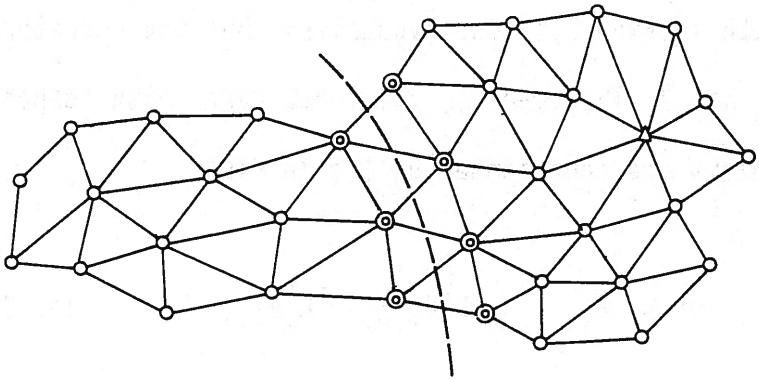
$$I-\Pi_{H_i}, \quad \text{with} \quad \Pi_{H_i} = H_i(H_i^T P_{ii} H_i)^{-1} H_i^T P_{ii}, \qquad i=1,2$$

encountered earlier.

10.6 Helmert blocking.

The procedure of section 10.5 is the theoretical basis for Helmert blocking. As an example take a network as depicted in the subsequent figure. Let ℓ_1 denote observations taken at stations to the east of the dashed line. Let ℓ_2 denote the observations taken at stations to the right. Let h_1 comprise coordinate increments of stations marked by single circles and situated to the east of the dividing line. Such stations are called inner stations of the eastern block. Include in h_1 also orientation unknowns of observations out of ℓ_1. Define h_2

accordingly. Finally let x denote coordinate increments of junction stations marked by double circles. The procedure of section 10.5 gives rigorously adjusted values of the junction station coordinate increments.



11. Complementary extremum principles in least squares adjustment.

11.1 The basic geometric principle.

Let V be an n-dimensional vector space equipped with an inner product. Let V_A, V_B be orthocomplementary subspaces. The following theorem is a near triviality.

Theorem: Let $a \in V_A$. The vector $b \in V_B$ closest to a is the zero vector b=0.

Proof: $d(a,b)^2 = \|a-b\|^2 = (a-b,a-b) = (a,a)+(b,b)$, because $(a,b) = 0$, for $a \in V_A$, $b \in V_B$. Thus $d(a,b) = \|a\|^2 + \|b\|^2$. Obviously this is minimal for b=0.

11.2 Reformulation for linear manifolds.

We shift the problem slightly away from triviality by considering linear manifolds instead of subspaces.

Definition: Let V_A be a subspace and let a_0 be any fixed vector in V. The linear manifold M_A comprises all vectors u which may be represented as

$$u = a_0 + a, \quad a \in V_A$$

Similarly, the linear manifold M_B is introduced. It comprises all vectors representable as

$$v = b_0 + b, \quad b \in V_B$$

b_0 is again a fixed vector in V. It is seen that a linear manifold is generally not a subspace. The zero vector may not be a member of M_A or M_B. However,

difference vectors of the vectors in a linear manifold form a vector-subspace.

Next we show that the linear manifolds M_A and M_B have only one vector in common if the participating subspaces are orthocomplementary. Let

$$w = a_0 + a = b_0 + b, \quad a \in V_A, \; b \in V_B$$

It follows that

$$b_0 - a_0 = a - b, \quad a \in V_A, \; b \in V_B$$

The decomposition of $b_0 - a_0$ into vectors of V_A and V_B is unique. This shows existence and uniqueness of w. It also shows that

$$a = \Pi_A(b_0 - a_0)$$
$$b = \Pi_B(a_0 - b_0)$$

Here Π_A and Π_B are the (orthogonal) projection operators onto the subspaces V_A, V_B.

The translation

$$x' = x + w$$

carries the linear subspaces V_A, V_B over into the manifolds M_A, M_B. Because distances are translation-invariant, the theorem of section 11.1 is reformulated

as follows.

Theorem: Let $u \in M_A$. Then the vector $v \in M_B$ closest to u is the vector w, representing the intersection of M_A and M_B.

The roles of M_A and M_B may be interchanged. Thus it is seen that w is the solution of two extremum problems:

(I) Given $v \in M_B$, find $u \in M_A$ such that

$$d(u,v)^2 = \text{Minimum}$$

(II) Given $u \in M_A$, find $v \in M_B$ such that

$$d(u,v)^2 = \text{Minimum}$$

The two extremum problems have different admissible sets, namely M_A and M_B. The only vector common to both admissible sets solves both extremum problems.

Suppose that u' is admissible for (I), but not necessarily optimal. Similarly let v' be admissible for (II). Then

$$d(v,w)^2 \leq d(v,u')^2$$
$$d(u,w)^2 \leq d(u,v')^2$$

By the theorem by Pythagoras

$$d(u,v)^2 = d(u,w)^2 + d(v,w)^2$$

Using the second of the above inequalities, one finds

$$d(v,w)^2 \geq d(u,v)^2 - d(u,v')^2$$

Combining with the first of the above inequalities one gets a lower and an upper bound on the optimum value of (I):

$$d(u,v)^2 - d(u,v')^2 \leq d(v,w)^2 \leq d(v,u')^2$$

Similarly, an inclusion of the extremum of (II) is obtained. This is also obvious if one notes that both extrema sum up to $d(u,v)^2$.

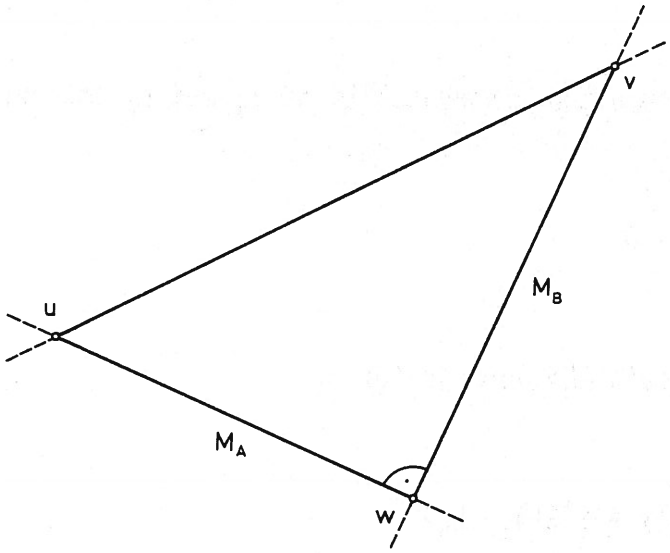
Remark: In the literature, the second problem is frequently posed as follows:
(II*) Given $u \in M_A$ and $v \in M_B$, find $w \in M_B$ such that

$$d(u,v)^2 - d(u,w)^2 = \text{Maximum}$$

This redefinition causes the two optima to coincide. ("The energy equals the complementary energy"). Besides this, the second problem now searches for a maximum. The solution is therefore a minimum for (I) and a maximum for (II*). It is a saddle point. It is also more obvious now, why admissible vectors for (I)

and (II*) give upper and lower bounds on the optimum. On the other hand, a certain lack of symmetry in (I) and (II*) is noted. Therefore we prefer our original setup.

Remark: In 2 dimensions the complementary extremum principles and their solutions can be read off from a rectangular triangle



The points u, v form the end points of the hypotenuse. Problem (I) searches for a point on the straight line containing the short side from u to w. The point shall be nearest to v. Obviously the solution is w. Problem (II) is obtained by symmetry.

Remark: If the subspace V_A is spanned by the linearly independent columns of the matrix A, and if V_B is spanned similarly by the columns of B, then w is represented as

$$w = a_0 + Ax = b_0 + By$$

we get

$$Ax - By = b_0 - a_0$$

Premultiplying by A^TG, where G represents the inner product, we obtain the normal equations for x:

$$(A^TGA)x = A^TG(b_0 - a_0)$$

This is so, because the orthogonality of V_A and V_B implies

$$A^TGB = 0$$

Similarly we obtain the normals for y

$$(B^TGB)y = B^TG(a_0 - b_0)$$

11.3 Adjustment by minimizing the norm of the residuals.

Let L be the vector space of observations $\underline{l}$. Let L_A, L_B be the orthocomplementary subspaces denoted V_A, V_B earlier. Let M_A be the manifold

$$M_A = \{\underline{l} \in L \mid \underline{l} = a_0 + a,\ a \in L_A\}$$

Here a_0 is a fixed vector in L. M_A is called the manifold of adjusted observations. An observation vector in M_A fulfills a number of geometrical or physical constraints, such as triangle closures or equations of motion.

The familiar problem of least squares adjustment searches for a vector $\tilde{\ell} \in M_A$ of adjusted observations. $\tilde{\ell}$ is chosen in a way that the (squared) norm of the residual vector

$$v = \tilde{\ell} - \ell$$

is minimized. Thus we have problem

(I) Given ℓ, find $\tilde{\ell} \in M_A$ such that

$$d(\ell,\tilde{\ell})^2 = \text{Minimum}$$

In order to formulate the complementary problem, we need a manifold M_B whose participating subspace is L_B, the orthocomplement of L_A. The vector b_0 must be chosen in a way that $\ell \in M_B$. The simplest choice is

$$b_0 = \ell$$

One could deviate from this simplest choice and reformulate problem (I) accordingly. However, we just take $b_0=\ell$. We now have problem

(II) Given a_0, find $\tilde{\ell} \in M_B$ such that

$$d(a_0,\tilde{\ell})^2 = \text{Minimum}$$

As we know, both problems have identical solutions. They are obtained as

$$\tilde{\ell} = a_0 + \Pi_A(\ell - a_0)$$
$$\tilde{\ell} = \ell + \Pi_B(a_0 - \ell)$$

If A, B are the matrices whose columns space L_A and L_B, and if the inner product in L is represented by the weight matrix P, then we may proceed as follows

$$\tilde{\ell} = a_0 + Ax$$
$$\tilde{\ell} = \ell + Bk$$

Normal equations

$$(A^T P A)x = A^T P(\ell - a_0)$$
$$(B^T P B)k = B^T P(a_0 - \ell)$$

The optima fulfill:

$$d(\tilde{\ell}, a_0)^2 + d(\tilde{\ell}, \ell)^2 = d(a_0, \ell)^2$$

Lower and upper bounds for the optimum of (I) follow from

$$d(a_0, \ell)^2 - d(a_0, \ell'')^2 \leq d(\ell, \tilde{\ell})^2 \leq d(\ell, \ell')^2$$

provided that

$$\ell' \in M_A,\ \ell'' \in M_B$$

It remains to characterize the linear manifold M_B. It consists of all observation vectors ℓ'' which result in the same adjusted values $\tilde{\ell}$ as ℓ does. It is even more instructive to characterize L_B. It consists of all observation vectors $b \in L$ whose adjustment results in the zero vector. (The vector of residuals is then the negative of the observation vector.) Hence problem (II) is formulated in words as follows:

(II): Find an observation vector of minimal norm which results in the same adjusted vector $\tilde{\ell}$ as the original vector ℓ does.

11.4 Adjustment by minimizing variances.

We take some time and space to review (in different notation) concepts introduced earlier (confer chapter B.4). The observations ℓ are now viewed as random variables. L is the space of their realizations, also called the sample space. The expectation $E(\ell) = \lambda$ is restricted to the submanifold M_A.

$$E(\ell) = \lambda \in M_A$$

The covariance matrix of the observations ℓ is

$$\Sigma(\ell) = Q\sigma^2$$

The positive definite matrix Q is known, the factor σ^2, called mean square unit weight error, is either known or unknown.

The inner product of L is represented by the weight matrix P. As pointed out in section 8.4.1, there is an isometry between L and its dual L'. The inner product in L' is represented by Q. Q is also called the reproducing kernel of L (not of L'). It holds that

$$Q = P^{-1}$$

The isometry is established by the representation of functionals as vectors. Confer sections 4.8 and 4.9! Let f be a functional $f \in L'$. If f is applied to a vector $\underline{l} \in L$ we write

$$f(\underline{l}) = f^T \underline{l} = f_1 \underline{l}_1 + \ldots + f_n \underline{l}_n$$

The representor r of f is a vector fulfilling

$$f(\underline{l}) = (r, \underline{l})$$

It follows that

$$f(\underline{l}) = f^T \underline{l} = r^T P \underline{l} = (r, \underline{l})$$

holds, if and only if

$$f = Pr$$

Equivalently

$$r = Qf$$

If f is represented by r and g by s, then

$$(f,g) = f^T Qg = (r,s) = r^T Ps$$

The isometry preserves the inner product.

Now let f be a functional out of L'. Writing

$$F(\ell) = f(\ell) + f_0 = f^T \ell + f_0 = \sum_{i=1}^{n} f_i \ell_i + f_0$$

this can also be viewed as a linear inhomogeneous function of the random variables ℓ. As such $F(\ell)$ has the variance

$$\sigma^2(F(\ell)) = \sigma^2(f^T \ell) = f^T Qf \sigma^2$$

Up to the factor σ^2 this equals $\|f\|^2$, the squared norm of f.

Another linear inhomogeneous function $\hat{F}(\ell) = \hat{f}^T \ell + \hat{f}_0$ is called an unbiased estimator of f, provided that

$$E\{\hat{F}(\ell)\} = E\{F(\ell)\}$$

Whatever the value of $E(\ell) = \lambda \in M_A$ may be. We obtain

$$E\{\hat{f}^T \ell\} + \hat{f}_0 = E\{f^T \ell\} + f_0$$

$$\hat{f}^T E\{\ell\} + \hat{f}_0 = f^T E\{\ell\} + f_0$$

$$\hat{f}^T \lambda + \hat{f}_0 = f^T \lambda + f_0 \quad \text{for any } \lambda \in M_A$$

Replacing λ by $\lambda = a_0 + a$, $a \in L_A$, it follows that

$$\hat{f}^T a = f^T a \quad \text{for any } a \in L_A, \quad \text{and} \quad \hat{f}^T a_0 + \hat{f}_0 = f^T a_0 + f_0$$

or

$$(f-\hat{f})^T a = 0 \quad \text{for any } a \in L_A, \quad \text{and} \quad \hat{f}_0 = (f-\hat{f})^T a_0 + f_0$$

Once, an unbiased functional $\hat{f}$ is established, the constant f_0 is easily determined from the last relation.

In practice $F(\ell)$ is related to a socalled "derived" quantity such as for example a distance between two remote points in a network. Replacing $F(\ell)$ by $\hat{F}(\ell)$ gives a random function $\hat{F}(\ell)$ having the same expectation as $F(\ell)$. One may exploit this fact trying to replace $F(\ell)$ by an $\tilde{F}(\ell)$ having a variance as small as possible. This optimal $\tilde{F}(\ell)$ will be called best unbiased estimator. Representing $\tilde{F}(\ell)$ as $\tilde{f}^T(\ell) + \tilde{f}_0$, the decisive problem is to find $\tilde{f}$. This problem is the following one:

Given $f \in L'$, find $\tilde{f} \in L'$ such that

$$(f-\tilde{f})a = 0 \quad \text{for any } a \in L_A$$

and

$$\|\tilde{f}\|^2 = \text{Minimum}$$

We further exploit the isometry between L and L'. The orthocomplementary subspaces L_A and L_B have their orthocomplementary counter-parts in L' as L'_A, L'_B. If L_A and L_B are spanned by the columns of A and B, then L'_A, L'_B are spanned by the columns of A' and B'. Thereby

$$A' = PA \qquad A = QA'$$

$$B' = PB \qquad B = QB'$$

Functionals g fulfilling

$$(f-g)a = 0 \quad \text{for } a \in L_A$$

are recognized to be precisely functionals representable as

$$g = f + h \qquad h \in B'$$

The proof is obtained by using representing vectors:

$$h(a) = (b,a)$$

with b representing h. The inner product (b,a) is zero if and only if $b \in L_B$, i.e. $h \in L'_B$. We call the set of all functionals g fulfilling the above relation M'_B:

$$M_B' = \{g \in L' \mid g = f + h,\ h \in L_B'\}$$

M_B' is the set of functionals $\hat{f}$ leading to unbiased estimators $\hat{F} = \hat{f}^T \ell + \hat{f}_0$ for $F(\ell)$ after choosing a suitable constant $\hat{f}_0$. Thus we arrive at the following problem, which we call

(II') Find $\tilde{f} \in M_B'$ having minimal norm.

It is seen that f plays the role of the vector b_0 in section 11.2, and that the zero vector now plays the role of the earlier vector a_0. Thus the complementary problem is immediately obtained by taking $M_A' = L_A'$.

(I') Given f find $\tilde{f} \in L_A'$ such that

$$\|f-\tilde{f}\|^2 = \text{Minimum}$$

We see that

$$\tilde{f} = \Pi_A' f$$
$$\tilde{f} = f + \Pi_B'(-f) = (I-\Pi_B')f$$

The projectors Π_A', Π_B' are represented by the matrices

$$\Pi_A' = A'(A'^T Q A)^{-1} A'^T Q$$
$$\Pi_B' = B'(B'^T Q B)^{-1} B'^T Q$$

Recalling A' = PA, B' = PB, one verifies

$$P'_A = P_A^T \ldots P_A = A(A^TPA)^{-1}A^TP$$
$$P'_B = P_B^T \ldots P_B = B(B^TPB)^{-1}B^TP$$

This brings out the fact that Π'_A is the adjoint operator of Π_A:

$$(\Pi'_A f)(\ell) = f(\Pi_A \ell)$$

Similarly Π'_B is the adjoint of P_B. This closes the gap between the two ways to perform an adjustment. An adjusted functional (a best linear unbiased estimator i.e. a BLUE) applied to a vector ℓ is the same as the original functional applied to the adjusted vector of observations.

Remark: A useful application of the complementary problems I' and II' is the a-priori specification of upper and lower bounds on the variance of the BLUE $\tilde{f}^T\ell$ of a linear function $f^T\ell$ of the observations. Any unbiased estimator $\hat{f}^T\ell$ will give an upper bound by II':

$$\sigma^2(\tilde{f}^T\ell) \leq \sigma^2(\hat{f}^T\ell)$$

On the other hand, by I', any functional $\overset{v}{f} \in L_A$, will give a lower bound

$$\sigma^2(\tilde{f}^T\ell) \geq \sigma^2(f^T\ell) - \sigma^2(\overset{v}{f}^T\ell)$$

If L_A is spanned by the columns of A, then any functional in $L_{A'}$ is a linear combination of the columns of PA.

Thus, one can frequently specify useful upper and lower bounds on $\sigma^2(\tilde{f}^T\underline{\ell})$ before an adjustment is actually carried out.

12. Generalized inverses.

12.1. Range space and null space of a linear operator.

Let V_m and V_n be vector spaces of dimensions m and n respectively. We do not impose any restrictions such as $m \leq n$ onto the dimensions. Let Λ be a linear operator from V_m into V_n. After a choice of bases in V_m and V_n, the operator is represented by an nxm matrix A.

The set of vectors $x \epsilon V_m$ which is mapped onto the zero vector forms a vector subspace $N(\Lambda)$ or briefly N of V_m. N is called null space of Λ.

$$N = N(\Lambda) = \{x \epsilon V_m \mid \Lambda(x) = 0\}$$

The set of vectors $y \epsilon V_n$, such that y is the image of some vector $x \epsilon V_m$, is called the range space $R(\Lambda)$ or briefly R. R is a vector subspace of V_n:

$$R = R(\Lambda) = \{y \epsilon V_n \mid y = \Lambda(x) \text{ for some } x \epsilon V_m\}$$

A basis of N is obtained by identifying a maximal linearly independent set of solutions x to the homogeneous system

$$Ax = 0$$

This can be accomplished by the Gauss-Jordan procedure (cf. section A.1.5).

A basis of R is given by a maximal number of linearly independent columns of A.

We know from section A.2.5 that an inverse operator Λ^{-1} exists if and only if

$$m = n, \quad N = 0, \quad R = V_n$$

In this case the matrix A is nxn, its rank is n. The matrix is regular and possesses an inverse A^{-1}, the matrix representation of Λ^{-1}.

The theory of generalized inverses attempts to extend the notion of an inverse operator and an inverse matrix to situations where Λ^{-1} and A^{-1} no longer exist. Of course, some requirements of an inverse operator (inverse matrix) have to be relaxed.

12.2. The g-inverse.

A linear operator Λ^g is called a g-inverse (generalized inverse), if it maps any $y \in R$ back onto a pre-image x of y:

$$\text{if } y \in R \text{ and } x = \Lambda^g(y) \text{ then } \Lambda(x) = y$$

This is equivalent to

$$\Lambda \circ \Lambda^g(y) = y \quad \text{for} \quad y \in R$$

Since any $y \in R$ may be represented as

$$y = \Lambda(x) \quad \text{for some } x \epsilon V_n$$

We also have

$$\Lambda \circ \Lambda^g \circ \Lambda(x) = \Lambda(x) \quad \text{for any } x \epsilon V_n$$

It follows that

$$\Lambda \circ \Lambda^g \circ \Lambda = \Lambda$$

is the necessary and sufficient requirement for Λ^g to be a generalized inverse of Λ.

Let A^g be the matrix representation of Λ^g. Note that Λ^g maps V_n into V_m. Hence A^g is an mxn matrix, while A was nxm. We call A^g a generalized inverse matrix of A. It is characterized by

$$A\, A^g\, A = A$$

The matrix A^g is also characterized by the following property: If the system

$$Ax = y$$

is consistent (i.e. if y is in the span of the columns of A), then

$$x = A^g y$$

yields a solution of this system. There may be other solutions for the same right hand side y.

The operator Λ^g and its matrix A^g are generally not unique. This is plausible because the solution to a consistent system $Ax = y$ is generally not unique. Let θ_1 be an operator from V_n into V_m whose range space $R(\theta_1)$ lies in $N = N(\Lambda)$. Let θ_2 be an operator from V_n into V_m whose null space contains $R = R(\Lambda)$. If Λ^g is any particular generalized inverse, then

$$\Lambda^g + \theta_1 + \theta_2$$

is also a generalized inverse. One readily verifies

$$\Lambda\circ(\Lambda^g + \theta_1 + \theta_2)\circ\Lambda = \Lambda$$

because $\Lambda\circ\theta_1 = 0$ and $\theta_2\circ\Lambda = 0$

<u>A general theorem on projectors.</u> Let Π be an operator from a vector space V into itself. Necessary and sufficient for Π to be a projector is the relation

$$\Pi\circ\Pi = \Pi$$

<u>Remark.</u> Let P be the matrix representation of Π. Then the above relation

presents itself as $PP = P$.

Definition. An operator (matrix) fulfilling $\Pi\circ\Pi = \Pi$ ($PP = P$) is called idempotent.

Proof of the theorem on projectors. Recall that a projector Π induces a decomposition of V into a direct sum of subspaces, the range space $R(\Pi)$ and the null space $N(\Pi)$. Vectors in the range space are reproduced: $\Pi(x) = x$ if $x = \Pi(y)$ for any y. Thus $\Pi(\Pi(y)) = \Pi(y)$ for any y. This proves $\Pi\circ\Pi = \Pi$, i.e. necessity. To prove sufficiency, assume that $\Pi\circ\Pi = \Pi$ holds. Consider $R(\Pi)$ and $N(\Pi)$. If $x\in R(\Pi)$, then $x = \Pi(y)$ for some y. From $\Pi\circ\Pi = \Pi$ we infer that $\Pi(x) = \Pi\circ\Pi(y) = \Pi(y) = x$. Thus vectors in $R(\Pi)$ are reproduced. Next we show that any vector x can be represented as $x = x_1 + x_2$ with $x_1\in R(\Pi)$, $x_2\in N(\Pi)$. Just put $x_1 = \Pi(x)$ and $x_2 = x - \Pi(x)$. Then obviously $x_1\in R(\Pi)$ and $x_2\in N(\Pi)$. It remains to prove that $R(\Pi)$ and $N(\Pi)$ have only the zero vector in common. If $x\in R(\Pi)$ then $x = \Pi(x)$. If at the same time $x\in N(\Pi)$, then $\Pi(x) = 0$, i.e. $x = 0$. This was to be shown.

Remark. We are not talking about orthogonal projectors. An inner product may not even be defined in V.

Theorem. The operators $\Lambda\circ\Lambda^g$, $\Lambda^g\circ\Lambda$, represented by AA^g, A^gA, are projectors in V_m, V_n respectively. It holds that

$$R(\Lambda\circ\Lambda^g) = R(\Lambda), \quad N(\Lambda^g\circ\Lambda) = N(\Lambda)$$

Proof. Both operators are verified to be idempotent. Hence they are projectors. Any vector y in $R(\Lambda)$ is represented as $\Lambda(x)$ for some x. From $\Lambda\circ\Lambda^g\circ\Lambda = \Lambda$, i.e. $\Lambda\circ\Lambda^g\circ\Lambda(x) = \Lambda(x)$ for any x, it follows that $\Lambda\circ\Lambda^g(y) = y$. Hence $R(\Lambda)$ includes $R(\Lambda\circ\Lambda^g)$. The reverse inclusion is trivial. Thus $R(\Lambda\circ\Lambda^g) = R(\Lambda)$. We turn to $\Lambda^g\circ\Lambda$. Obviously $N(\Lambda^g\circ\Lambda)$ includes $N(\Lambda)$. Suppose that there is a vector z in $N(\Lambda^g\circ\Lambda)$ which is not in $N(\Lambda)$. Then $(\Lambda^g\circ\Lambda)(z) = 0$, but $\Lambda(z) \neq 0$. Put these equations into matrix form

$$A^g(Az) = 0, \quad Az \neq 0$$

This tells us that x=0 is a solution to the consistent system

$$Ax = Az, \quad Az \in R(\Lambda), \quad Az \neq 0$$

This is impossible. Hence $N(\Lambda^g\circ\Lambda) = N(\Lambda)$, as was to be shown.

Remark. If Λ^{-1} is the ordinary inverse of Λ, then Λ is the ordinary inverse of Λ^{-1}. This is generally not true in the case of Λ and Λ^g: Λ may not be a $(\Lambda^g)^g$. This failure will be repaired in the next subsection by imposing further restrictions on Λ^g.

12.3. Reflexive generalized inverse. In addition to

$$\Lambda\circ\Lambda^g\circ\Lambda = \Lambda, \quad \text{i.e.} \quad AA^gA = A$$

we also require

$$\Lambda^g \circ \Lambda \circ \Lambda^g = \Lambda^g, \text{ i.e. } A^g A A^g = A^g$$

Then Λ is also a generalized inverse of Λ^g. The roles of Λ and Λ^g can be interchanged. We call such a generalized inverse "reflexive". It will be denoted by Λ^r. Its matrix representation is A^r. We repeat the above equations in the new notation:

$$\Lambda \circ \Lambda^r \circ \Lambda = \Lambda, \text{ i.e. } AA^rA = A$$
$$\Lambda^r \circ \Lambda \circ \Lambda^r = \Lambda^r, \text{ i.e. } A^rAA^r = A$$

It follows that the projectors $\Lambda \circ \Lambda^r$ and $\Lambda^r \circ \Lambda$ fulfill:

$$R(\Lambda \circ \Lambda^r) = R(\Lambda), \quad N(\Lambda \circ \Lambda^r) = N(\Lambda^r)$$
$$R(\Lambda^r \circ \Lambda) = R(\Lambda^r), \quad N(\Lambda^r \circ \Lambda) = N(\Lambda)$$

From these equations and a dimension argument one can infer that A and A^g have equal rank. One can also show that this requirement is sufficient for A^r to be a reflexive inverse.

12.4. Generalized inverse with least squares property.

Assume an inner product in V_n. Let it be represented by the matrix G_n. We may form the orthocomplement $R^\perp$ of R. We consider a generalized inverse Λ^g, i.e. we require

$$\Lambda \circ \Lambda^g \circ \Lambda = \Lambda, \quad \text{i.e.} \quad AA^gA = A$$

In addition we postulate

$$N(\Lambda \circ \Lambda^g) = R^\perp, \quad \text{i.e.} \quad AA^gy = 0 \text{ is equivalent to } y \in R^\perp.$$

The projector $\Lambda \circ \Lambda^g$ is thus required to be an orthogonal projector. We denote such a generalized inverse by Λ^l, and its matrix by A^l. The importance of A^l is stressed in the following

Theorem. Consider the (generally) inconsistent system

$$Ax = y$$

Let y be arbitrary. Then

$$x = A^gy$$

fulfills the least squares reqirement

$$\underset{z \in V_m}{\text{Min}} \|y - Az\| = \|y - Ax\|$$

if and only if $A^g = A^l$.

Proof. Given $y \epsilon V_n$, decompose it as

$$y = y_1 + y_2, \quad y_1 \epsilon R, \quad y_2 \epsilon R^{\perp}$$

Suppose that $N(\Lambda \circ \Lambda^g) = R^{\perp}$, then the projector $\Pi = \Lambda \circ \Lambda^g$ onto R is an orthogonal projector. Calculate

$$x = \Lambda^g(y)$$

then

$$z = \Lambda(x) = \Lambda \circ \Lambda^g(y) = \Pi(y)$$

is the orthogonal projection of y onto R. Thus z is the solution of the stated extremum problem.

Suppose now that

$$x = \Lambda^g(y)$$

gives the least squares solution for any y. Then

$$\Lambda(x) = \Lambda \circ \Lambda^g(y)$$

must be the orthogonal projection of y onto R. Thus $\Lambda \circ \Lambda^g$ is the orthogonal

projector onto R.

A general theorem on projectors.

Let V be an inner product space. Necessary and sufficient for an operator Π to be an orthogonal projector is

$$\Pi \circ \Pi = \Pi \quad \text{and} \quad \Pi^* = \Pi$$

Remark. Π^* is the adjoint operator of Π in the sense of section A.4.10, i.e. it holds that $(\Pi(x),y) = (x,\Pi^*(y))$. If the inner product in V is represented by the matrix G, and if Π is represented by P, then the above conditions are restated as

$$PP = P$$
$$P^* = P, \quad P^* = G^{-1}P^TG$$

Proof. It was shown above that $\Pi \circ \Pi = \Pi$ is necessary and sufficient for Π to be a projector onto $R(\Pi)$, and that V is spanned by $R(\Pi)$ and $N(\Pi)$. It suffices to show that $\Pi = \Pi^*$ is equivalent to $R(\Pi)^\perp = N(\Pi)$. Assume that $\Pi = \Pi^*$. From the defining equation for the adjoint operator [which is $(\Pi(x),y) = (x,\Pi^*(y))$], we get

$$(\Pi(x),y) = (x,\Pi(y))$$

This shows: If $y \in N(\Pi)$ and $x \in R(\Pi)$ then $0 = (x,0) = (x,\Pi(y)) = (\Pi(x),y) = (x,y) = 0$. Thus $N(\Pi)$ is included in $R(\Pi)^\perp$.

If x is arbitrary and $y \in R(\Pi)$, then $(x,\Pi(y)) = (\Pi(x),y) = 0$. This shows that $\Pi(y) = 0$. Hence $R(\Pi)$ is included in $N(\Pi)$. Thus $\Pi = \Pi^*$ indeed implies $R(\Pi) = N(\Pi)$.

Now assume $R(\Pi) = N(\Pi)$, i.e. assume that Π is an orthogonal projector. It was shown in section A.5.6 that $\Pi = \Pi^*$. We give another proof as follows. From

$$(\Pi^*(x),y) = (x,\Pi(y))$$

we deduce for <u>any</u> x:

(1) if $y \in R(\Pi)$: $(\Pi^*(x),y) = (x,\Pi(y)) = (x,y) = (\Pi(x) + (I-\Pi)(x),y) = (\Pi(x),y)$

(2) if $y \in R(\Pi)^\perp$: $(\Pi^*(x),y) = (x,\Pi(y)) = (x,0) = 0 = (\Pi(x),y)$

Thus $(\Pi^*(x),y) = (\Pi(x),y)$ holds for all x and all y. Hence $\Pi^* = \Pi$.

Using the theorem we arrive at the following characterization of a least squares inverse $\Lambda^!$ represented by $A^!$:

$$\Lambda \circ \Lambda^! \circ \Lambda = \Lambda \qquad \text{or } AA^!A = A$$
$$(\Lambda \circ \Lambda^!)^* = \Lambda \circ \Lambda^! \quad \text{or } G_n^{-1}(AA^!)^T G_n = AA^!$$

<u>12.5. Generalized inverse with minimum norm property.</u>

Suppose that V_m is equipped with an inner product. V_n does not necessarily have an inner product. In addition to

$$\Lambda \circ \Lambda^g \circ \Lambda = \Lambda$$

we require

$$R(\Lambda^g \circ \Lambda) = N$$

Because $N(\Lambda^g \circ \Lambda) = N$, we thus require that $\Lambda^g \circ \Lambda$ is an orthogonal projector. We call a g-inverse fulfilling these conditions a minimum norm inverse Λ^m. Its matrix is denoted by A^m.

Theorem. Suppose that the system

$$Ax = y$$

is consistent. Otherwise let y be arbitrary. Then

$$x = A^g y$$

fulfills the minimum norm requirement

$$\min_{\substack{z \in V_m \\ Az = y}} \|z\| = \|x\|$$

if and only if $A^g = A^m$.

Proof. Consider the consistent system

$$Az = y$$

Represent z as

$$z = z_1 + z_2, \quad z_1 \epsilon N^{\perp}, \; z_2 \epsilon N$$

All solutions to the system Az = y are obtained by keeping z_1 fixed (as a particular solution to the homogeneous system) and by letting z_2 vary over N (as the general solution to the homogeneous system Az = 0). Because $\|z\|^2 = \|z_1\|^2 + \|z_2\|^2$, the minimal solution is obviously $z = z_1$. It is obtained as $z_1 = A^g y$ if and only if the image of $A^g y$ is in $N^{\perp}$ for any $y \epsilon R$. Such y are represented as y = Au. Thus $A^g A$ u must be in $N^{\perp}$ for any u. Because $A^g A$ is a projector whose null space is N, $A^g A$ must be the projector onto $N^{\perp}$, i.e. it must be an orthogonal projector.

The characterization of a minimum norm inverse is

$$\Lambda \circ \Lambda^m \circ \Lambda = \Lambda \qquad \text{or } AA^mA = A$$
$$(\Lambda^m \circ \Lambda)^* = \Lambda^m \circ \Lambda \text{ or } (A^mA)^* = A^mA \text{ with } (A^mA)^* = G_m^{-1}(A^mA)^T G_m$$

12.6. The minimum norm least squares inverse.

Assume that an inner product is specified in V_m as well as in V_n. Given an

operator Λ from V_m into V_n, we are searching for a Λ^g giving a least squares solution of minimal norm. We shall call such an inverse Λ^{lm}. Translated into the language of matrices, we start from a (generally) inconsistent system

$$Ax = y$$

and we search for x fulfilling

$$\|x\| = \min_{z \in Z} \|z\|$$

where Z is the set of least squares solutions defined by

$$Z = \{z \in V_m \mid \|y - Az\| \leq \|y - Au\| \text{ for any } u \in V_m\}$$

We shall show that the solution to this problem is unique. Thus Λ^{lm} and A^{lm} will be unique. We decompose V_m and V_n into direct sums:

$$V_m = N^{\perp} + N$$
$$V_n = R + R^{\perp}$$

Decomposing y as

$$y = y_1 + y_2, \quad y_1 \in R, \quad y_2 \in R^{\perp}$$

the set Z of least squares solutions z is given by

$$Az = y_1$$

[If $Az = y_1 + \eta_1$ with $\eta_1 \epsilon R$, then we would have $\|y - Az\|^2 = \|\eta_1\|^2 + \|y_2\|^2$. Obviously this is minimal for $\eta_1 = 0$.] We decompose

$$z = z_1 + z_2, \quad z_1 \epsilon N^{\perp}, \quad z_2 \epsilon N$$

Because

$$\|z\|^2 = \|z_1\|^2 + \|z_2\|^2$$

the minimal z is obviously $z = z_1$. This concludes the proof. From the proof it is clear that

(1) The operator Λ maps $N^{\perp}$ onto R and N onto 0

(2) The operator Λ^{lm} maps R back onto $N^{\perp}$ and $R^{\perp}$ onto 0.

If the operators Λ, Λ^{lm} are restricted to the subspaces $N^{\perp}$ and R, then Λ^{lm} is the conventional inverse of Λ.

The operator Λ^{ln} and its matrix A^{lm} may be uniquely characterized by the following equations:

(1) $A \circ A^{lm} \circ A = A$ or $AA^{lm}A = A$

(2) $A^{lm} \circ A \circ A^{lm} = A^{lm}$ or $A^{lm}AA^{lm} = A^{lm}$

(3) $(A \circ A^{lm})^* = A \circ A^{lm}$ or $(AA^{lm})^* = AA^{lm}$

(4) $(A^{lm} \circ A)^* = A^{lm} \circ A$ or $(A^{lm}A)^* = A^{lm}A$

The necessity of these relations is easily proved: (1) holds for any A^g, (3) holds for a least squares inverse, (4) holds for a minimum norm inverse. (2) is proved directly from the above geometric characterization of A and A^{lm}. (Just verify what A^{lm} does to vectors in R and $R^\perp$).

Sufficiency of (1), (3), (4) is also clear from earlier sections. The sufficiency of (2) is an interesting question. Note that a minimum norm inverse, as considered in section 12.4, guarantees a minimum norm solution only for a consistent system. However, in this section we want minimum norm solutions also for inconsistent systems. Here condition (2) steps in, excluding inverses which would not give minimum solutions to inconsistent systems. We do not further elaborate but leave it with the hint that one must be concerned with the images of vectors in $R^\perp$ under A^g.

12.7. The pseudo inverse.

Assume that the inner products in V_m, V_n are represented by the identity matrix (of appropriate dimensions m and n). The matrix representing the minimum norm least squares inverse is then denoted by A^+. It is called the pseudo inverse of A. Sometimes also the name "Moore-Penrose inverse" is used. A^+ is unique and uniquely characterized by the following relations

(1) $AA^+A = A$

(2) $A^+AA^+ = A^+$

(3) $(AA^+)^T = AA^+$

(4) $(A^+A)^T = A^+A$

Thus the pseudo inverse represents the minimum norm least squares inverse in case of orthonormal bases in V_m and V_n.

References.

RAO, C.R. and S.K. Mitra (1971): Generalized inverse of matrices and its applications. Wiley, IX + 240 pages.

BJERHAMMAR, A. (1973): Theory of errors and generalized matrix inverses. Elsevier. XII + 420 pages.

13. Adjustment of rank deficient systems.

13.1. Formulation of the problem.

We introduce the following spaces

L ... n-dimensional space of observations (sample space)

L_A ... r-dimensional subspace of adjusted observations

X ... m-dimensional parameter space, $m \geq r$

There is a mapping Λ from X onto L_A. It is represented by the nxm matrix A. The mapping is not unique if $m > r$. The rank of A is r.

The inner product in L is represented by the weight matrix P. The inner product in X shall be represented by G.

The adjustment problem is formulated as follows. Given a vector $\ell \in L$ of observations, find corrections $v \in L$ and parameters $x \in X$ such that

$$\ell + v = A\,x$$

$$v^T P v = \text{minimum}$$

The solution $\tilde{x}$ for x is generally not unique. However, the corrections are unique.

13.2. Solution via generalized inverses of A.

Let $A^!$ be a least squares inverse of A. Then

$$\tilde{x} = A^{\mathbf{l}}\mathbf{l}$$

is a solution of the adjustment problem. This is the statement of the theorem in section 12.4.

Let $A^{\mathbf{lm}}$ be the (unique) minimum norm least squares inverse of A, then

$$\tilde{x} = A^{\mathbf{lm}}\mathbf{l}$$

is the solution of the least squares problem having minimal norm. This is the result of section 12.6.

13.3. A minimum property of the covariance of the adjusted parameters $\tilde{x} = A^{\mathbf{ln}}\mathbf{l}$.

Using any least squares inverse $A^{\mathbf{l}}$, the covariance of the adjusted parameters $\tilde{x}$ is

$$\Sigma(\tilde{x}) = A^{\mathbf{l}}P^{-1}(A^{\mathbf{l}})^{T}\sigma^2$$

Let $M = (m_{ij})$ be an nxn matrix. The trace of M is defined as

$$tr(M) = \sum_{i=1}^{n} m_{ii}$$

The following facts about the trace of a matrix are needed in the subsequent theorem:

(*) if M is positive semidefinite, then $\mathrm{tr}(M) \geq 0$. (This is clear, because a positive semidefinite matrix has nonnegative diagonal elements. See section A.4.5 for the definition of positive definite matrices.)

(*) $\mathrm{tr}(MN) = \mathrm{tr}(NM)$. The proof is an easy exercise. Note that M and N need not be square matrices, only MN must be square.

Theorem. It holds that

$$\mathrm{tr}\{GA^{lm} P^{-1} (A^{lm})^T\} \leq \{GA^{l} P^{-1} (A^{l})^T\}$$

Thus the trace of $G\,\Sigma(\tilde{x})$ is minimal if A^l is chosen as A^{lm}.

Proof. Let $x = A^l y$. Decompose $x = x_1 + x_2$ with $x_1 \in N^\perp$ and $x_2 \in N$ ($N = N(A)$ being the null space of A). Then $x_1 = A^{lm} y$. We call $B = A^l - A^{lm}$. The range $R(B)$ is in N. Thus $B^T G A^{lm} = 0$ holds (recall that G represents the inner product in X). Next observe that

$$\begin{aligned}
&\mathrm{tr}\{G A^l P^{-1} (A^l)^T\} = \\
&= \mathrm{tr}\{G [A^{lm} + B] P^{-1} [A^{lm} + B]^T\} = \\
&= \mathrm{tr}\{G A^{lm} P^{-1} (A^{lm})^T\} + \mathrm{tr}\{G A^{lm} P^{-1} B^T\} + \\
&\quad \mathrm{tr}\{G B P^{-1} (A^{lm})^T\} + \mathrm{tr}\{G B P^{-1} B^T\} = \\
&= \mathrm{tr}\{G A^{lm} P^{-1} (A^{lm})^T\} + \mathrm{tr}\{G B P^{-1} B^T\}
\end{aligned}$$

because $tr\{G\, A^{lm}\, P^{-1}\, B^T\} = tr\{B^T\, G\, A^{lm}\, P^{-1}\} = 0$, since $B^T G\, A^{lm} = 0$. Similarly $tr\{G\, B\, P^{-1}\, (A^{lm})^T\} = 0$.

Now we focus attention on the trace

$$tr\{G\, B\, P^{-1}\, B^T\}$$

Decomposing $G = RR^T$ into Cholesky factors, we have

$$tr\{G\, B\, P^{-1}\, B^T\} = tr\{R\, B\, P^{-1}\, B^T R^T\}$$

Now $R\, B\, P^{-1}\, B^T R^T = (RB)\, P^{-1}\, (RB)^T$ is positive semidefinite. (If M is positive semidefinite, so is WMW^T for any W. The proof of this just uses the definition of positive semidefiniteness: For arbitrary x we have $x^T\, (WMW^T)\, x = (W^T x)^T\, M \cdot (W^T x) = y^T My$, with $y = W^T x$. However, $y^T My > 0$, because M is positive semidefinite.) Thus $tr\{R\, B\, P^{-1}\, B^T R^T\} = tr\{G\, B\, P^{-1}\, B^T\} > 0$ and the theorem is proved.

13.4. Solution via singular normal equations.

We form the normal equations

$$(A^T P\, A)\, \tilde{x} = A^T P l$$

They are singular if m>r. Any solution to these normals is a least squares solution. This follows from the fact that the normals require nothing else but

the orthogonality of $v = A\tilde{x} - \ell$ and the columns of A. The normal equations are always consistent, because the projection of ℓ onto the space spanned by the columns of A must exist. (An explicit consistency proof is given in section 13.6.) Thus, if $(A^TPA)^g$ is any generalized inverse of A^TPA, then

$$\tilde{x} = (A^TPA)^g A^TP\ell = A^{l}\ell$$

It follows that

$$A^{l} = (A^TPA)^g A^TP$$

This equation means that if $(A^TPA)^g$ runs through all generalized inverses of A^TPA then, in any case, a least squares inverse A^{l} is obtained. If $(A^TPA)^m$ is any minimum norm inverse, then

$$\tilde{x} = (A^TPA)^m A^TP\ell$$

must be the least squares solution having minimal norm. Thus

$$A^{lm} = (A^TPA)^m A^TP$$

We find for $\tilde{x} = A^{lm}\ell$:

$$\Sigma(\tilde{x}) = (A^TPA)^m A^TPP^{-1}PA (A^TPA)^m \sigma^2 =$$
$$= (A^TPA)^m (A^TPA) (A^TPA)^m \sigma^2$$

If $(A^TPA)^m$ is chosen as $(A^TPA)^{\ell m}$, then, due to the reflexivity of $(A^TPA)^{\ell m}$, we have

$$\Sigma(\tilde{x}) = (A^TPA)^{\ell m}\ \sigma^2$$

This, by the way, shows

$$(A^TPA)^{\ell m} = (A^TPA)^m\ (A^TPA)\ (A^TPA)^m$$

More important is the minimum trace property derived in section 13.3:

$\Sigma(\tilde{x})$ is minimal if $(A^TPA)^m$ is chosen as $(A^TPA)^{\ell m}$

13.5. Calculation of the ℓm-inverse $A^{\ell m}$.

There are many ways to calculate the unique inverse $A^{\ell m}$. We consider two procedures. The first one is recommended if rank(A) is small compared to the size of the nxm matrix A, more precisely if

$$\text{rank}(A) \ll \min(n,m)$$

The second procedure works well for matrices which are nearly square and whose rank is nearly equal to the size, more precisely

$$\text{rank}(A) \doteq m \doteq n$$

Method (1). Let r be the rank of A. Consider a rank factorization of A, i.e. a decomposition

$$A = BC^T$$

where B ... nxr and C^T...rxm are both of rank m. Then

$$A^{\dagger m} = GC\,(C^TGC)^{-1}\,(B^TPB)^{-1}\,B^TP$$

The proof follows by verifying the four conditions for $A^{\dagger m}$ given in section 12.6. (Recall that G represents the inner product in the domain space of A.)

Remark. A rank factorization for A may be deduced from the last stage of the Gauss-Jordan procedure (exercise).

Method (2). Consider matrices S, N of size n x (n-r) and m x (m-r), such that S spans R(A) and N spans N(A):

$$A^TPS = 0$$
$$AN = 0$$

Both S and N have the maximal number of linearly independent columns fulfilling the two homogeneous systems above.

Consider the square matrix of size (n+m-r) x (n+m-r)

$$H = \begin{bmatrix} A & S \\ N^T G & 0 \end{bmatrix}$$

We show that this matrix is regular by showing that Hz = 0 implies z = 0.

Decompose z as

$$z = \begin{bmatrix} x \\ y \end{bmatrix}$$

Then Hz = 0 means

$$Ax + Sy = 0$$
$$N^T Gx = 0$$

The spaces spanned by the columns of A and S are orthocomplementary. Hence Ax = 0 and Sy = 0 must hold. Since S has linearly independent columns, we infer y = 0. We are left with

$$Ax = 0$$
$$N^T Gx = 0$$

Such a vector x must be in N(A) and, at the same time, it must be orthogonal to N(A) which is spanned by the columns of N. Thus x = 0 and H is indeed regular.

We form the inverse of H, and we denote it as

$$H^{-1} = \begin{bmatrix} Q & K \\ L^T P & M \end{bmatrix}$$

Our aim is to show that $Q = A^{\#m}$. As a first step we show that $M = 0$. From $HH^{-1} = I$ we deduce

$$AK + SM = 0$$

In the regularity proof for H we have seen that these equations imply $M = 0$. Thus

$$H^{-1} = \begin{bmatrix} Q & K \\ L^T P & 0 \end{bmatrix}$$

We now write out the equations for $HH^{-1} = I$ in full:

$$AQ + SL^T P = I \quad (a)$$
$$AK = 0 \quad (b)$$
$$N^T GQ = 0 \quad (c)$$
$$N^T GK = I \quad (d)$$
$$QA + KN^T G = I \quad (a')$$
$$QS = 0 \quad (b')$$
$$L^T PA = 0 \quad (c')$$
$$L^T PS = I \quad (d')$$

Post-multiplying (a) by A and minding (c') we find the first condition of

section 12.6 for a minimum norm least square inverse Q

$$AQA = A \quad (1)$$

Post-multiplying (a') by Q and minding (c) we get

$$QAQ = Q \quad (2)$$

We see that Q is a reflexive inverse of A. Thus

$$\text{rank}(Q) = \text{rank}(A) = r$$

We are done if we can show that the projectors AQ and QA are orthogonal projectors. This is equivalent to conditions (3) and (4) for A^{lm} as given in section 12.6:

$$(AQ)^* = AQ \quad (3)$$
$$(QA)^* = QA \quad (4)$$

What we need to show is that the null space of AQ is S and that the range space of QA is $N^\perp$. The first assertion may be deduced from (b'), the second one from (c). In both cases it is necessary to observe that the rank of AQ and QA is r. This implies that the null space of AQ cannot be larger than the space spanned by S, and that the range space of QA cannot be larger than the orthocomplement of N. Thus we have proved that the submatrix Q of H^{-1} is indeed identical to A^{lm}.

Remark. From A being the Q^{lm}-inverse of Q, and from reasons of symmetry, we may deduce that the submatrices K and L of H^{-1} span $R(Q)^{\perp}$ and N(Q). This also follows algebraically from $R(Q)^{\perp} = N(A)$ together with (b), and from $N(Q) = R(A)^{\perp}$ together with (c'). Thus $R(A)^{\perp} = N(Q)$ is spanned by S as well as by L, and $N(A) = R(Q)^{\perp}$ is spanned by N as well as by K. Confer the geometric characterization of A^{lm} given in section 12.6.

13.6. Application to free network adjustment.

We illustrate the principle by assuming, as a special example, a network in the plane involving distance measurements. There may also be a number of measured angles or unoriented directions. However, no azimuth measurements shall be available. Also measurements of absolute positions shall be absent. The coordinates of the network points are denoted by x_i, i=1,...,n. As usual they are represented as

$$x_i = x_i^{(0)} + \Delta x_i, \quad y_i = y_i^{(0)} + \Delta y_i, \quad i=1,\ldots,n$$

whereby $x_i^{(0)}$, $y_i^{(0)}$ denote known approximate values and Δx_i, Δy_i denote small unknown increments. We introduce the vector Δz by $\Delta z^T = (\Delta x_1, \Delta y_1, \Delta x_2, \Delta y_2, \ldots .. \Delta x_n, \Delta y_n)^T$. We do not assign fixed coordinates to any of the network points, nor do we fix any of the azimuths within the network. As a result the observation equations

$$\Delta\ell + v = A\,\Delta z$$

are singular, i.e. the linear system

$$A\,\Delta z = 0$$

has nonzero solutions. If we form the normal equations

$$(A^T P A)\,\Delta z = A^T P\,\Delta\ell$$

then also these equations are singular. The system

$$(A^T P A)\,\Delta z = 0$$

has nonzero solutions. It is not too difficult to show that the solutions for $A\,\Delta z = 0$ and $(A^T P A)\,\Delta z = 0$ coincide, i.e. that the null space of A equals the null space of $A^T P A$. (If $A\,\Delta z = 0$ then trivially $A^T P A\,\Delta z = 0$. If $A\,\Delta z \neq 0$ then $(A\,\Delta z)^T P (A\,\Delta z) > 0$, since P is positive definite. It follows that $\Delta z^T (A^T P A)\,\Delta z > 0$, implying $(A^T P A)\,\Delta z \neq 0$. Q.e.d.)

On the other hand, the normal equations are consistent, i.e. a solution Δz fulfilling $(A^T P A)\,\Delta z = A^T P \Delta\ell$ exists for any choice of $\Delta\ell$. For a proof start from $A\,\Delta z = \Delta\ell + v$, decompose $\Delta\ell$ into $\Delta\ell_1 + \Delta\ell_2$ with $\Delta\ell_1 \in R(A)$ and $\Delta\ell_2 \in R(A)^\perp$. Choose $v = -\Delta\ell_2$. Then $A\,\Delta z = \Delta\ell_1$ is consistent; so is $(A^T P A)\,\Delta z = (A^T P)\,\Delta\ell_1 = A^T P\,\Delta\ell$.

The proof just given demonstrates that any solution to the normal equations gives a least squares solution to the generally inconsistent observation equations

$$A\,\Delta z = \Delta \ell$$

However, the solution to the normal equations is not unique. All kinds of solutions are obtained if we write

$$\Delta z = (A^T P A)^g\, A^T P\, \Delta \ell$$

where $(A^T P A)^g$ is any generalized inverse of the normal equation matrix $A^T P A$.

The various solutions Δz obtained in this way differ by solutions Δz_n to the homogeneous system

$$(A^T P A)\, \Delta z_n = 0$$

or also

$$A\, \Delta z_n = 0$$

These solutions are easy to specify from geometrical considerations. Since

$$\Delta \ell = A \, \Delta z$$

are the changes of the observables ℓ if the coordinates are changed by Δz, we must look for such coordinate changes which leave the observables unchanged. Such coordinate changes are implied by a translation and a rotation of the whole set of points.

If the whole set of points is translated by Δc_x, Δc_y and rotated by a small angle $\Delta\varphi$, then the coordinate changes are given by

$$\Delta z = \begin{bmatrix} 1 & 0 & -y_1 \\ 0 & 1 & x_1 \\ 1 & 0 & -y_2 \\ 0 & 1 & x_2 \\ . & . & . \\ . & . & . \\ . & . & . \\ 1 & 0 & -y_n \\ 0 & 1 & x_n \end{bmatrix} \begin{bmatrix} \Delta c_x \\ \Delta c_y \\ \Delta\varphi \end{bmatrix}$$

$$= N \, \Delta t, \text{ say}$$

It follows that

$$\Delta \ell = A \, N \, \Delta t = 0, \quad \text{i.e.} \quad AN = 0$$

The three columns of N span the null space of the matrix A. As we know the null space of A coincides with that one of A^TPA.

We assume an inner product in the parameter space implied by the unit matrix: G = I. Then the ℓm-inverse of A^TPA reduces to the pseudo inverse $(A^TPA)^+$. Applying the theorem of section 13.5, we get this pseudo inverse by taking the appropriate submatrix of

$$\begin{bmatrix} A^TPA & N \\ N^T & 0 \end{bmatrix}^{-1} = \begin{bmatrix} (A^TPA)^+ & K \\ K^T & 0 \end{bmatrix}$$

In this case

$$\Sigma(\tilde{x}) = (A^TPA)^+ \sigma^2$$

has minimal trace among all covariance matrices $\Sigma(\tilde{x})$ resulting from all choices of least squares solutions to the rank deficient network adjustment problem.

<u>References</u> (see also the references at the end of section 12!)

KOCH, K.R. (1979): Parameterschaetzung und Hypothesentests in linearen Modellen. Duemmler Verlag, Bonn.

KRARUP, T. (1979): S-transformation or how to live without the generalized inverse - almost. Geodetic Institute, Charlottenlund, Denmark.

MEISSL, P. (1962): Die innere Genauigkeit eines Punkthaufens. OeZfV, Jg.50, Nr.5,6.

MEISSL, P. (1969): Zusammenfassung und Ausbau der inneren Fehlertheorie eines Punkthaufens. In Rinner/Killian/Meissl, Beitraege zur Theorie der geodaetischen Netze im Raum, DGK, Reihe A, Heft 61, pp.8-21.

MOLENAAR, M. (1981): A further inquiry into the theory of S-transformations and criterion matrices. International Institute for Survey and Earth Sciences, Enschede, Netherlands.

POPE, A.J. (1971): Transformation of covariance matrices due to changes in minimal control. National Ocean Survey, Geodetic Research and Development Laboratory, Rockville, MD.

SCHMID, H. (1980): Vom freien zum gelagerten Netz. Mitteilungen des Institutes fuer Geodaesie und Photogrammetrie an der ETH Zuerich, Nr.29.

WOLF, H. (1972): Helmerts Loesung zum Problem der freien Netze mit singulaerer Normalgleichungsmatrix. ZfV, Jg.97, Nr.5.

WIESER, M. (1981): Wesen und Nutzen der inneren Fehlertheorie. Diplomarbeit an der TU Graz.

B. THE STOCHASTIC APPROACH TOWARD LEAST SQUARES ADJUSTMENT

1. Probabilities.

1.1 Relative frequencies.

Imagine a box filled with N identically shaped cards. Each of the individual cards carries one out of the letters A,B,C. Suppose that the corresponding absolute frequencies are N_A, N_B, N_C. Of course

$$N_A + N_B + N_C = N$$

We also introduce the relative frequencies

$$f_A = \frac{N_A}{N}, \quad f_B = \frac{N_B}{N}, \quad f_C = \frac{N_C}{N}$$

It follows that

$$f_A + f_B + f_C = 1$$

Suppose that the contents of the box are thoroughly mixed and that one card is drawn at random. We call this experiment an elementary event.

An event is defined as a set of elementary events. All events related to the present experiment are quickly listed as follows:

Φ	the empty set ... the impossible event
A	the writing on the drawn card shows the letter A
B	similar
C	similar
$A \cup B$	the letter A or the letter B is written on the card
$A \cup C$	similar
$B \cup C$	similar
$\Omega = A \cup B \cup C$	any letter is written on the card ... the certain event

We do not hesitate to assign probabilities to these events as follows

$$p(\Phi) = 0$$

$$p(A) = f_A, \quad p(B) = f_B, \quad p(C) = f_C$$

$$p(A \cup B) = f_A + f_B, \quad p(A \cup C) = f_A + f_C, \quad p(B \cup C) = f_B + f_C,$$

$$p(\Omega) = 1$$

There are alternative ways to identify an event. The symbol $\bar{A}$ denotes the event "Not A", i.e. the letter A does not appear on the drawn card. Obviously $\bar{A} = B \cup C$. Consequently $p(\bar{A}) = p(B \cup C) = 1 - f_A = f_B + f_C$.

1.2. Probability space.

Let Ω be a set of elements. The elements are called elementary events. Let Σ be a collection of subsets of Ω. Σ is a set of sets. The subsets of Ω which belong to Σ are called events. Not all possible subsets of Ω may be in Σ. We require the following properties of Σ.

(1) $\Omega \in \Sigma$

(2) Σ is closed with respect to complementation (if $A \in \Sigma$ then $\bar{A} \in \Sigma$), and with respect to the formation of countable unions (if $A_1, A_2, \ldots \in \Sigma$, then $A_1 \cup A_2 \cup \ldots \in \Sigma$).

It follows that $\Phi \in \Sigma$. (Ω is in Σ, hence $\Phi = \bar{\Omega}$ must be in Σ). Furthermore, if $A_1, A_2, \ldots$ is a countable sequence of events, $\bigcap_{i=1}^{\infty} A_i = A_1 \cap A_2 \cap \ldots$ must be in Σ. This holds because

$$\overline{\bigcap_{i=1}^{\infty} A_i} = \bigcup_{i=1}^{\infty} \bar{A}_i$$

A collection of subsets having the indicated properties is called a sigma-field of subsets (events). Frequently Σ is also called a Borel field of sets.

Let p be a functional defined on the subsets in Σ. Thus p assigns a number to any $A \in \Sigma$. This number is denoted p(A). It is called the probability of the event A. (There is no point asking whether the functional p is linear. The domain Σ is not necessarily a vector space!).

We require the following properties of the functional p.

(1) $0 \leq p(A) \leq 1$

(2) $p(\Omega) = 1$

(3) Let A_1, A_2... be a sequence of mutually nonintersecting events, i.e.

$A_i \cap A_j = 0$, if $i \neq j$.

Then

$$p(A_1 \cup A_2 \cup \ldots) = p(A_1) + p(A_2) + \ldots$$

i.e.

$$p\left\{\bigcup_{n=1}^{\infty} A_n\right\} = \sum_{n=1}^{\infty} p(A_n), \qquad \text{if} \quad A_i \cap A_j = 0 \text{ for } i \neq j$$

1.3. Examples.

1.3.1. Drawing cards out of a box.

Events and probabilities are defined in section 1.1.

1.3.2. Shooting against a target butt fixed to a wall.

Events are Borel sets of points in the plane. Borel sets in the plane are defined as follows. Suppose that a Cartesian coordinate system is chosen in the plane. The Borel field of sets in the plane is then defined as the smallest Σ field containing all rectangles of the form

$$a \leq x \leq b$$

$$c \leq y \leq d$$

It can be shown that most two dimensional sets which "can be imagined" are in Σ. It is an amusing and nontrivial task to show that for example the set of all points x, y fulfilling $ax + by < c$ is in Σ. The interior of a region bounded by a nonintersecting smooth curve is also a Borel set. Consider now a function $f(x,y)$, called a probability density function, having the following properties

$$f(x,y) \geq 0$$

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x,y)\, dxdy = 1$$

The density assigns a probability to any rectangle $a \leq x \leq b$, $c \leq y \leq d$ by means of

$$p\{a \leq x \leq b,\ c \leq y \leq d\} = \int_a^b \int_c^d f(x,y)\, dxdy$$

The properties of Borel sets and probabilities propagate the functional $p(A)$ from rectangles to any Borel set in the plane.

1.4. Calculus of probabilities.

We restrict ourselves to the following simple rules. Some of them have been anticipated in section 1.1.

$$p(\bar{A}) = 1 - p(A), \quad \bar{A} = \Omega - A$$

$$p(A \cup B) = p(A) + p(B) - p(A \cap B)$$

$$A \subset B \text{ implies } p(A) \leq p(B)$$

The calculus of probabilities is modelled after the calculus of relative frequencies.

2. Random variables.

2.1. One dimensional random variables.

Let Ω, Σ, p be the 3 ingredients of a probability space. Let X be a function mapping Ω into the real line R. X is called measureable, if for any real x the set

$$\{\omega \in \Omega \mid X(\omega) \leq x\}$$

is a member of Σ.

Remark: The notation $\xi(\omega)$ would be more appropriate, because $\xi(\omega)$ is a real number. However, we use the conventional notation $X(\omega)$.

To the mathematician a random variable is nothing but a measureable function. To the practician any number resulting from a random experiment is a random variable. Random variables quantify the outcome of an experiment. Instead of results like "A", "B", "C", "head", "tail", "male", "female", "win", "loss", "wet", "dry", "low", "medium", "big" etc. we get numbers. Any number resulting from a geodetic measurement will be considered as a random variable.

Since the above set $\{\omega \in \Omega \mid X(\omega) \leq x\}$ is in Σ, a probability may be assigned to it. We introduce the distribution function

$$F(x) = p\{\omega \in \Omega \mid X(\omega) \leq x\}$$

For the sake of brevity we write this as

$$F(x) = p\{X \leq x\}$$

The following properties of F(x) follow

$$0 \leq F(x) \leq 1$$

$$F(x_1) \leq F(x_2) \ \ldots \text{ if } x_1 \leq x_2 \quad \text{(monotonicity)}$$

$$\lim_{x \to -\infty} F(x) = 0$$

$$\lim_{x \to \infty} F(x) = 1$$

The distribution function makes it easy to specify probabilities for events $a \leq X \leq b$

$$p\{a \leq X \leq b\} = F(b) - F(a)$$

It follows that probabilities can be defined for any event of the form

$$\{X \in A\}$$

where A is a one-dimensional Borel set.

2.2. Probability density function.

Assume that F(x) is differentiable. Then

$$f(x) = F'(x)$$

$$F(x) = \int_{-\infty}^{x} f(y)dy$$

f(x) is called the probability density function of X. It follows that

$$p\{a \leq X \leq b\} = \int_a^b f(x)dx$$

The following properties hold

$$f(x) \geq 0$$

$$\int_{-\infty}^{\infty} f(x)\, dx = 1$$

2.3. n-dimensional random variables.

Let $X_1, X_2, \ldots, X_n$ be n one-dimensional random variables. Any X_i maps Ω into the real line. We may introduce the n-dimensional random variable or random vector

$$X = \begin{bmatrix} X_1 \\ \cdot \\ \cdot \\ \cdot \\ X_n \end{bmatrix}$$

It is a mapping (function) from Ω into R^n. Because any X_i is measureable, any set of the form

$$\{\omega \in \Omega \mid X_1(\omega) \leq x_1, \ldots, X_n(\omega) \leq x_n\}$$

is an element of Σ. We may introduce the joint distribution function of $X_1, \ldots, X_n$ by defining

$$F(x_1, \ldots, x_n) = p\{\omega \in \Omega \mid X_1(\omega) \leq x_1, \ldots, X_n(\omega) \leq x_n\}$$

shorter

$$F(x_1,\ldots,x_n) = p\{X_1 \leq x_1,\ldots,X_n \leq x_n\}$$

$F(x_1,\ldots,x_n)$ is a scalar function of n independent variables having the following properties

$$0 \leq F(x_1,\ldots,x_n) \leq 1$$

$$F(x_1,\ldots,x_i,\ldots,x_n) \leq F(x_1,\ldots,\bar{x}_i,\ldots,x_n) \quad \text{if} \quad x_i \leq \bar{x}_i$$

(monotonicity in each variable separately)

$$\lim_{x_i \to -\infty} F(x_1,\ldots,x_n) = 0, \quad i=1,\ldots,n$$

$$\lim_{\substack{x_1 \to \infty \\ \vdots \\ x_n \to \infty}} F(x_1,\ldots,x_n) = 1$$

The probability of an event of the form

$$p\{a_i \leq X_i \leq b_i,\ i=1,\ldots,n\}$$

is obtained as

$$p\{a_i \leq X_i \leq b_i,\ i=1,\ldots,n\} = \Delta_1\Delta_2 \ldots \Delta_n F(x_1,\ldots,x_n)$$

Here Δ_i is the difference operator

$$\Delta_i \phi(x_1,\ldots,x_i,\ldots,x_n) = \phi(x_1,\ldots,b_i,\ldots,x_n) - \phi(x_1,\ldots,a_i,\ldots,x_n)$$

For example, in case of n=2 we have

$$p\{a_1 \leq x_1 \leq b_1,\ a_2 \leq x_2 \leq b_2\} =$$
$$F(b_1,b_2) - F(a_1,b_2) - F(b_1,a_2) + F(a_1,a_2)$$

The mechanism of Σ-fields and Borel sets propagates the assignment of probabilities to any set of the form

$$\{X \in A\}$$

where A is a Borel set in R^n. (Borel sets in R^n are the smallest Σ-field including all rectangular boxes with faces parallel to the coordinate planes. It turns out that the choice of a coordinate system does not affect the field of Borel sets.)

If $F(x_1,\ldots,x_n)$ is differentiable, we may introduce the density function $f(x_1,\ldots,x_n)$ by

$$f(x_1,\ldots,x_n) = \frac{\partial}{\partial x_1}\ldots\frac{\partial}{\partial x_n} F(x_1,\ldots,x_n)$$

It follows that

$$F(x_1,\ldots,x_n) = \int_{-\infty}^{x_1}\ldots\int_{-\infty}^{x_n} f(y_1,\ldots,y_n)\, dy_1\ldots dy_n$$

The probability

$$p\{a_i \leq X_i \leq b_i,\ i=1,\ldots,n\}$$

is given by the integral

$$\int_{a_1}^{b_1} \ldots \int_{a_n}^{b_n} f(x_1,\ldots,x_n)\, dx_1 \ldots dx_n$$

Moreover, we have for any Borel set

$$p\{X \in A\} = \int\int\ldots\int_A f(x_1,\ldots,x_n)\, dx_1 \ldots dx_n$$

(The integral is always defined in the sense of Lebesgue. In most cases it is identical to the familiar Riemann-integral.)

2.4. Functions of random variables.

Let $\varphi(x_1,\ldots,x_m)$ be a vector valued function mapping R^m into R^n.

$$y = \varphi(x)$$

is a shorthand notation for

$$\begin{aligned} y_1 &= \varphi(x_1,\ldots,x_m) \\ y_2 &= \varphi(x_1,\ldots,x_m) \\ &\ldots\ldots\ldots\ldots \\ y_n &= \varphi(x_1,\ldots,x_m) \end{aligned}$$

The function is called measureable, provided that the sets

$$\{x \in R^m \mid \varphi_1(x) \leq y_1, \ldots, \varphi_n(x) \leq y_n\}$$

are Borel sets in R^m for any choice of $y_1, \ldots, y_n$.

If X is an m-dimensional random variable, then

$$Y = \varphi(X)$$

is an n-dimensional random variable which is the image of X under the mapping φ.

If $F(x_1, \ldots, x_m)$ is the distribution function of X, then the distribution function $G(y_1, \ldots, y_n)$ of Y may in principle be deduced from

$$G(y_1, \ldots, y_n) = p\{x \in R^m \mid \varphi_1(x) \leq y_1, \ldots, \varphi_n(x) \leq y_n\}$$

The probability of the Borel set on the right hand side may be deduced from $F(x_1, \ldots, x_m)$.

Remark: If n=m and if the mapping $y = \varphi(x)$ is one to one and differentiable, and if X has a probability density $f(x_1, \ldots, x_n)$, then also Y has a probability density $g(y_1, \ldots, y_n)$ which is given by

$$g(y_1, \ldots, y_n) = f(x_1, \ldots, x_n) \left| \frac{\partial \varphi}{\partial x} \right|^{-1}$$

Here

$$\left|\frac{\partial \varphi}{\partial x}\right| = \begin{vmatrix} \frac{\partial \varphi_1}{\partial x_1} & \cdots\cdots & \frac{\partial \varphi_1}{\partial x_n} \\ \cdots\cdots & & \cdots\cdots \\ \cdots\cdots & & \cdots\cdots \\ \cdots\cdots & & \cdots\cdots \\ \frac{\partial \varphi_n}{\partial x_1} & \cdots\cdots & \frac{\partial \varphi_n}{\partial x_n} \end{vmatrix}$$

is the Jacobian determinant of the mapping. The proof follows from the familiar rule of substituting variables in an n-dimensional integral.

2.5. Marginal distribution.

Let

$$X = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} .$$

be a 2-dimensional random variable. Let $F(x_1, x_2)$ be the distribution function and $f(x_1, x_2)$ the density. Suppose that we are interested in the distribution of X_1 alone. We denote by $F_{(1)}(x_1)$ the distribution of X_1 and call it the marginal distribution. Similarly, we call the corresponding density $f_{(1)}(x_1)$ the marginal density. Obviously we have

$$F_{(1)}(x_1) = p\{X_1 \leq x_1\} = p\{X_1 \leq x_1,\ X_2 \leq \infty\} = F(x_1, \infty)$$

The corresponding density is obtained as

$$f_{(1)}(x_1) = \int_{-\infty}^{\infty} f(x_1, x_2)\, dx_2$$

The procedure generalizes to higher dimensions in an obvious way. Let X be n-dimensional and partitioned as

$$X = \begin{bmatrix} X_1 \\ \cdot \\ \cdot \\ X_m \\ \hline X_{m+1} \\ \cdot \\ \cdot \\ X_n \end{bmatrix} = \begin{bmatrix} X_{(1)} \\ X_{(2)} \end{bmatrix} \text{, say}$$

The marginal distribution function of the m-dimensional random variable $X_{(1)}$ is obtained as

$$F_{(1)}(x_1, \ldots, x_m) = F(x_1, \ldots, x_m, \infty, \ldots, \infty)$$

The density follows from

$$f_{(1)}(x_1, \ldots, x_m) = \int_{-\infty}^{\infty} \ldots \int_{-\infty}^{\infty} f(x_1, \ldots, x_m, x_{m+1}, \ldots, x_n)\, dx_{m+1} \ldots dx_n$$

2.6. Stochastic independence.

Assume again a 2-dimensional random variable

$$X = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}$$

Suppose that the density decomposes as

$$f(x_1,x_2) = f_{(1)}(x_1)\, f_{(2)}(x_2)$$

The two 1-dimensional random variables X_1, X_2 are then called stochastically independent. The probability of the joint event $\{X_1 \in A_1,\ X_2 \in A_2\}$ is

$$p\{X_1 \in A_1,\ X_2 \in A_2\} = \int_A \int_A f(x_1,x_2)\, dx_1 dx_2 =$$
$$= \int_A \int_A f_{(1)}(x_1) f_{(2)}(x_2)\, dx_1 dx_2 = \int_A f_{(1)}(x_1)\, dx_1 * \int_A f_{(2)}(x_2)\, dx_2 =$$
$$= p\{X_1 \in A_1\}\, p\{X_2 \in A_2\}$$

This allows one to view X_1 and X_2 as the outcomes of two completely independent random experiments. There is no coupling between these two experiments. Knowing the outcome of experiment 1 tells us nothing about the outcome of experiment 2.

The concept of stochastic independence carries over to more than two dimensions. Represent an n-dimensional random variable as

$$X = \left[\begin{array}{c} X_1 \\ \cdot \\ \cdot \\ X_m \\ \hline X_{m+1} \\ \cdot \\ \cdot \\ X_n \end{array}\right] = \begin{bmatrix} X_{(1)} \\ X_{(2)} \end{bmatrix}$$

If the density decomposes as

$$f(x_1,\ldots,x_m,x_{m+1},\ldots,x_n) = f_{(1)}(x_1,\ldots,x_m)\, f_{(2)}(x_{m+1},\ldots,x_n)$$

Then the two subvectors $X_{(1)}, X_{(2)}$ are called stochastically independent. Again it holds that

$$p\{X_{(1)} \in A_1,\ X_{(2)} \in A_2\} = p\{X_{(1)} \in A_1\}\, p\{X_{(2)} \in A_2\}$$

for any Borel sets A_1, A_2 of appropriate dimensions.

An interesting special case arises if

$$f(x_1,\ldots,x_n) = f_{(1)}(x_1)\, f_{(2)}(x_2) \ldots f_{(n)}(x_n)$$

The components $X_1, \ldots, X_n$ are then mutually independent. We have

$$p\{X_1 \in A_1, \ldots, X_n \in A_n\} = p\{X_1 \in A_1\}\, p\{X_2 \in A_2\} \ldots p\{X_n \in A_n\}$$

Another point of interest is the following one. Suppose that the two subvectors $X_{(1)}$, $X_{(2)}$ of X are stochastically independent. Let

$$Y_{(1)} = \varphi_1(X_{(1)})$$
$$Y_{(2)} = \varphi_2(X_{(2)})$$

be two functions, where $Y_{(i)}$ depends on $X_{(i)}$ only. Then $Y_{(1)}$, $Y_{(2)}$ are also stochastically independent.

3. Expectation, variances and covariances.

3.1. Expectation of a one-dimensional random variable.

Let X be a 1-dimensional random variable having a probability density f(x). The expectation E(X) is defined as

$$E(X) = \int_{-\infty}^{\infty} x\, f(x)\, dx$$

Obviously E(X) represents some mean value of the random variable X. The various possible outcomes of X are averaged in agreement with f(x) as weight function. If a random variable is observed many times, and if the arithmetic mean of all outcomes is taken, one expects that the arithmetic mean is very close to E(X).

The above integral may not exist. However, in our applications existence is always tacitly assumed.

Let the one-dimensional random variable Y be a function of X

$$Y = \varphi(X)$$

Y has an expectation of its own. It may be defined in two different ways.

(1) The density g(y) of Y may be calculated as indicated in section 2.4. One then defines

$$E(Y) = \int_{-\infty}^{\infty} y\, g(y)\, dy$$

(2) One may directly define

$$E(Y) = \int_{-\infty}^{\infty} \varphi(x) f(x)\, dx$$

It may be shown that both definitions coincide. The equivalence proof is easy if X and Y have densities and if the mapping $Y = \varphi(X)$ is one to one and smooth.

3.2. The variance of a 1-dimensional random variable.

We denote for the sake of brevity:

$$\mu = E(X) = \int_{-\infty}^{\infty} x\, f(x)\, dx$$

We put $\varphi(X) = (X-\mu)^2$ and calculate the expectation of $\varphi(X)$. The result is the socalled variance of X:

$$\sigma^2(X) = E((X-\mu)^2) = \int_{-\infty}^{\infty} (x-\mu)^2\, f(x)\, dx$$

Obviously $\sigma^2(X)$ measures in some way, how strongly X varies around its expectation. If the density f(x) is very much concentrated around $\mu = E(X)$, we anticipate a small variance. If f(x) has very wide and strong tails, the variance will be large.

3.3. Various kinds of observation errors.

In geodesy an observation ℓ is imagined as the superposition of a true value λ plus an observation error ε.

$$\ell = \lambda + \varepsilon$$

The observation error is a random variable, so is the observation ℓ. The true value λ is an (unknown) constant.

The expectation of the observation error

$$E(\varepsilon)$$

is called the systematic error of the observation. It systematically falsifies the observation, because

$$E(\ell) = E(\lambda+\varepsilon) = \int_{-\infty}^{\infty}(\lambda+\varepsilon)f(\varepsilon)d\varepsilon =$$
$$\lambda\underbrace{\int_{-\infty}^{\infty}f(\varepsilon)d\varepsilon}_{1} + \int_{-\infty}^{\infty}\varepsilon f(\varepsilon)d\varepsilon = \lambda + E(\varepsilon)$$

The underlying assumption in least squares adjustment is

$$E(\varepsilon) = 0$$

This assumption is frequently violated. Hence least squares adjustment is not always optimal.

The variance of ε

$$\sigma^2(\varepsilon) = \int_{-\infty}^{\infty} (\varepsilon - E(\varepsilon))^2 \, f(\varepsilon) \, d\varepsilon$$

is called mean square error. The square root

$$\sigma(\varepsilon) = \sqrt{\int_{-\infty}^{\infty} (\varepsilon - E(\varepsilon))^2 \, f(\varepsilon) \, d\varepsilon}$$

is called root mean square error. The variance of ℓ is also $\sigma^2(\varepsilon)$, because

$$\begin{aligned} \sigma^2(\ell) &= \int_{-\infty}^{\infty} [\lambda+\varepsilon-(\lambda+E(\varepsilon))]^2 \, f(\varepsilon) \, d\varepsilon \\ &= \int_{-\infty}^{\infty} (\varepsilon-E(\varepsilon))^2 \, f(\varepsilon) \, d\varepsilon = \sigma^2(\varepsilon) \end{aligned}$$

3.4. Simple computational rules for $E(X)$, $\sigma^2(X)$.

Some of these rules have been used in the previous subsection. Let c, c_1, c_2 denote any constants. Then:

$$\begin{aligned} E(cX) &= c\,E(X) \\ E(c+X) &= c + E(X) \\ E(c_1+c_2X) &= c_1 + c_2E(X) \\ \sigma^2(cX) &= c^2\sigma^2(X) \\ \sigma(cX) &= c\,\sigma(X) \\ \sigma^2(c+X) &= \sigma^2(X) \\ \sigma^2(c_1+c_2X) &= c_2{}^2\sigma^2(X) \end{aligned}$$

3.5. The case of higher dimensional random variables.

Let $X = (X_1,\ldots,X_m)^T$ be an m-dimensional random variable. Let $f(x_1,\ldots,x_m)$ be the joint density function. The expectation of X is defined as a vector

$$E(X) = \begin{bmatrix} E(X_1) \\ E(X_2) \\ . \\ . \\ E(X_m) \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ . \\ . \\ \mu_m \end{bmatrix}$$

With

$$E(X_i) = \mu_i = \int_{-\infty}^{\infty} \ldots \int_{-\infty}^{\infty} x_i f(x_1,\ldots,x_m)\, dx_1 \ldots dx_m$$
$$= \int_{-\infty}^{\infty} x_i f_{(i)}(x_i)\, dx_i$$

Here

$$f_{(i)}(x_i) = \int_{-\infty}^{\infty} f(x_1,\ldots,x_{i-1},x_i,x_{i+1},\ldots,x_m)\, dx_1 \ldots dx_{i-1}, dx_{i+1} \ldots dx_m$$

is the socalled marginal density of x_i. It is the density of X_i considered as a one-dimensional random variable. Confer section 2.5.

Let

$$y = \varphi(x)$$

denote a mapping from R_m into R_n. In component notation we have

$$y_1 = \varphi_1(x_1,\ldots,x_m)$$
$$y_2 = \varphi_2(x_1,\ldots,x_m)$$
$$\ldots\ldots\ldots\ldots$$
$$y_n = \varphi_n(x_1,\ldots,x_m)$$

The mapping also maps the random variable X onto the random variable Y.

$$Y = \varphi(X)$$

The expectation of Y may be computed in two different ways:

(1) The joint density $g(y_1,\ldots,y_n)$ of Y may be calculated. One then defines

$$E(Y) = \begin{bmatrix} E(Y_1) \\ \vdots \\ E(Y_n) \end{bmatrix} = \begin{bmatrix} \nu_1 \\ \vdots \\ \nu_n \end{bmatrix}$$

with

$$E(Y_i) = \int_{-\infty}^{\infty}\ldots\int_{-\infty}^{\infty} y_i g(y_1,\ldots,y_n)\, dy_1\ldots dy_n$$

$$= \int_{-\infty}^{\infty} y_i g_{(i)}(y_i)\, dy_i$$

Here $g_{(i)}$ is again the marginal density of the component Y_i.

(2) One defines directly

$$E(Y_i) = \int_{-\infty}^{\infty} \ldots \int_{-\infty}^{\infty} \varphi_i(x_1, \ldots, x_m)\, dx_1 \ldots dx_m$$

Both definitions are consistent.

3.6. Covariance matrix.

The covariance matrix of the random vector X is defined as

$$\Sigma = \Sigma(X) = \begin{bmatrix} \sigma_{11} & \cdots & \sigma_{1m} \\ \vdots & & \vdots \\ \sigma_{m1} & \cdots & \sigma_{mm} \end{bmatrix}$$

with

$$\sigma_{ij} = E\{(X_i - \mu_i)(X_j - \mu_j)\} =$$

$$= \int_{-\infty}^{\infty} \ldots \int_{-\infty}^{\infty} (x_i - \mu_i)(x_j - \mu_j)\, f(x_1 \ldots x_m)\, dx_1 \ldots dx_m$$

The diagonals

$$\sigma_{ii} = \sigma^2(X_i) = E((X_i - \mu_i)^2)$$

are just the variances of the individual components of X considered as onedimensional random variables. The off-diagonals

$$\sigma_{ij} = E\{(X_i - \mu_i)(X_j - \mu_j)\}$$

are something new and deserve discussion. Clearly, σ_{ij} measures in some way a coupling between the deviations of X_i from its expectation μ_i and the deviations of X_j from its expectation μ_j. If X_i and X_j have a tendency to deviate either both positively or both negatively from their expectations then σ_{ij} will be positive. This does not mean that a positive $X_i - \mu_i$ cannot occur together with a negative $X_j - \mu_j$. However, in the majority of cases the signs will be coupled as indicated. Similarly, σ_{ij} will be negative, if a positive $X_i - \mu_i$ prefers to be coupled to a negative $X_j - \mu_j$ and vice versa.

3.7. Propagation of expectations and covariances.

It suffices to consider linear inhomogeneous mappings

$$Y = AX + b$$

Here X is m-dimensional and Y is n-dimensional. A is a known n×m matrix and b is a known n-vector. From the linearity of integrals we derive at once the following important relations.

$$E(Y) = E(AX+b) = A\,E(X) + b$$
$$\Sigma(Y) = \Sigma(AX+b) = A\,\Sigma(X)\,A^T$$

Proof of the first law.

$$E(Y_i) = E\left(\sum_{j=1}^{m} a_{ij}X_j + b_i\right) =$$

$$= \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\left(\sum_j a_{ij}x_j + b_i\right) f(x_1,\ldots,x_m)\, dx_1\ldots dx_m$$

$$= \sum_{j=1}^{m} a_{ij} \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} x_j f(x_1,\ldots,x_m)\, dx_1\ldots dx_m +$$

$$+ b_i \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} f(x_1,\ldots,x_m)\, dx_1\ldots dx_m = \sum_{j=1}^{m} a_{ij}E(X_j) + b_i$$

as was to be shown.

Proof of the second law.

Denote

$$\Sigma' = \Sigma(Y) = \begin{bmatrix} \sigma'_{11} & \cdots & \sigma'_{1n} \\ \vdots & & \vdots \\ \sigma'_{n1} & \cdots & \sigma'_{nn} \end{bmatrix}$$

Then

$$\sigma'_{ij} = E\left\{(Y_i - \nu_i)(Y_j - \nu_j)\right\} =$$

$$= E\left\{\left(\sum_{k=1}^{n} a_{ik}X_k + b_i - \sum_{k=1}^{m} a_{ik}\mu_k - b_i\right)\left(\sum_{l=1}^{m} a_{jl}X_l + b_j - \sum_{l=1}^{m} a_{jl}\mu_l - b_j\right)\right\}$$

$$= E\left\{\sum_k a_{ik}(X_k - \mu_k)\sum_l a_{jl}(X_l - \mu_l)\right\}$$

$$= \sum_{kl} a_{ik}a_{jl}E\left\{(X_k - \mu_k)\cdot(X_l - \mu_l)\right\} = \sum_{kl} a_{ik}a_{jl}\sigma_{kl}$$

This is equivalent to

$$\Sigma' = A\Sigma A^T$$

which was to be shown.

Remark: The second law is called the law of propagation of covariances. In geodesy it is usually and simply called "the" error propagation law.

3.8. Important special cases.

If X_1 and X_2 are two random variables having a joint distribution then

$$E(X_1 + X_2) = E(X_1) + E(X_2)$$

Also

$$E\{\lambda_1 X_1 + \lambda_2 X_2\} = \lambda_1 E(X_1) + \lambda_2 E(X_2)$$

This is the linearity property of the expectation.

If X is a random vector, and if $\Sigma(X)$ is of diagonal form

$$\Sigma(X) = \begin{bmatrix} \sigma_{11} & 0 & \dots & 0 \\ \cdot & \sigma_{22} & & \cdot \\ \cdot & & & \cdot \\ \cdot & & & \cdot \\ 0 & \dots & \dots & \sigma_{nn} \end{bmatrix}$$

Then the components of X are called mutually uncorrelated. If Y is a onedimensional function

$$Y = a_1 x_1 + \ldots + a_n x_n = a^T X$$

Then

$$\sigma^2(Y) = a^T \Sigma(X) a = \sum_{i=1}^{n} a_i^2 \sigma^2(X_i)$$

3.9. Zero correlation and stochastic independence.

Suppose that X is 2-dimensional

$$X = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}$$

Let the covariance matrix be

$$\Sigma(X) = \begin{bmatrix} \sigma_{11} & 0 \\ 0 & \sigma_{22} \end{bmatrix}$$

The two components are then uncorrelated.

Assume now that X_1, X_2 are stochastically independent. As we know from section 2.6, this is equivalent to

$$f(x_1, x_2) = f_{(1)}(x_1)\, f_{(2)}(x_2)$$

We show that stochastic independence implies zero correlation:

$$\sigma_{12} = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x_1-\mu_1)(x_2-\mu_2)\, f(x_1,x_2)\, dx_1 dx_2, \quad \mu_i = E(x_i), \quad i=1,2$$

$$\sigma_{12} = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x_1-\mu_1)(x_2-\mu_2)\, f_{(1)}(x_1)\, f_{(2)}(x_2)\, dx_1 dx_2$$

$$= \int_{-\infty}^{\infty} (x_1-\mu_1)\, f_{(1)}(x_1)\, dx_1 \int_{-\infty}^{\infty} (x_2-\mu_2)\, f_{(2)}(x_2)\, dx_2$$

$$= E(X_1-\mu_1)\, E(X_2-\mu_2) = 0.0 = 0$$

The converse is not true. Zero correlation does not necessarily imply stochastic independence.

The concept of zero correlation generalizes to more than two dimensions. Represent an n-dimensional random vector as

$$X = \left[\begin{array}{c} X_1 \\ \cdot \\ \cdot \\ X_m \\ \hline X_{m+1} \\ \cdot \\ \cdot \\ X_n \end{array}\right] = \left[\begin{array}{c} X_{(1)} \\ X_{(2)} \end{array}\right]$$

Suppose that the covariance matrix looks as

$$\Sigma(X) = \left[\begin{array}{ccc|ccc} \sigma_{11} & \cdots & \sigma_{1m} & & & \\ \vdots & & \vdots & & 0 & \\ \sigma_{m1} & \cdots & \sigma_{mm} & & & \\ \hline & & & \sigma_{m+1,m+1} & \cdots & \sigma_{m+1,n} \\ & 0 & & \vdots & & \vdots \\ & & & \sigma_{n,m+1} & \cdots & \sigma_{nn} \end{array}\right] = \left|\begin{array}{cc} \Sigma_{11} & 0 \\ 0 & \Sigma_{22} \end{array}\right|$$

The two subvectors $X_{(1)}$, $X_{(2)}$ are then called uncorrelated.

In section 2.6 we called $X_{(1)}$, $X_{(2)}$ stochastically independent if

$$f(x_1,\ldots,x_n) = f_{(1)}(x_1,\ldots,x_m)\, f_{(2)}(x_{m+1},\ldots,x_n)$$

One shows again that stochastic independence implies zero correlation. The converse is, however, generally not true.

4. The Gauss-Markoff model of least squares adjustment.

4.1. The stochastic model.

Remark on notation: From now on it will be completely impossible to adhere to the convention of using Latin letters for vectors and Greek letters for coordinates.

Let $\mathbf{l}$ be the vector of observations. Its components are denoted l_i, as in section A.6.1.

$$\mathbf{l} = \begin{bmatrix} l_1 \\ l_2 \\ . \\ . \\ l_n \end{bmatrix}$$

$\mathbf{l}$ is viewed as an n-dimensional random variable. In the mathematical sense, $\mathbf{l}$ comprises n measurable functions $l_i(\omega)$ mapping the set Ω introduced in section 1.2 into R^n. The image space R^n will be denoted L in the following. It is a vector space. Sometimes L is called sample space or space of realizations of $\mathbf{l}$. Once the observations are taken, the result are n numbers, the coordinates of a vector in L. Although it is logically unsatisfactory, this vector will occasionally be denoted by the same letter $\mathbf{l}$.

We introduce the vector ε of observation errors and the vector λ of "true" observables by the equation

$$\mathbf{l} = \lambda + \varepsilon$$

The components of λ are constants which are generally unknown. They are true angles, distances, height differences, i.e. observable quantities unaffected by observation errors. As a vector of constants, λ can be viewed as a vector in the n-dimensional space L. Although λ is generally unknown, some a priori information on it is available. It is known that the n unknown quantities λ_i are linearly expressible in terms of m unknown parameters $x_1, \ldots, x_m$, whereby $m < n$.

$$\lambda_1 = a_{11}x_1 + a_{12}x_2 + \ldots + a_{1m}x_m$$
$$\lambda_2 = a_{21}x_1 + a_{22}x_2 + \ldots + a_{2m}x_m$$
$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$
$$\lambda_n = a_{n1}x_1 + a_{n2}x_2 + \ldots + a_{nm}x_m$$

Shortly

$$\lambda = Ax$$

with

$$A = \begin{bmatrix} a_{11} & \cdots & a_{1m} \\ \cdot & & \cdot \\ \cdot & & \cdot \\ \cdot & & \cdot \\ \cdot & & \cdot \\ a_{n1} & \cdots & a_{nm} \end{bmatrix} \qquad x = \begin{bmatrix} x_1 \\ \cdot \\ \cdot \\ \cdot \\ x_m \end{bmatrix}$$

The n×m matrix A is assumed of rank m. Note that we use Latin letters now for the elements of A and the coordinates of x.

Example: Consider a leveling network involving the stations P_0, P_1, P_2, P_3. The height $H_0 = 0$ of P_0 is known. The heights H_1, H_2, H_3 of stations P_1, P_2. P_3 are unknown. Assume height difference measurements ℓ_{01}, ℓ_{03}, ℓ_{12}, ℓ_{13}, ℓ_{23}.

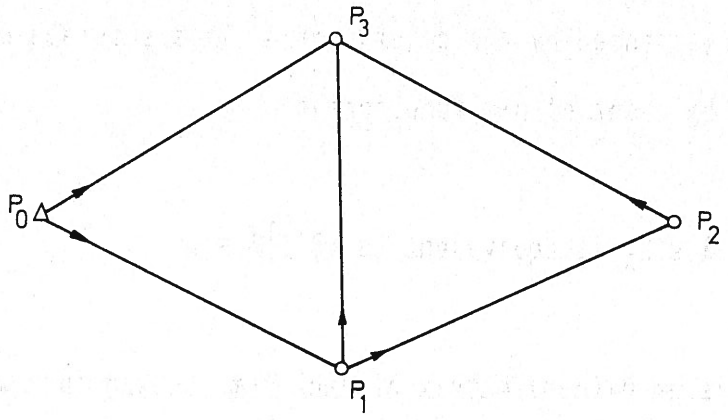


Fig. 4.1.

The true (unknown) height differences are

$$h_{01} = H_1 - H_0 = H_1$$
$$h_{03} = H_3 - H_0 = H_3$$
$$h_{12} = H_2 - H_1$$
$$h_{13} = H_3 - H_1$$
$$h_{23} = H_3 - H_2$$

Thus we have

$$\begin{bmatrix} h_{01} \\ h_{03} \\ h_{12} \\ h_{13} \\ h_{23} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ -1 & 1 & 0 \\ -1 & 0 & 1 \\ 0 & -1 & 1 \end{bmatrix} \begin{bmatrix} H_1 \\ H_2 \\ H_3 \end{bmatrix}$$

corresponding to $\lambda = Ax$.

Remark: The requirement $\lambda = Ax$ is equivalent to restricting the vector λ to the

subspace L_A spanned by the columns of A. This subspace could equally well be described by a set of n-m functionals:

$$\lambda \in L_A \text{ is equivalent to } (B')^T\lambda = 0$$

where B' is an $n\times(n-m)$ matrix of rank n-m, chosen in a way that

$$(B')^T A = 0$$

This would lead to the concept of adjustment by conditions.

Example: Refering to the above introduced leveling network, we have two condition equations of the form

$$h_{01} + h_{13} - h_{03} = 0$$
$$h_{12} + h_{23} - h_{13} = 0$$

i.e.

$$\begin{bmatrix} 1 & -1 & 0 & 1 & 0 \\ 0 & 0 & 1 & -1 & 1 \end{bmatrix} \begin{bmatrix} h_{01} \\ h_{03} \\ h_{12} \\ h_{13} \\ h_{23} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

This corresponds to

$$B^T\lambda = 0$$

We adhere to the model of adjustment by parameters x in

$$\lambda = Ax$$

Recall the representation of the observations ℓ in terms of true observables λ and observation errors ε.

$$\ell = \lambda + \varepsilon$$

We now introduce the important requirement

$$E(\varepsilon) = 0$$

It implies

$$E(\ell) = \lambda$$

or

$$E(\ell) = Ax$$

The requirement $E(\varepsilon) = 0$ means that the systematic error of the observations is assumed as zero. The observations are "unbiased". The postulate of unbiased observations is far reaching in theory. Unfortunately it is rarely fulfilled in practice. Many difficulties encountered during the practical application of least squares adjustment are caused by the failure of the observations to be truly unbiased.

We assume the covariance matrix of ε in the form

$$\Sigma(\varepsilon) = Q\sigma^2$$

Here Q is a known symmetric and positive definite n×n matrix. In most applications Q will be a diagonal matrix. The scalar factor σ^2 is called mean square unit weight error. It may be assumed either as known or as unknown.

Since the random vectors $\mathbf{l}$ and ε differ only by the constant vector λ, i.e. $\mathbf{l} = \lambda+\varepsilon$, the covariance matrix of $\mathbf{l}$ is identical to that of ε:

$$\Sigma(\mathbf{l}) = \Sigma(\varepsilon) = Q\sigma^2$$

We summarize the basic assumptions of the Gauss-Markoff model as follows

$$\boxed{\begin{aligned} E(\mathbf{l}) &= Ax \\ \Sigma(\mathbf{l}) &= Q\sigma^2 \end{aligned}}$$

The vector $\mathbf{l}$ is the vector of observations. The known matrix A is sometimes called design matrix. The unknown vector x is the vector of parameters. The known matrix Q is positive definite. Its inverse

$$P = Q^{-1}$$

is called the matrix of observational weights. The scalar σ^2 is either known or unknown. It is called the mean square unit weight error.

4.2. Unbiased estimates.

Let φ denote a linear functional on L_A. Because any vector in L_A is identified by its coordinates $x = (x_1,\ldots,x_m)^T$ with respect to the bases represented by the columns of A, the functional φ may be represented as

$$\varphi = \varphi^T x = \varphi_1 x_1 + \varphi_2 x_2 + \ldots + \varphi_m x_m$$

Thus a functional on L_A is a linear homogeneous function of the unknown parameters.

Example: Refering to the above introduced leveling network, we have an example for a functional by

$$\varphi^T H = \left(\frac{1}{4} \frac{1}{4} \frac{1}{4} \right) \begin{bmatrix} H_1 \\ H_2 \\ H_3 \end{bmatrix} = = \frac{1}{4} (H_1 + H_2 + H_3)$$

It is the mean value of the heights $H_0 = 0$ and H_1, H_2, H_3. Another example is simply

$$\varphi^T H = \left(0\ 1\ 0 \right) \begin{bmatrix} H_1 \\ H_2 \\ H_3 \end{bmatrix} = H_2$$

the height of the station H_2.

A third example is:

$$\varphi^T H = (0 \; -1 \; 1) \begin{bmatrix} H_1 \\ H_2 \\ H_3 \end{bmatrix} = H_3 - H_2 = h_{23}$$

the height difference $H_3 - H_2$.

Example: Consider a network adjustment. The observations are angles, distances, azimuths, Doppler positions etc. The vector ℓ represents the observation increments after linearization of the observation equations. The parameters x represent coordinate increments. A functional φ may be the (linearized) distance between two remote stations. Alternatively it may refer to an azimuth between any pair of stations, or to an angle between a triplet of stations. Also the area defined by a polygon whose corners are a subset of all stations is (after linearization) a functional of the considered type.

Remark: It is important to note that also any component x_i of the parameter vector x may be viewed as a functional on L_A. Recall section 2.8 where we pointed out that coordinates may be viewed as functionals. The functional x_i is represented by

$$\varphi^T = (0,\ldots,0,1,0,\ldots,0)$$

i.e. by the i-th row of the m×m unit matrix.

Remark: It is further important to note that any component λ_i of the vector $\lambda = E(\ell)$ may be viewed as a functional on L_A. Indeed we have

$$\lambda_i = a_{i1}x_1 + \ldots + a_{im}x_m$$

This means that λ_i is represented by

$$\varphi^T = (a_{i1},\ldots,a_{im})$$

i.e. by the i-th row of the matrix A.

The parameters x are unknown. Hence the value of the functional $\varphi(x) = \varphi^T x$ is unknown. The purpose of statistical estimation is to give estimates of $\varphi = \varphi(x)$ in terms of the observed values of ℓ. Such an estimate is denoted $\hat{\varphi}$. It is a function of the random vector ℓ. Thus $\hat{\varphi}$ is a random variable. It has an expectation $E(\hat{\varphi})$ and a variance $\sigma^2(\hat{\varphi})$.

We restrict ourselves to linear estimates

$$\hat{\varphi} = \beta_1\ell_1 + \beta_2\ell_2 + \ldots + \beta_n\ell_n$$

Shortly

$$\hat{\varphi} = \beta^T\ell, \quad \beta = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_n \end{bmatrix}$$

Remark: The vector β^T can be viewed as the 1×n matrix representation of a linear

functional defined on L.

The expectation of $\hat{\varphi}$ is

$$E(\hat{\varphi}) = E(\beta^T \ell) = \beta^T E(\ell) = \beta^T \lambda = \beta^T A x$$

It is seen that the expectation $E(\hat{\varphi})$ is a linear form in the unknowns x. The requirement

$$E(\hat{\varphi}) = \beta^T A x = \varphi^T x \quad \text{for any } x$$

results in the fundamental concept of a linear unbiased estimate (abbreviated LUE).

The requirement

$$\beta^T A x = \varphi^T x \quad \text{for any } x$$

is equivalent to

$$\beta^T A = \varphi^T$$

4.3. Best linear unbiased estimation.

The variance of the linear estimate

$$\hat{\varphi} = \beta^T \ell$$

is given by

$$\sigma^2(\hat{\varphi}) = \beta^T \Sigma(\ell) \beta = \beta^T Q \beta \, \sigma^2$$

We are looking for an unbiased estimate having a minimal variance. It will be denoted $\tilde{\varphi}$ and called best linear unbiased estimate (BLUE).

The BLUE is the solution of the following extremum problem. Find β such that

$$\beta^T Q \beta = \text{Minimum}$$

subject to

$$\beta^T A = \varphi^T$$

The minimum problem is purely algebraic and can be attacked by means of Lagrange-multipliers. However, we may also make use of the fact that the minimum problem was essentially solved in an earlier section 5.5, using a different notation.

As indicated above, nothing can prevent us from viewing β^T as the matrix representer of a linear functional β on L. The space L was called the sample space. It is the space of all possible realizations of the random vector ℓ of observations. Recall that λ was assumed in L and that L_A was assumed as a subspace of L.

We assign a norm to our functional β by putting

$$\|\beta\|^2 = (\beta,\beta) = \beta^T Q \beta$$

Up to the factor σ^2 the norm of the functional b is nothing but the variance of $\beta^T \ell$ viewed as a linear function of the random vector ℓ.

The matrix Q is interpreted as representing the inner product of functionals on L:

$$(\beta,\gamma) = \beta^T Q \gamma$$

In agreement with section A.4.9, Q is called the reproducing kernel of L. It follows that the inner product in L is represented by the matrix

$$P = Q^{-1}$$

The functional $\varphi = \varphi^T x$ is only defined on L_A. The requirement

$$\beta^T A = \varphi^T$$

is equivalent to requiring that the functional β shall coincide with the functional φ if applied to vectors out of L_A. The functional β is an extension of φ. Confer section A.5.5. Among all such functionals we search for one having minimal norm. According to section A.5.5. on projection of functionals the solution is given by a functional $\tilde{\varphi}$ defined on the whole of L, coinciding with φ

on L_A and vanishing on L_B, the orthocomplement of L_A in L:

$$\tilde{\varphi}(\ell) = \varphi(P_A\ell), \qquad \ell \in L$$

Given a vector $\ell \in L$, we form $P_A\ell$. We find the coordinates $\tilde{x}$ of $P_A\ell$:

$$P_A\ell = A\tilde{x}$$

and we form

$$\tilde{\varphi}(\ell) = \varphi^T\tilde{x}$$

The projection $P_A\ell$ is given by

$$P_A\ell = A(A^TPA)^{-1}A^TP\ell$$

Comparing with

$$P_A\ell = A\tilde{x}$$

we recognize

$$\tilde{x} = (A^TPA)^{-1}A^TP\ell$$

Hence

$$\tilde{\varphi}(\ell) = \varphi^T(A^TPA)^{-1}A^TP\ell$$

This is also obtained by the following rule. Solve the normals

$$(A^T P A)\tilde{x} = A^T P \ell$$

and apply the functional φ to the adjusted parameters $\tilde{x}$:

$$\tilde{\varphi}(\ell) = \varphi^T \tilde{x}$$

The minimal variance $\sigma^2(\tilde{\varphi}(\ell))$ of the best estimater for $\varphi = \varphi^T x$ is given by

$$\sigma^2(\tilde{\varphi}(\ell)) = \beta^T Q \beta \, \sigma^2, \quad \beta = \varphi^T (A^T P A)^{-1} A^T P$$

Because $PQ = I$, one obtains

$$\sigma^2(\tilde{\varphi})) = \varphi^T (A^T P A)^{-1} \varphi \, \sigma^2$$

4.4. Error calculus.

The adjusted parameters $\tilde{x}$ are linear functions of the observations

$$\tilde{x} = (A^T P A)^{-1} A^T P \ell = B\ell$$

The expectation of $\tilde{x}$ is x. The estimators $\tilde{x}$ are unbiased

$$\begin{aligned} E(\tilde{x}) &= E(B\ell) = BE(\ell) = BAx \\ &= (A^T P A)^{-1} A^T P A x = x \end{aligned}$$

The $\tilde{x}$ are BLUE for x (i.e. $\tilde{x}_i$ is the BLUE of x_i). This is also seen by putting $\varphi^T = (0,\ldots,0,1,0,\ldots,0)$ where the 1 appears at the i-th position.

The covariance matrix of $\tilde{x}$ is obtained as

$$\begin{aligned}\Sigma(\tilde{x}) &= B\Sigma(\ell)B^T = BQB^T\sigma^2 = \\ &= (A^TPA)^{-1}A^TPQPA(A^TPA)^{-1}\sigma^2 = (A^TPA)^{-1}(A^TPA)(A^TPA)^{-1}\sigma^2 \\ &= (A^TPA)^{-1}\sigma^2\end{aligned}$$

The covariance matrix of the adjusted parameters is the inverse of the normal equation matrix multiplied by σ^2.

Remark: Note the validity of the error propagation law

$$\tilde{\varphi}(\ell) = \varphi^T\tilde{x}$$

Hence

$$\sigma^2(\tilde{\varphi}(\ell)) = \sigma^2(\varphi^T\tilde{x}) = \varphi^T\Sigma(\tilde{x})\varphi = \varphi^T(A^TPA)^{-1}\varphi\,\sigma^2$$

as before.

Instead of a linear functional $\varphi = \varphi^T x$, we now consider a set of p functionals

$$\Phi = \Phi x$$

Here Φ is a p×m matrix

$$\Phi = \begin{bmatrix} \varphi_{11} & \cdots\cdots & \varphi_{1m} \\ \vdots & & \vdots \\ \varphi_{p1} & \cdots\cdots & \varphi_{pm} \end{bmatrix}$$

Any row of Φ represents one linear functional. Hence we immediately obtain that the best unbiased estimators for all functionals comprised by Φ are given by

$$\tilde{\Phi} = \Phi\tilde{x}$$

Their covariance matrix is

$$\Sigma(\tilde{\Phi}) = \Phi\Sigma(\tilde{x})\Phi^T = \Phi(A^TPA)^{-1}\Phi^T\sigma^2$$

One example for Φ is given by

$$\Phi = Ax$$

The best estimators are

$$\tilde{\Phi} = A\tilde{x} = \tilde{l}$$

It is usual to call them "adjusted observations". Their covariance matrix is

$$\Sigma(\tilde{l}) = A(A^TPA)^{-1}A^T\sigma^2$$

This is sometimes called the "a posteriori covariance matrix" of the

observations. The "a priori covariance matrix" is, of course, given by $\Sigma(\underline{l}) = Q\sigma^2$.

The residuals v are defined as the difference between adjusted and observed values.

$$v = \tilde{\underline{l}} - \underline{l} = A\tilde{x} - \underline{l}$$

Inserting for $\tilde{x}$, we obtain

$$v = -(I-A(A^TPA)^{-1}A^TP)\underline{l}$$
$$= -(I-P_A)\underline{l} = -P_B\underline{l}$$

The covariance matrix of v is obtained as

$$\Sigma(v) = P_B Q P_B^T \sigma^2 =$$
$$= (Q-A(A^TPA)^{-1}A^T)\sigma^2$$

One notices:

$$\Sigma(\underline{l}) = \Sigma(\tilde{\underline{l}}) + \Sigma(v)$$

This is the familiar theorem of Pythagoras in a new disguise.

It is also interesting to calculate the common covariance matrix of $\tilde{\ell}$ and v.

$$\tilde{\ell} = A\tilde{x} = A(A^T PA)^{-1}A^T P\ell = P_A\ell$$

$$v = -(I-A(A^T PA)^{-1}A^T P)\ell = -(I-P_A)\ell$$

One finds

$$\Sigma\begin{bmatrix}\tilde{\ell}\\ v\end{bmatrix} = \begin{bmatrix}\Sigma(\tilde{\ell}) & 0\\ 0 & \Sigma(v)\end{bmatrix}$$

where $\Sigma(\tilde{\ell})$ and $\Sigma(v)$ are the expressions derived above. The remarkable thing is the zero correlation between $\tilde{\ell}$ and v. Zero correlation is the stochastic counterpart to the geometric concept of orthogonality.

5. Applications of the error propagation law.

5.1. Triangle with three measured sides.

Consider the triangle of fig. 5.1.

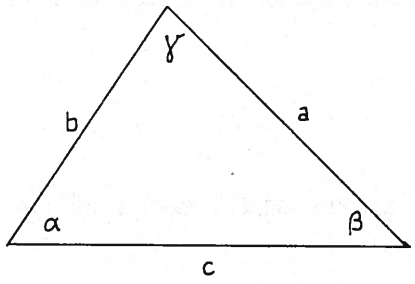


Fig. 5.1.

Assume that a,b,c are measured. Let the covariance matrix of the measurements be

$$\sum \begin{bmatrix} a \\ b \\ c \end{bmatrix} = \begin{bmatrix} m_a^2 & 0 & 0 \\ 0 & m_b^2 & 0 \\ 0 & 0 & m_c^2 \end{bmatrix}$$

What is the variance of the angle α?

By the law of cosine we have

$$a^2 = b^2 + c^2 - 2bc \cos \alpha$$

or

$$\cos \alpha = \frac{b^2 + c^2 - a^2}{2bc}, \quad \alpha = \arccos \frac{b^2 + c^2 - a^2}{2bc}$$

We have expressed α as a function of the observations a,b,c. However, the function is not linear, and does not allow an application of the error propagation law in its standard form. We must linearize the dependence of α on a,b,c.

We assume that m_a, m_b, m_c are small compared to a,b,c. We represent

$$a = a_0 + \Delta a, \quad b = b_0 + \Delta b, \quad c = c_0 + \Delta c$$

Here a_0, b_0, c_0 are fixed values, Δa, Δb, Δc are random variables. Because they differ from a,b,c only by constants a_0, b_0, c_0, the covariance matrix of Δa, Δb, Δc is the same as that one of a,b,c:

$$\sum \begin{bmatrix} \Delta a \\ \Delta b \\ \Delta c \end{bmatrix} = \sum \begin{bmatrix} a \\ b \\ c \end{bmatrix} = \begin{bmatrix} m_a^2 & 0 & 0 \\ 0 & m_b^2 & 0 \\ 0 & 0 & m_c^2 \end{bmatrix}$$

Call

$$\alpha_0 = \arccos \frac{b_0^2 + c_0^2 - a_0^2}{2b_0c_0}$$

and let

$$\alpha = \alpha_0 + \Delta\alpha$$

then

$$a^2 = b^2 + c^2 - 2bc \cos \alpha$$

goes over into

$$(a_0 + \Delta a)^2 = (b_0 + \Delta b)^2 + (c_0 + \Delta c)^2 - 2(b_0 + \Delta b)(c_0 + \Delta c) \cos(\alpha_0 + \Delta \alpha)$$

Applying Taylor's formula, and keeping only the linear terms, gives

$$2a_0 \Delta a = 2b_0 \Delta b + 2c_0 \Delta c - 2(c_0 \Delta b + b_0 \Delta c) \cos \alpha_0 + 2b_0 c_0 \sin \alpha_0 \Delta \alpha$$

The constant terms cancel due to the consistency of a_0, b_0, c_0 and α_0. We solve for $\Delta \alpha$ obtaining

$$\Delta \alpha = \frac{1}{b_0 c_0 \sin \alpha_0} \left\{ a_0 \Delta a - (b_0 - c_0 \cos \alpha_0) \Delta b - (c_0 - b_0 \cos \alpha_0) \Delta c \right\}$$

We abbreviate this as

$$\Delta \alpha = C_{\alpha a} \Delta a + C_{\alpha b} \Delta b + C_{\alpha c} \Delta c$$

This is the desired linearized relationship. We immediately get the variance of $\Delta \alpha$ by

$$m_\alpha^2 = C_{\alpha a}^2\, m_a^2 + C_{\alpha b}^2\, m_b^2 + C_{\alpha c}^2\, m_c^2$$

The geometric meaning of $C_{\alpha a}$, $C_{\alpha b}$, $C_{\alpha c}$ is seen from

$$C_{\alpha a} = \frac{2a_o}{F}, \quad C_{\alpha b} = \frac{2p_{ba}}{F}, \quad C_{\alpha c} = \frac{2p_{ca}}{F}$$

with F being the area of the triangle, and p_{ba} and p_{ca} being the projections of b and c onto a as shown in fig. 5.2.

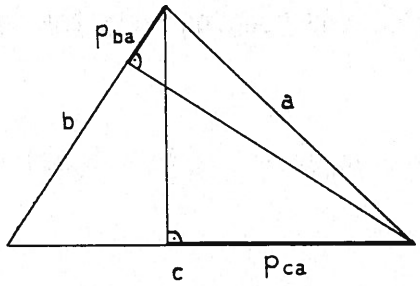


Fig. 5.2.

The covariance matrix of all angles α,β,γ (which is the same as that one of $\Delta\alpha$, $\Delta\beta$, $\Delta\gamma$) is obtained

$$\sum \begin{bmatrix} a \\ b \\ c \end{bmatrix} = \begin{bmatrix} C_{\alpha a} & C_{\alpha b} & C_{\alpha c} \\ C_{\beta a} & C_{\beta b} & C_{\beta c} \\ C_{\gamma a} & C_{\gamma b} & C_{\gamma c} \end{bmatrix} \cdot \begin{bmatrix} m_a^2 & 0 & 0 \\ 0 & m_b^2 & 0 \\ 0 & 0 & m_c^2 \end{bmatrix} \cdot \begin{bmatrix} C_{\alpha a} & C_{\beta a} & C_{\gamma a} \\ C_{\alpha b} & C_{\beta b} & C_{\gamma b} \\ C_{\alpha c} & C_{\beta c} & C_{\gamma c} \end{bmatrix}$$

It is interesting to calculate the variance of

$$\alpha + \beta + \gamma$$

Because this sum equals π, which is a constant, we must have

$$\sigma^2(\alpha+\beta+\gamma) = (1\ 1\ 1)\ \Sigma(\alpha,\beta,\gamma)\ (1\ 1\ 1)^T = 0$$

This is easily verified to be correct. From the geometric meaning of the C's one easily recognizes that e.g.

$$C_{\alpha a} + C_{\beta a} + C_{\gamma a} = 0$$

Remark. If b+c = a, the area F of the triangle becomes zero. The quantities $C_{\alpha a}$, $C_{\alpha b}$, $C_{\alpha c}$ then are infinite. The mean square error m_α^2 becomes infinitely large. This is not entirely meaningful since one could argue that an angle of a triangle is bounded within the interval $[0,\pi]$. The reason for m_α^2 tending to infinity for $b+c \rightarrow a$ is the linearization of the dependency of α on a,b,c. For $b+c \rightarrow a$ the higher order terms are no longer negligible. The degeneracy of a triangle into a line segment is a critical configuration. The angle α becomes very poorly determined. The linearized theory signals this degeneracy. However, it exaggerates somewhat by letting the error of α tend to infinity.

5.2. The first fundamental problem in the plane.

Let a point P with coordinates x_0, y_0 be given. Assume that its coordinates are known and free of any error. Let the distance s and the azimuth α to another point P with unknown coordinates x,y be measured. Let the covariance matrix of s,α be

$$\sum \begin{bmatrix} s \\ \alpha \end{bmatrix} = \begin{bmatrix} m_s^2 & 0 \\ 0 & m_\alpha^2 \end{bmatrix}$$

The coordinates x,y are computed by

$$x = x_0 + s \cos\alpha$$
$$y = y_0 + s \sin\alpha$$

What is the covariance matrix of x,y ? Linearizing the relationship we get

$$\Delta x = \cos\alpha \, \Delta s - s \sin\alpha \, \Delta\alpha$$
$$\Delta y = \sin\alpha \, \Delta s + s \cos\alpha \, \Delta\alpha$$

For simplicity we have refrained from distinguishing between s,α and their approximate values. Applying the error propagation law we find

$$\sum \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} \cos\alpha & -s \sin\alpha \\ \sin\alpha & -s \cos\alpha \end{bmatrix} \cdot \begin{bmatrix} m_s^2 & 0 \\ 0 & m_\alpha^2 \end{bmatrix} \cdot \begin{bmatrix} \cos\alpha & \sin\alpha \\ -s \sin\alpha & s \cos\alpha \end{bmatrix}$$

Calling

$$\sum \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} m_{xx} & m_{xy} \\ m_{yx} & m_{yy} \end{bmatrix}$$

we find

$$m_{xx} = m_s^2 \cos^2\alpha + (s m_\alpha)^2 \sin^2\alpha$$
$$m_{xy} = [m_s^2 - (s m_\alpha)^2] \cos\alpha \sin\alpha$$

$$m_{yy} = m_s^2 \sin^2\alpha + (s m_\alpha)^2 \cos^2\alpha$$

Discussion. We see that x and y are correlated unless

$$m_s = s\, m_\alpha$$

i.e. unless the error of the distance s equals the error of the lateral deviation in P due to the azimuth error. If this equality holds, it also implies

$$m_{xx} = m_{yy}$$

The accuracy of P is equally good in all directions. We shall make this statement precise in the next subsection.

5.3. Error ellipses.

In this subsection we prefer to call the coordinates in the plane x_1 and x_2 instead of x and y. Suppose then that

$$X = (X_1\ X_2)^T$$

are the random coordinates of a point in the plane. Let the covariance matrix of X be

$$\Sigma(X) = M = \begin{bmatrix} m_{11} & m_{12} \\ m_{21} & m_{22} \end{bmatrix}$$

M must be positive semidefinite. We assume that M^{-1} exists, i.e. we assume that M is positive definite.

Definition. Let $x = (x_1, x_2)^T$ be points on the curve

$$x^T M^{-1} x = 1$$

This curve is an ellipse. It is called the error ellipse of the point.

The curve is an ellipse because M^{-1} is positive definite. (The inverse of a positive definite matrix is also positive definite. The proof runs as follows: $x^TM^{-1}x = x^TM^{-1}MM^{-1}x = (M^{-1}x)^TM(M^{-1}x) = y^TMy$, with $y = M^{-1}x$. Now $y^TMy > 0$ because M is positive definite.)

Let

$$\xi = (\xi_1\ \xi_2)^T = (\cos\varphi\ \sin\varphi)^T$$

be a unit vector in the plane. ξ defines a direction which is also given by the direction angle φ. We call

$$\xi^TX = \xi_1X_1 + \xi_2X_2 = X_1\cos\varphi + X_2\sin\varphi$$

the error of the point in the direction ξ. Note: ξ^TX is the projection of X

onto the line having direction ξ.

By the error propagation law we get

$$\sigma^2(\xi^T X) = \xi^T M \xi = m_{11}\xi_1^2 + 2m_{12}\xi_1\xi_2 + m_{22}\xi_2^2 =$$
$$m_{11}\cos^2\varphi + 2m_{12}\cos\varphi\,\sin\varphi + m_{22}\sin^2\varphi = S^2(\varphi), \text{ say}$$

An ellipse is a closed convex curve. Any convex curve has a support function $p(\varphi)$ with respect to a chosen point located in its interior. See fig. 5.3 for the definition of the support function.

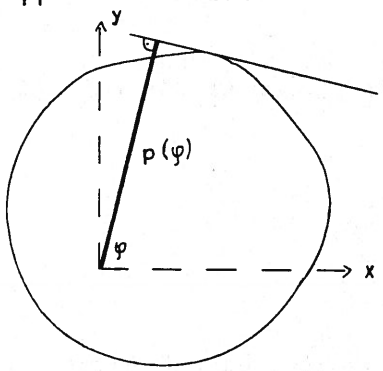


Fig. 5.3.

<u>Theorem.</u> $S(\varphi)$ is the support function of the error ellipse with respect to the center of the ellipse (see fig. 5.4).

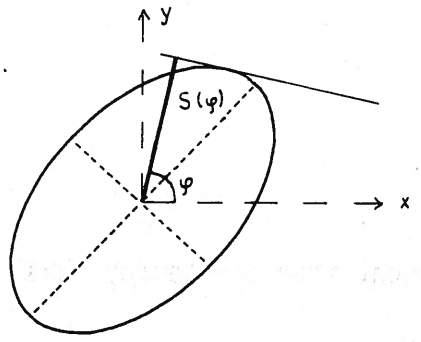


Fig. 5.4.

Proof. Let x_0 be a point on the error ellipse. The tangent line through this point is given by

$$x_0^T M^{-1} x = 1$$

[For a proof note that the gradient of $(x^T M^{-1} x - 1)$ taken at $x = x_0$ is a vector orthogonal to the ellipse. This vector is $2\, M^{-1} x_0$. We may cut its length by 1/2 obtaining $M^{-1} x_0$. Thus the tangent line is $(M^{-1} x_0)^T (x - x_0) = 0$, or $x^T M^{-1} x \;\; x^T M^{-1} x_0 = 0$ or $x_0^T M^{-1} x = 1$, because $x_0^T M^{-1} x_0 = 1$.]

Introducing the normal unit vector of the tangent

$$\xi = \| M^{-1} x_0 \|^{-1} \, M^{-1} x_0$$

we write the tangent line in its "Hesse" form

$$\xi^T x = \| M^{-1} x_0 \|^{-1}$$

we see that

$$p(\varphi) = \| M^{-1} x_0 \|^{-1}$$

is the distance of the tangent from the center. What we want to prove is $p(\varphi) = S(\varphi)$. Now

$$S(\varphi)^2 = \xi^T M \xi = \|M^{-1}x_0\|^{-2}\,(M^{-1}x_0)^T M(M^{-1}x_0) = \|M^{-1}x_0\|^{-2}\,x_0^T M^{-1}x_0 = \|M^{-1}x_0\|^{-2} = p(\varphi)^2$$

because $x_0^T M^{-1} x_0 = 1$. This proves the theorem.

Remark. The theorem and its proof carry over from two to n dimensions with hardly any change. (Of course it takes more than one angle φ to define a direction in n dimensions. A direction in R^n is in the best way defined by a unit vector ξ.)

Remark. The question for the maximal and minimal values of $S(\varphi)$ is equivalent to the question for the direction of the principal axes of the error ellipse. In two dimensions one finds the directions as solutions to the equation

$$\tan 2\varphi = \frac{2\,m_{12}}{m_{11} - m_{22}}$$

If $m_{11} = m_{22}$ then φ becomes indeterminate. The error ellipse becomes a circle.

5.4. Polar survey with redundancy.

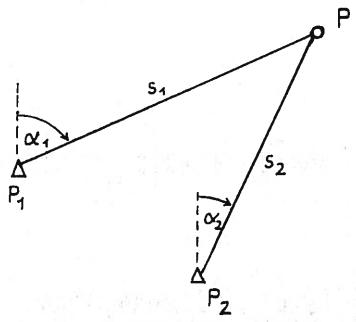


Fig. 5.5.

We return to the problem of section 5.2. However, this time there are two reference points P_1, P_2 whose coordinates are assumed free of any error. The measurements are s_1, α_1, s_2, α_2. Their common covariance matrix be diagonal having the values

$$\Sigma(s_1,\alpha_1,s_2,\alpha_2) = \mathrm{diag}(m_{s_1}^2,\ m_{\alpha_1}^2,\ m_{s_2}^2,\ m_{\alpha_2}^2)$$

5.4.1. The practician's solution.

The practician calculates the coordinates of point P twice, once from P_1 and once from P_2. He then takes the arithmetic mean of the two solutions. Thus he calculates

$$x = \tfrac{1}{2}\{x_1 + s_1\cos\alpha_1 + x_2 + s_2\cos\alpha_2\}$$
$$y = \tfrac{1}{2}\{y_1 + s_1\sin\alpha_1 + y_2 + s_2\sin\alpha_2\}$$

In order to propagate errors we linearize:

$$\Delta x = \tfrac{1}{2}\{\cos\alpha_1\,\Delta s_1 - s_1\sin\alpha_1\,\Delta\alpha_1 + \cos\alpha_2\,\Delta s_2 - s_2\sin\alpha_2\,\Delta\alpha_2\}$$
$$\Delta y = \tfrac{1}{2}\{\sin\alpha_1\,\Delta s_1 + s_1\cos\alpha_1\,\Delta\alpha_1 + \sin\alpha_2\,\Delta s_2 + s_2\cos\alpha_2\,\Delta\alpha_2\}$$

This gives

$$m_{xx} = \tfrac{1}{4}\left\{m_{s_1}^2\cos^2\alpha_1 + (s_1 m_{\alpha_1})^2\sin^2\alpha_1 + m_{s_2}^2\cos^2\alpha_2 + (s_2 m_{\alpha_2})^2\sin^2\alpha_2\right\}$$

$$m_{xy} = \tfrac{1}{4}\left\{[m_{s_1}^2-(s_1 m_{\alpha_1})^2]\cos\alpha_1\sin\alpha_1 + [m_{s_2}^2-(s_2 m_{\alpha_2})^2]\cos\alpha_2\sin\alpha_2\right\}$$

$$m_{yy} = \frac{1}{4}\left\{m_{s_1}^2 \sin^2\alpha_1 + (s_1 m_{\alpha_1})^2 \cos^2\alpha_1 + m_{s_2}^2 \sin^2\alpha_2 + (s_2 m_{\alpha_2})^2 \cos^2\alpha_2\right\}$$

One also recognizes that

$$\Sigma(x,y) = \frac{1}{4}\left[\Sigma_1(x,y) + \Sigma_2(x,y)\right]$$

where $\Sigma_1(x,y)$, $\Sigma_2(x,y)$ are the covariance matrices obtained for P in section 5.2, if the role of P_o is taken either by P_1 or by P_2.

5.4.2. The adjustment expert's solution.

The observation equations (in nonlinear and implicit form) are

$$x = x_1 + (s_1 + v_{s_1}) \cos(\alpha_1 + v_{\alpha_1})$$
$$y = y_1 + (s_1 + v_{s_1}) \sin(\alpha_1 + v_{\alpha_1})$$
$$x = x_2 + (s_2 + v_{s_2}) \cos(\alpha_2 + v_{\alpha_2})$$
$$y = y_2 + (s_2 + v_{s_2}) \sin(\alpha_2 + v_{\alpha_2})$$

We assume approximate values $x^{(o)}$, $y^{(o)}$ for x and y. We introduce coordinate increments Δx, Δy by

$$x = x^{(o)} + \Delta x$$
$$y = y^{(o)} + \Delta y$$

The linearized observation equations are

$$\Delta x = s_1 \cos\alpha_1 - (x^{(o)} - x_1) + \cos\alpha_1 \, v_{s_1} - \sin\alpha_1 s_1 v_{\alpha_1}$$
$$\Delta y = s_1 \sin\alpha_1 - (y^{(o)} - y_1) + \sin\alpha_1 \, v_{s_1} + \cos\alpha_1 s_1 v_{\alpha_1}$$
$$\Delta x = s_2 \cos\alpha_2 - (x^{(o)} - x_2) + \cos\alpha_2 \, v_{s_2} - \sin\alpha_2 s_2 v_{\alpha_2}$$
$$\Delta y = s_2 \sin\alpha_2 - (y^{(o)} - y_2) + \sin\alpha_2 \, v_{s_2} + \cos\alpha_2 s_2 v_{\alpha_2}$$

Although there are adjustment schemes which could handle observation equations in this form, we prefer to solve for v_{s_1}, v_{α_1}, v_{s_2}, v_{α_2}.

Multiplying the first equation by $\cos\alpha_1$, the second one by $\sin\alpha_1$ and adding, we obtain

$$s_1 - (x^{(o)} - x_1) \cos\alpha_1 - (y^{(o)} - y_1) \sin\alpha_1 + v_{s_1} = \cos\alpha_1 \, \Delta x + \sin\alpha_1 \, \Delta y$$

By similar manoeuvres we obtain three more equations

$$\frac{1}{s_1} (x^{(o)} - x_1) \sin\alpha_1 - \frac{1}{s_1} (y^{(o)} - y_1) \cos\alpha_1 + v_{\alpha_1} = \frac{-\sin\alpha_1}{s_1} \Delta x + \frac{\cos\alpha_1}{s_1} \Delta y$$
$$s_2 - (x^{(o)} - x_2) \cos\alpha_2 - (y^{(o)} - y_2) \sin\alpha_2 + v_{s_2} = \cos\alpha_2 \, \Delta x + \sin\alpha_2 \, \Delta y$$
$$\frac{1}{s_2} (x^{(o)} - x_2) \sin\alpha_2 - \frac{1}{s_2} (y^{(o)} - y_2) \cos\alpha_2 + v_{\alpha_2} = \frac{-\sin\alpha_2}{s_2} \Delta x + \frac{\cos\alpha_2}{s_2} \Delta y$$

The weights must be proportional to the reciprocal mean square errors. Thus we use

$$p_{s_1} = m_{s_1}^{-2}, \quad p_{\alpha_1} = m_{\alpha_1}^{-2}$$
$$p_{s_2} = m_{s_2}^{-2}, \quad p_{\alpha_2} = m_{\alpha_2}^{-2}$$

We exhibit the normal equation matrix. Its elements are

$$g_{11} = m_{s_1}^{-2} \cos^2\alpha_1 + (s_1 m_{\alpha_1})^{-2} \sin^2\alpha_1 + m_{s_2}^{-2} \cos^2\alpha_2 + (s_2 m_{\alpha_2})^{-2} \sin^2\alpha_2$$

$$g_{12} = \left\{ m_{s_1}^{-2} - (s_1 m_{\alpha_1})^{-2} \right\} \cos\alpha_1 \sin\alpha_1 + \left\{ m_{s_2}^{-2} - (s_2 m_{\alpha_2})^{-2} \right\} \cos\alpha_2 \sin\alpha_2$$

$$g_{22} = m_{s_1}^{-2} \sin^2\alpha_1 + (s_1 m_{\alpha_1})^{-2} \cos^2\alpha_1 + m_{s_2}^{-2} \sin^2\alpha_2 + (s_2 m_{\alpha_2})^{-2} \cos^2\alpha_2$$

The inverse matrix

$$G^{-1} = \begin{bmatrix} g_{11} & g_{12} \\ g_{21} & g_{22} \end{bmatrix}^{-1}$$

is the covariance matrix of the rigorously adjusted coordinates x,y.

5.4.3. Comparing the two solutions.

Call Σ_p the covariance matrix from the practician's solution and Σ_a that one from the adjustment, i.e. $\Sigma_a = G^{-1}$. Because Σ_a is the covariance of the best linear unbiased estimates, any linear function of the rigorously adjusted coordinates must have a variance less than or equal to the variance of the same linear function applied to the practician's coordinates. This must in particular be true for linear functions

$$\cos\varphi\ x + \sin\varphi\ y$$

which led to the concept of error ellipses. From this one may infer that the

error ellipse of the adjusted coordinates must be situated in the interior of the practician's error ellipse. See fig. 5.6.

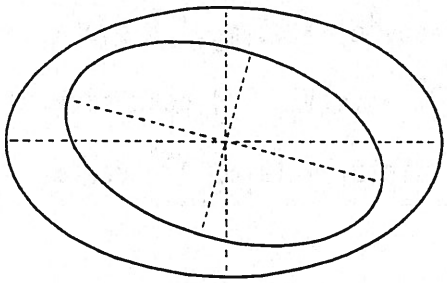


Fig. 5.6.

5.5. An area calculated from polar survey.

Consider the problem illustrated by fig. 5.7.

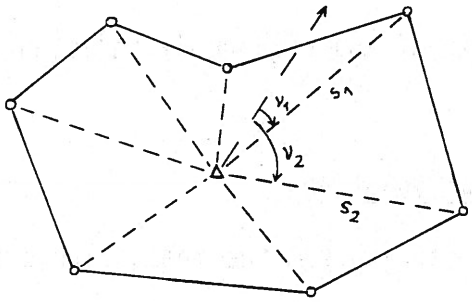


Fig. 5.7.

From a station P_0 whose coordinates in a local system we take as $x_0 = y_0 = 0$, we measure distances s_i and direction angles ν_i to the points on the circumference of a polygon. The direction angles are taken with respect to a local axis as shown in fig. 5.7.

The area is calculated by the following polygon

$$F = \frac{1}{2} \sum_{i=1}^{n} s_i \, s_{i+1} \sin(\nu_{i+1} - \nu_i) \quad \text{(node n+1 equals node 1)}$$

This formula is nonlinear. The linearization is

$$\Delta F = \frac{1}{2} \sum_{i=1}^{n} \{ \Delta s_i \, s_{i+1} \sin(\nu_{i+1} - \nu_i) + \Delta s_{i+1} \, s_i \sin(\nu_{i+1} - \nu_i) - \Delta \nu_i \, s_i \, s_{i+1} \cos(\nu_{i+1} - \nu_i) + \Delta \nu_{i+1} \, s_i \, s_{i+1} \cos(\nu_{i+1} - \nu_i) \}$$

or

$$\Delta F = \frac{1}{2} \sum_{i=1}^{n} \{ \Delta s_i \, [s_{i+1} \sin(\nu_{i+1} - \nu_i) + s_{i-1} \sin(\nu_i - \nu_{i-1})] + s_i \Delta \nu_i \, [-s_{i+1} \cos(\nu_{i+1} \nu_i) + s_{i-1} \cos(\nu_i - \nu_{i-1})] \}$$

$$= \frac{1}{2} \sum_{i=1}^{n} \{ a_i \, \Delta s_i + b_i \, (s_i \Delta \nu_i) \} , \quad \text{say.}$$

Assuming uncorrelated measurement errors with standard deviations m_{s_i}, m_{ν_i}, we find

$$m_F^2 = \frac{1}{4} \sum_{i=1}^{n} [a_i^2 \, m_{s_i}^2 + b_i^2 \, s_i^2 \, m_{\nu_i}^2]$$

The geometric meaning of a_i, b_i is explained in fig. 5.8.

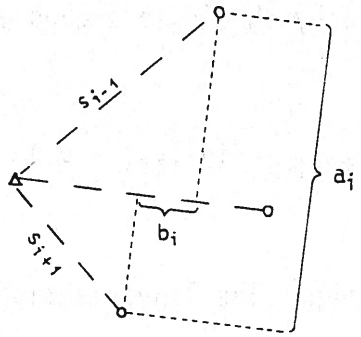


Fig. 5.8.

Thus, while a_i rarely becomes zero, the b_i may vanish under certain symmetry conditions. For example, if the polygon is regular, and if the station P_0 is located at the center, then m_F will depend only on the accuracy of the distances s_i. This again must be seen in the light of the first order Taylor expansion which underlies the error analysis.

5.6. Conventionally adjusted regular traverse.

In fig. 5.9, stations marked by triangles are fixed, those marked by circles are unknown.

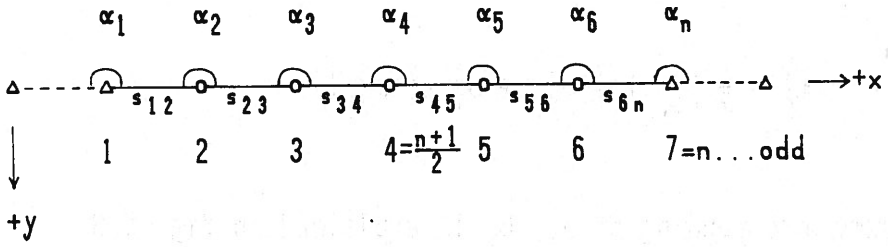


Fig. 5.9.

We assume that the distances $s_{i,i+1}$, i=1,...,n-1 are measured, as well as the angles α_i, i=1,...,n. Let m_s, m_α be the corresponding r.m.s. errors. We are

interested in the accuracy of coordinates ξ,μ of the central station (n+1)/2.

We must express those coordinates ξ,μ in terms of measurements, Because the measurements are redundant, we must know how the traverse is adjusted.

We assume that, as usual, the angles are adjusted first. They are constrained by

$$\alpha_1 + \alpha_2 + \ldots + \alpha_n = n\,\varkappa$$

Due to measurement errors, there will be a discrepancy

$$\alpha_1 + \alpha_2 + \ldots + \alpha_n = n\,\varkappa + w_\alpha$$

It is divided by n, and each angle α_i is replaced by

$$\bar{\alpha}_i = \alpha_i - \frac{w_\alpha}{n}$$

Now the coordinates of the last station are calculated by considering a free traverse leading from station 1 to station n, and having measurements $s_{i,i+1}$ and $\bar{\alpha}_i$.

One gets:

$$x_n = \sum_{i=1}^{n-1} s_{i,i+1} \cos \nu_{i,i+1}, \quad y_n = \sum_{i=1}^{n-1} s_{i,i+1} \sin \nu_{i,i+1}$$

with

$$\nu_{i,i+1} = \sum_{j=1}^{i} \bar{\alpha}_j - i\pi$$

We must linearize this, using approximate values $s_{i,i+1} = 1$, $\nu_{i,i+1} = 0$:

$$\Delta x_n = \sum_{i=1}^{n-1} \Delta s_{i,i+1}, \quad \Delta y_n = \sum_{i=1}^{n-1} \Delta \nu_{i,i+1}$$

For $\Delta \nu_{i,i+1}$, we substitute

$$\Delta \nu_{i,i+1} = \sum_{j=1}^{i} \Delta \alpha_j$$

For $\Delta \bar{\alpha}_j$, we substitute

$$\Delta \bar{\alpha}_j = \Delta \alpha_j - \frac{1}{n} \Delta w_\alpha$$

For Δw_α, we get

$$\Delta w_\alpha = \sum_{k=1}^{n} \Delta \alpha_k$$

Carrying out all these substitutions, we obtain:

$$\Delta y_n = \sum_{i=1}^{n-1} \sum_{j=1}^{i} \left\{ \Delta \alpha_j - \frac{1}{n} \sum_{k=1}^{n} \Delta \alpha_k \right\} =$$

$$= \sum_{j=1}^{n-1} \left\{ \sum_{i=j}^{n-1} 1 \right\} \Delta \alpha_j - \frac{1}{n} \sum_{i=1}^{n-1} \left\{ \sum_{j=1}^{i} 1 \right\} \sum_{k=1}^{n} \Delta \alpha_k$$

$$= \sum_{j=1}^{n-1} (n-j)\,\Delta\alpha_j - \left\{ \frac{1}{n} \sum_{i=1}^{n-1} i \right\} \sum_{k=1}^{n} \Delta\alpha_k$$

$$= \sum_{j=1}^{n} (n-j)\,\Delta\alpha_j - \frac{1}{n}\,\frac{n(n-1)}{2} \sum_{k=1}^{n} \Delta\alpha_k$$

$$= \sum_{j=1}^{n} \left\{ \frac{n+1}{2} - j \right\} \Delta\alpha_j$$

Summarizing:

$$\Delta x_n = \sum_{i=1}^{n-1} \Delta s_{i,i+1}, \quad \Delta y_n = \sum_{i=1}^{n} \left\{ \frac{n+1}{2} - i \right\} \Delta\alpha_i$$

A similar calculation for the midpoint $\frac{n+1}{2}$ gives:

$$\Delta\xi = \Delta x_{(n+1)/2} = \sum_{i=1}^{(n-1)/2} \Delta s_{i,i+1}$$

$$\Delta\eta = \Delta y_{(n+1)/2} = \sum_{i=1}^{(n-1)/2} \sum_{j=1}^{i} \left\{ \Delta\alpha_j - \frac{1}{n} \sum_{k=1}^{n} \Delta\alpha_k \right\}$$

$$= \sum_{j=1}^{(n-1)/2} \left\{ \sum_{i=j}^{(n-1)/2} 1 \right\} \Delta\alpha_j - \frac{1}{n} \left\{ \sum_{i=1}^{(n-1)/2} \sum_{j=1}^{i} 1 \right\} \sum_{k=1}^{n} \Delta\alpha_k$$

$$= \sum_{j=1}^{(n-1)/2} \left\{ \frac{n+1}{2} - j \right\} \Delta\alpha_j - \frac{1}{n}\,\frac{n-1}{2}\,\frac{n+1}{4} \sum_{k=1}^{n} \Delta\alpha_k$$

$$= \sum_{j=1}^{(n-1)/2} \left\{ \frac{3n^2+4n+1}{8n} - j \right\} \Delta\alpha_j - \frac{n^2-1}{8n} \sum_{k=(n+1)/2}^{n} \Delta\alpha_k$$

Summarizing:

$$\Delta\xi = \sum_{i=1}^{(n-1)/2} \Delta s_{i,i+1}$$

$$\Delta\eta = \sum_{i=1}^{(n-1)/2} \left\{ \frac{3n^2+4n+1}{8n} - i \right\} \Delta\alpha_i - \frac{n^2-1}{8n} \sum_{i=(n+1)/2}^{n} \Delta\alpha_i$$

The coordinate discrepancies at station n are now redistributed. This results in adjusted coordinates at the midpoint given by

$$\Delta\tilde{\xi} = \Delta\xi - \frac{1}{2}\Delta x_n = \frac{1}{2} \sum_{i=1}^{(n-1)/2} \Delta s_{i,i+1} - \frac{1}{2} \sum_{i=(n+1)/2}^{n-1} \Delta s_{i,i+1}$$

$$\Delta\tilde{\eta} = \Delta\eta - \frac{1}{2}\Delta y_n = \sum_{i=1}^{(n-1)/2} \left\{ \frac{3n^2+4n+1}{8n} - i - \frac{1}{2}\left(\frac{n+1}{2} - i \right) \right\} \Delta\alpha_i$$

$$+ \sum_{i=(n+1)/2}^{n} \left\{ -\frac{n^2-1}{8n} - \frac{1}{2}\left(\frac{n+1}{2} - i \right) \right\} \Delta\alpha_i$$

$$= \sum_{i=1}^{(n-1)/2} \left\{ \frac{(n+1)^2}{8n} - \frac{i}{2} \right\} \Delta\alpha_i + \sum_{i=(n+1)/2}^{n} \left\{ \frac{(n+1)^2}{8n} - \frac{n-i+1}{2} \right\} \Delta\alpha_i$$

One verifies that the coefficients of $\Delta\alpha_i$ and $\Delta\alpha_{n-i+1}$ are equal. This is desirable from reasons of symmetry.

From the above expressions follow the mean square errors of $\tilde{\xi},\tilde{\eta}$ as

$$m_{\tilde{\xi}}^2 = \frac{n-1}{4} m_s^2$$

$$m_{\tilde{\eta}}^2 = 2 \sum_{i=1}^{(n-1)/2} \left\{ \frac{(n+1)^2}{8n} - \frac{i}{2} \right\}^2 + \left\{ \frac{(n+1)^2}{8n} - \frac{n+1}{4} \right\}^2$$

$$= \frac{1}{2} \sum_{i=1}^{(n-1)/2} \left\{ \frac{(n+1)^2}{4n} - i \right\}^2 + \left\{ \frac{(n+1)(n-1)}{8n} \right\}^2$$

$$= \frac{1}{2} \left\{ \frac{(n+1)^4(n-1)}{32\,n^2} - \frac{(n+1)^2}{2n} \cdot \frac{(n+1)(n-1)}{8} + \frac{n-1}{2} \frac{n+1}{2} \frac{n}{6} \right\} +$$

$$\left\{ \frac{(n+1)(n-1)}{8n} \right\}^2$$

We obtain:

$$m_{\tilde{\eta}}^2 = \frac{(n^2-1)(n^2+3)}{192n}$$

The leading term is implied by

$$m_{\tilde{\eta}}^2 = \frac{n^3}{192}$$

Summarizing:

$$m_{\tilde{\xi}}^2 = \frac{n-1}{4} m_s^2, \qquad m_{\tilde{\xi}} \doteq \frac{\sqrt{n}}{2} m_s$$

$$m_{\tilde{\eta}}^2 = \frac{(n^2-1)(n^2+3)}{192n}, \qquad m_{\tilde{\eta}} \doteq \frac{n}{8} \frac{\sqrt{n}}{\sqrt{3}}$$

These results confirm our intuition: in a stretched traverse, the transversal accuracy is far inferior to the longitudinal accuracy.

5.7. Rigorously adjusted regular traverse.

It will turn out that, in case of a regular traverse, conventional adjustment is equivalent to rigorous adjustment.

Rigorous adjustment is most easily done by conditions. The three linearized conditions are

$$\sum_{i=1}^{n-1} \Delta s_{i,i+1} = 0 \qquad (\text{from } \Delta x_n = 0)$$
$$\sum_{i=1}^{n} \Delta\alpha_i = 0 \qquad (\text{from } \sum_{i=1}^{n} \alpha_i = n\pi)$$
$$\sum_{i=1}^{n-1} \sum_{j=1}^{i} \Delta\alpha_j = 0 \qquad (\text{from } \Delta y_n = 0)$$

The adjustment problem decomposes into one for distances and one for angles. We deal only with the problem for angles which is the more difficult one. The conditions are rewritten as

$$\Delta\alpha_1 + \Delta\alpha_2 + \ldots + \Delta\alpha_n = 0$$
$$(n-1)\Delta\alpha_1 + (n-2)\Delta\alpha_2 + \ldots + 0\cdot\Delta\alpha_n = 0$$

We orthogonalize the conditions by subtracting a proper multiple of the first one from the second, obtaining:

$$\Delta\alpha_1 + \Delta\alpha_2 + \ldots + \Delta\alpha_n = 0$$
$$\frac{n-1}{2}\Delta\alpha_1 + \frac{n-3}{2}\Delta\alpha_2 + \ldots - \frac{n-1}{2}\Delta\alpha_n = 0$$

The normal equations are

$$\begin{bmatrix} n & 0 \\ 0 & \sum_{i=-(n-1)/2}^{+(n-1)/2} i^2 \end{bmatrix} \cdot \begin{bmatrix} k_1 \\ k_2 \end{bmatrix} = \begin{bmatrix} w_1 \\ w_2 \end{bmatrix}$$

or

$$\begin{bmatrix} n & 0 \\ 0 & \frac{n(n^2-1)}{12} \end{bmatrix} \cdot \begin{bmatrix} k_1 \\ k_2 \end{bmatrix} = \begin{bmatrix} w_1 \\ w_2 \end{bmatrix}$$

The covariance matrix of the adjusted angles is

$$\Sigma(\Delta\tilde{\alpha}) = I - \begin{bmatrix} 1 & \frac{n-1}{2} \\ 1 & \frac{n-3}{2} \\ \vdots & \vdots \\ 1 & -\frac{n-1}{2} \end{bmatrix} \cdot \begin{bmatrix} \frac{1}{n} & 0 \\ 0 & \frac{12}{n(n^2-1)} \end{bmatrix} \cdot \begin{bmatrix} 1 & 1 & \dots & 1 \\ \frac{n-1}{2} & \frac{n-3}{2} & \dots & -\frac{n-1}{2} \end{bmatrix}$$

The best estimate for $\Delta\eta$ is

$$\Delta\tilde{\eta} = \frac{n-1}{2}\,\Delta\tilde{\alpha}_1 + \frac{n-3}{2}\,\Delta\tilde{\alpha}_2 + \dots + \Delta\tilde{\alpha}_{(n-1)/2}$$

In view of the condition

$$\frac{n-1}{2}\,\Delta\tilde{\alpha}_1 + \frac{n-3}{2}\,\Delta\tilde{\alpha}_2 + \dots - \frac{n-1}{2}\,\Delta\tilde{\alpha}_n = 0$$

we may rewrite the best estimate $\Delta\tilde{\eta}$ as:

$$\Delta\tilde{\eta} = \frac{n-1}{4}\,\Delta\tilde{\alpha}_1 + \frac{n-3}{4}\,\Delta\tilde{\alpha}_2 + \dots \quad 0\,\Delta\tilde{\alpha}_{(n+1)/2} + \dots + \dots + \frac{n-3}{4}\,\Delta\tilde{\alpha}_{n-1} + \frac{n-1}{4}\,\Delta\tilde{\alpha}_n$$

The mean square error of $\Delta\tilde{\eta}$ is obtained as

$$m_{\tilde{\eta}}^2 = \frac{1}{4}\, 2 \sum_{i=1}^{(n-1)/2} i^2 - \frac{1}{4} \left\{ 2 \sum_{i=1}^{(n-1)/2} i \right\}^2 . \frac{1}{n} =$$

$$= \frac{1}{2}\, \frac{n-1}{2}\, \frac{n+1}{2}\, \frac{n}{6} - \frac{1}{4}\, 4 \left\{ \frac{n-1}{2}\, \frac{n+1}{4} \right\}^2 \frac{1}{n} =$$

$$= \frac{n(n^2-1)}{48} - \frac{(n^2-1)^2}{64n} = \frac{(n^2-1)(n^2+3)}{192n} \doteq \frac{n^3}{192}.$$

This is exactly the same result as it was obtained in the previous section.

5.8. Systematic errors in a regular traverse.

We shall answer the following question. Suppose that the actual error of any of the angles α_i is bounded by a small quantity ε. What is the maximal error in $\tilde{\eta}$ that can arise under this assumption ?

From the previous sections we know the best estimator for $\Delta\eta$ as

$$\Delta\tilde{\eta} = \sum_{i=1}^{(n-1)/2} \left\{ \frac{(n+1)^2}{8n} - \frac{i}{2} \right\} \Delta\alpha_i + \sum_{i=(n+1)/2}^{n} \left\{ \frac{(n+1)^2}{8n} - \frac{n-i+1}{2} \right\} \Delta\alpha_i$$

The maximal error obviously is

$$\Delta\tilde{\eta}_{MAX} = \left\{ \sum_{i=1}^{(n-1)/2} \left| \frac{(n+1)^2}{8n} - \frac{i}{2} \right| + \sum_{i=(n+1)/2}^{n} \left| \frac{(n+1)^2}{8n} - \frac{n-i+1}{2} \right| \right\} \varepsilon$$

For large n, this sum is approximated by

$$\Delta\tilde{\eta}_{MAX} \doteq \left\{ \sum_{i=1}^{(n-1)/2} \left| \frac{n}{8} - \frac{i}{2} \right| + \sum_{i=(n+1)/2}^{n} \left| \frac{n}{8} - \frac{n-i+1}{2} \right| \right\} \varepsilon$$

$$\doteq 4 \sum_{i=1}^{n/4} \left(\frac{n}{8} - \frac{i}{2} \right) \varepsilon = 2 \sum_{i=1}^{n/4} \left(\frac{n}{4} - i \right) \varepsilon \doteq \frac{n^2}{16} \varepsilon$$

Thus

$$\Delta\tilde{\eta}_{MAX} \doteq \frac{n^2}{16} \varepsilon$$

Comparing this with the r.m.s. error derived earlier

$$m_{\tilde{\eta}} \doteq \frac{n \sqrt{n}}{8 \sqrt{3}}$$

we arrive at the following conclusion:

For n tending to infinity, the ratio

$$\Delta\tilde{\eta}_{MAX} / m_{\tilde{\eta}} \doteq \frac{\sqrt{3}}{2} \sqrt{n} \varepsilon$$

tends to infinity. In a large traverse, the maximal effect due to systematic errors eventually outgrows the effect of random errors. This statement, at least in a quantitative way, carries over to other large adjustment schemes, e.g. those of large continental networks.

C. CONFIDENCE REGIONS AND TESTS OF LINEAR HYPOTHESES.

1. Probability distributions used in statistical tests.

1.1. One dimensional Gauss distribution (normal distribution).

Let X be a one- dimensional random variable having the probability density

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]$$

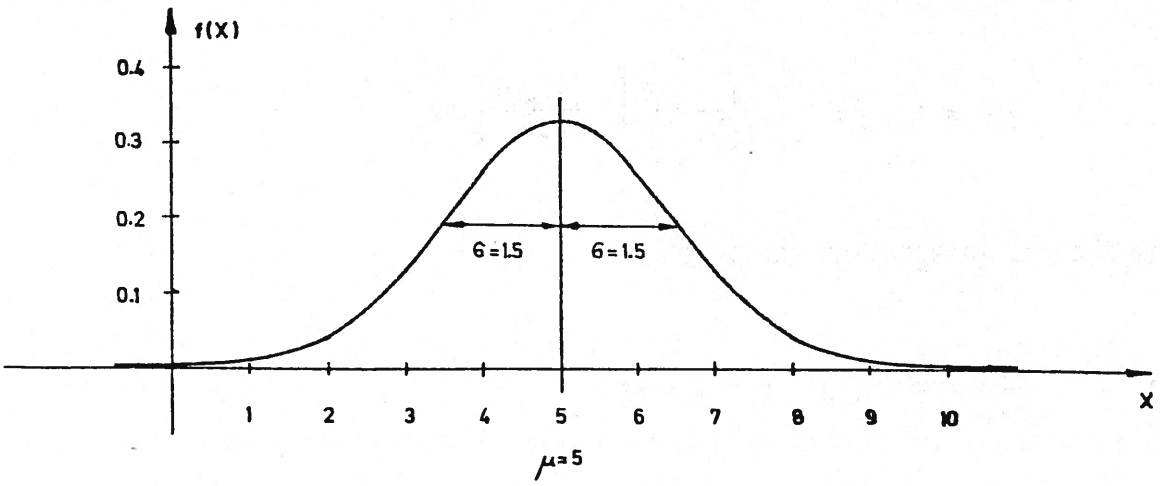
One can prove that

$$E(X) = \int_{-\infty}^{+\infty} x\, f(x)\, dx = \mu$$

$$\sigma^2(X) = \int_{-\infty}^{+\infty} (x-\mu)^2\, f(x)\, dx = \sigma^2$$

Hence μ is the expectation or mean value of X, σ^2 is its variance. The square root of σ^2 is denoted σ. It is called standard deviation.

The graph of f(x) is bell- shaped:



Remark: It is frequently assumed that observation errors are normally distributed. There is no completely rigorous justification for this assumption. However strong support comes from the central limit theorem of probability theory. It can be shown that the sum of a large number of small random variables has a tendency to be normally distributed. This holds under fairly general assumptions. The small elementary random effects may be arbitrarily distributed.

Important special case: The normalized Gauss distribution. It relies on the special choice

$$\mu = 0, \quad \sigma = 1$$

consequently

$$f(x) = \frac{1}{\sqrt{2\pi}} \exp\left| -\frac{x^2}{2} \right|$$

All tabulations of the Gauss distribution refer to the normalized case. That this is completely sufficient is demonstrated by the following example. Suppose that the Gauss distribution under consideration is not normalized, and suppose that one wishes to calculate the probability of $\alpha \leq X \leq \beta$

$$p\{\alpha \leq X \leq \beta\} = \int_{\alpha}^{\beta} \frac{1}{\sqrt{2\pi}\sigma} \exp\left| -\frac{(x-\mu)^2}{2\sigma^2} \right| dx$$

One changes integration variables

$$\frac{x-\mu}{\sigma} = \xi, \quad \text{i.e.} \quad x = \mu + \sigma\xi$$

One obtains

$$p\{\alpha \leq x \leq \beta\} = \int_{(\alpha-\mu)/\sigma}^{(\beta-\mu)/\sigma} \frac{1}{\sqrt{2\pi}} \exp\left|-\frac{\xi^2}{2}\right| d\xi$$

It is seen that one calculates the probability

$$p\left\{\frac{\alpha-\mu}{\sigma} \leq \Xi \leq \frac{\beta-\mu}{\sigma}\right\}$$

thereby Ξ has the normalized Gauss distribution. In tabulation one usually finds values of the distribution function

$$F(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} \exp\left|-\frac{\xi^2}{2}\right| d\xi$$

Hence

$$p\{\alpha \leq X \leq \beta\} = F\left|\frac{\beta-\mu}{\sigma}\right| - F\left|\frac{\alpha-\mu}{\sigma}\right|$$

Remark: The probability of the event

$$-k\sigma \leq X-\mu \leq +k\sigma$$

depends on k only. It equals the probability of

$$-k \leq \Xi \leq +k$$

where Ξ has the normalized Gauss distribution.

Usually one focuses interest on the complementary event, i.e.

$$|X-\mu| > k\sigma$$

The probability of $X-\mu$ exceeding the $k\sigma$- limits is small for moderately large k. For k=3 one obtains

$$p\{|X-\mu| > 3\sigma\} = 0.0027$$

This is about 3 parts per thousand.

Example: Although we are currently collecting the theoretical ingredients needed for statistical tests to be described in detail later on, we briefly pause, taking a look at a very simple example of a statistical test.

Suppose that a carefully maintained base line is known to have a length of 151.723 m. This value has been verified so many times that we regard it as free of any error. Suppose that a newly delivered distance meter gives a reading of 151.745. The manufacturers specified a standard deviation (root mean square error) of 5 mm, if the distance is in the range of 150 m. We make the following hypothesis: (1) The distance meter is free of any systematic error. (2) No severe blunder occurred during the measurement. (3) The r.m.s. error of 5 mm specified by the company is correct. (4) The observations are normally distributed.

We test the hypothesis as follows. If it is true, our observation is normally distributed with mean 151.723 and standard deviation 5 mm. The observed value 151.745 is outside the 3σ- boundaries (which are 151.723 ± 0.015 = 151.708 and 151.738 respectively). Hence we reject the hypothesis.

Statistical tests to be described later are modelled after this simple case. However the distribution functions involved are more complicated, and we have to learn more about them. Nevertheless, some preliminary questions are posed here:

*) Why do we choose the critical area of rejection as $|X-\mu| > 3\sigma$, and not otherwise, for example as $X-\mu > 3\sigma$, or even as $|X-\mu| >$ some constant?

*) A not rejected hypothesis is not always considered as accepted without reservations. Additional information may lead to rejection at a later time.

*) The probability of rejecting a true hypothesis is in our case given by 0.0027. A more difficult question is: How large is the probability to accept a wrong hypothesis. Obviously the probability depends on how wrong the hypothesis is.

1.2. The multidimensional Gauss distribution (normal distribution).

Let the random variable X be n- dimensional having a density function:

$$f(x) = f\left|x_1,\ldots,x_n\right| =$$

$$= \frac{1}{(2\pi)^{n/2}} |A|^{1/2} \exp\left[- \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij} \left(x_i - \mu_i\right) \left(x_j - \mu_j\right) \right]$$

$$= \frac{1}{(2\pi)^{n/2}} |A|^{1/2} \exp\left[- \frac{1}{2} (x-\mu)^T A (x-\mu) \right]$$

Thereby the matrix $A = (a_{ij})$ is symmetric and positive definite. The symbol $|A|$ denotes the determinant of A, which is positive. The vector

$$\mu = \left(\mu_1, \ldots, \mu_n\right)^T$$

can be verified to be the vector of mean values

$$E(X) = \mu$$

The inverse

$$\Sigma = A^{-1}$$

can be verified to be the covariance matrix of X

$$\Sigma(X) = \Sigma = A^{-1}$$

Remark: The case n=1 reduces to the one- dimensional Gauss distribution described in the previous section. Just identify $\mu=\mu_1$, $\Sigma=a_{11}{}^{-1}=\sigma^2$.

Theorems on the multidimensional Gauss distribution: (without proofs)

(1) Marginal distribution. Each component X_i of X can be viewed as a one-dimensional random variable (cf. section B.2.5.). As such, X_i has a one-dimensional Gauss distribution with mean $E(X_i) = \mu_i$ and variance $\sigma^2(X_i) = \sigma_{ii}$, the i-th diagonal element of Σ.

More generally: Partition X as

$$X = \begin{bmatrix} X_{(1)} \\ X_{(2)} \end{bmatrix}$$

where $X_{(1)}$ and $X_{(2)}$ are vectors of size n_1, n_2; $n_1+n_2=n$.

Partition μ, accordingly

$$\mu = \begin{bmatrix} \mu_{(1)} \\ \mu_{(2)} \end{bmatrix}, \qquad \Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}$$

Then X_i has an n_i- dimensional Gauss distribution with mean μ_i and covariance matrix Σ_{ii}; i=1,2

$$f(x_{(i)}) = \frac{1}{(2\pi)^{n_i/2}} \exp\left[-\frac{1}{2}\left(x_{(i)}-\mu_{(i)}\right)^T A_{ii}\left(x_{(i)}-\mu_{(i)}\right)\right],$$

$$A_{ii} = \Sigma_{ii}^{-1}, \quad i=1,2$$

(2) Linear functions of normally distributed random variables. Let

$$Y = BX + b$$

be a linear, inhomogeneous function of X, which is assumed normally distributed with $E(X)=\mu$, $\Sigma(X)=\Sigma$. Then Y is also normally distributed with

$$E(Y) = B\mu + b, \quad \Sigma(Y) = B\Sigma(X)B^T = B\Sigma B^T$$

Note that E(Y), Σ(Y) follow from the propagation laws for expectations and covariances given in section B.3.7.

(3) Meaning of zero correlation in case of normally distributed random variables. Assume that $\Sigma=\Sigma(X)$ is of the form

$$\Sigma(X) = \begin{bmatrix} \sigma_{11} & & & 0 \\ & \sigma_{22} & & \\ & & \ddots & \\ 0 & & & \sigma_{nn} \end{bmatrix} = \begin{bmatrix} \sigma_1^2 & & & 0 \\ & \sigma_2^2 & & \\ & & \ddots & \\ 0 & & & \sigma_n^2 \end{bmatrix}$$

Then the components of X are mutually uncorrelated. As a consequence

$$A = \Sigma^{-1} = \begin{bmatrix} a_{11} & & & 0 \\ & a_{22} & & \\ & & \ddots & \\ 0 & & & a_{nn} \end{bmatrix} = \begin{bmatrix} 1/\sigma_1^2 & & & 0 \\ & 1/\sigma_2^2 & & \\ & & \ddots & \\ 0 & & & 1/\sigma_n^2 \end{bmatrix}$$

The probabilty density presents itself as

$$f(x) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left| - \frac{|x_i-\mu_i|^2}{2\sigma_i^2} \right|$$

This is true because

$$|A| = \left[\sigma_1^2 \, \sigma_2^2 \, \ldots \, \sigma_n^2\right]^{-1}$$

and

$$\exp\left| - \tfrac{1}{2} (x-\mu)^T A^{-1} (x-\mu) \right| = \exp\left| - \tfrac{1}{2} \sum_{i=1}^{n} a_{ii} \, |x_i-\mu_i|^2 \right|$$

$$= \prod_{i=1}^{n} \exp\left| - \tfrac{1}{2} \frac{|x_i-\mu_i|^2}{\sigma_i^2} \right|$$

It is seen that f(x) decomposes as

$$f(x) = f_{(1)}|x_1| \; f_{(2)}|x_2| \; \ldots \; f_{(n)}|x_n|$$

where $f_{(i)}|x_i|$ is the marginal density of the component X_i. In view of section B.2.6. we recognize that the components X_i of X are mutually stochastically <u>independent</u>.

Thus: In case of normally distributed random variables zero correlation means stochastic independence.

This generalizes to the following situation. Split X, μ again as

$$X = \begin{bmatrix} X_{(1)} \\ X_{(2)} \end{bmatrix}, \qquad \mu = \begin{bmatrix} \mu_{(1)} \\ \mu_{(2)} \end{bmatrix}$$

Assume that Σ is of the form

$$\Sigma = \begin{bmatrix} \Sigma_{11} & 0 \\ 0 & \Sigma_{22} \end{bmatrix}$$

Then the random variables $X_{(1)}$, $X_{(2)}$, whose marginal densities $f_{(1)}(x_{(1)})$, $f_{(2)}(x_{(2)})$ have been specified under (1), are independent. It holds that

$$f(x) = f_{(1)}(x_{(1)})\, f_{(2)}(x_{(2)})$$

Remark: An important special case arises if $\mu_i=0$, $\sigma_{ii}=\sigma^2$, $\sigma_{ij}=0$, $i\neq j$. The components X_i of X are then identically and independently distributed. We have $\mu=0$, $\Sigma=I\sigma^2$. Any component has the Gauss distribution with mean zero and variance σ^2. If observations of the same kind are taken under identical circumstances, then the observation errors are frequently assumed to be distributed in this way.

1.3. The chi- squared distribution (χ^2- distribution).

Let $X_1,\ldots,X_n$ be mutually independent and normally distributed random variables, having mean $E(X_i)=0$ and $\sigma^2(X_i)=1$. Thus any X_i has the normalized Gauss

distribution, and the random vector $X=(X_1,\ldots,X_n)^T$ has mean and covariance matrix given by

$$E(X) = 0 \ , \qquad \Sigma(X) = I$$

We consider the function of X

$$\chi^2_n = \Phi(X) = X_1^2 + X_2^2 + \ldots + X_n^2$$

This is a nonlinear function which may also be written as

$$\chi^2_n = X^T X$$

The random variable χ^2_n has a distribution which is called chi- squared (χ^2) distribution with n degrees of freedom. An analytical expression may be specified, but is not needed here. It holds that

$$E(\chi^2_n) = n$$
$$\sigma^2(\chi^2_n) = 2n$$

The proof of $E(\chi^2_n)=n$ is easy: Since any X_i has the normalized Gauss distribution, we have $E(X_i)=0$ and

$$E(X_i^2) = E\{(X_i - E(X_i))^2\} = \sigma^2(X_i) = 1$$

Hence the random variable X_i^2 has expectation $E\{X_i^2\}=1$. The random variable χ^2_n is the sum of n such random variables. Hence its expectation must be n- times as large (Remember: the expectation operator is linear; cf. section B.3.7.).

The distribution functions $F_{\chi^2_n}(x)$ are tabulated for moderately large values of n ($n<200$, for example). For large values of n, $F_{\chi^2_n}(x)$ approximates a Gauss distribution with mean n and variance 2n.

Note that $\chi^2_n \geq 0$. Hence $F_{\chi^2_n}(x)=0$ for $x<0$.

1.4. Student's distribution (t- distribution).

Let X_i, $i=1,\ldots,n$ again be random variables being mutually independent, and obeying the normalized Gauss distribution. Let Y be another random variable having the normalized Gauss distribution, and let Y be independent of any X_i. Thus the partitioned random vector

$$\begin{bmatrix} X \\ Y \end{bmatrix} = \begin{bmatrix} X_1 \\ \cdot \\ \cdot \\ X_n \\ Y \end{bmatrix}$$

has expectation and covariance given by

$$E\begin{bmatrix} X \\ Y \end{bmatrix} = 0, \qquad \Sigma\begin{bmatrix} X \\ Y \end{bmatrix} = I$$

We consider the function

$$t_n = \frac{Y}{\sqrt{(X_1^2+X_2^2+\ldots+X_n^2)/n}}$$

The distribution of this nonlinear function t_n of X, Y is called Student's distribution, or t- distribution with n degrees of freedom.

Remark: Note that the denominator can be written as

$$\sqrt{(X_1^2+\ldots+X_n^2)/n} = \sqrt{\chi^2_n/n}$$

where χ^2_n is a random variable having the χ^2- distribution introduced in the previous subsection.

Student's distribution has a density whose analytical representation is not given here. It is symmetric with respect to the origin $t_n=0$. The distribution function is tabulated for moderately large n (again, $n<200$ is reasonable). For large n it approximates the normalized Gauss distribution.

1.5. Fisher's distribution (F- distribution).

Let $X_1,\ldots,X_m$, $Y_1,\ldots,Y_n$ be mutually independent random variables having the normalized Gauss distribution. The quantity

$$F_{m,n} = \frac{(X_1^2+\ldots+X_m^2)/m}{(Y_1^2+\ldots+Y_n^2)/n}$$

is a random variable whose distribution is called Fisher's F- distribution with m and n degrees of freedom.

Remark: Note that $F_{m,n}$ is represented as

$$F_{m,n} = \frac{\chi^2_m/m}{\chi^2_n/n}$$

i.e. as the quotient of two normalized and independent χ^2- distributed random variables having m and n degrees of freedom.

The F- distribution is tabulated for moderately large n and m (n,m<200, say). If n becomes very large, the denominator can be considered to be a constant with value 1. The numerator is then χ^2_m divided by m. If m becomes very large, the argument can be applied to $1/F_{m,n}=F_{n,m}$.

2. Canonical transformation.

2.1. Preliminaries.

We consider an adjustment problem in the Gauss- Markoff form

$$E(\ell) = Ax, \qquad \Sigma(\ell) = Q\sigma^2$$

Alternatively, we also consider the conventional form

$$\ell+v = Ax, \qquad \Sigma(\ell) = Q\sigma^2$$

which shows the corrections explicitely. We assume ℓ of dimension n, x of dimension m, such that A is an n*m matrix, whereby m<n and rank(A)=m.

We subject the problem to a series of transformations such that from the final appearance not only the solution can be read off immediately, but also various statistical quantities needed in tests to be described later. Our transformation will be a more sophisticated version of the transformation described in section A.8.5. It will only serve the purpose of mathematical proofs and theoretical insight. In practical application such transformations are never carried out explicitely.

Our transformations will not only involve the quantities ℓ, v, Ax, Q, σ^2 showing up in the above problem formulation. We also consider a set of p linear functionals

$$\Phi = \Phi x$$

defined on the subspace L_A which is spanned by the columns of A. The rows of the p*m matrix Φ represent linear functions of the unknown parameters x.

Statistical tests to be described later will be concerned with hypotheses like

$$\Phi x = c$$

where c is a vector of p constants.

2.2. Making the functionals a part of the parameters.

We augment the p*m matrix Φ by an (m-p)*m matrix Ψ such that

$$C = \begin{bmatrix} \Phi \\ \Psi \end{bmatrix}$$

becomes an m*m regular matrix. We introduce new parameters

$$y = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix}$$

by

$$y = Cx, \quad \text{or} \quad y_1 = \Phi x, \quad y_2 = \Psi x$$

The inverse transformation is

$$x = C^{-1}y$$

and our adjustment problem transforms as

$$\ell+v = AC^{-1}y$$

or

$$\ell+v = \bar{A}y, \quad \bar{A} = AC^{-1}$$

or

$$\ell+v = \bar{A}_1 y_1 + \bar{A}_2 y_2$$

The first set of parameters refers now directly to the p functionals: $y_1=\Phi x$.

2.3. Orthogonal decomposition of the space L_A.

As usual, we view the realizations of ℓ, as well as $\mathbb{L}=E(\ell)$, the columns of A, $\bar{A}$, as members of an inner product space L. The inner product is represented by

$$P = Q^{-1}$$

Note that A and $\bar{A}$ span the subspace L_A (the columns of $\bar{A}$ are linear combinations of those of A). In an analogous way as described in section A.9.3. on partial reduction, we decompose the space $L_A=L_{\bar{A}}$ into ortho- complementary subspaces $L_{\bar{A}_1}$ and $L_{\bar{A}_2}$. The only change with respect to section A.9.3. is an additional overbar over the A_i's, and an interchange of the subscripts 1 and 2. As we know from the

cited section, the orthogonal decomposition goes along with a parameter transformation

$$y_1 = y_1$$
$$z_2 = y_2 - \left(\bar{A}_2^T P \bar{A}_2\right)^{-1} \bar{A}_2^T P \bar{A}_1 \, y_1$$

The result is the transformed problem

$$\ell + v = \bar{\bar{A}}_1 y_1 + \bar{A}_2 z_2 ,$$

with

$$\bar{\bar{A}}_1 = \bar{A}_1 - \bar{A}_2 \left(\bar{A}_2^T P \bar{A}_2\right)^{-1} \bar{A}_2^T P \bar{A}_1 = \left(I - P_{\bar{A}_2}\right) \bar{A}_1$$

and

$$\bar{\bar{A}}_1^T P \bar{A}_2 = 0$$

2.4. Orthogonal decomposition of L into L_A and L_B.

Such a transformation has been used in several earlier sections. It is accomplished by

$$\ell = \left(\bar{\bar{A}}_1, \bar{A}_2, B\right) \ell'$$

or

$$\ell' = \begin{bmatrix} \ell_1' \\ \ell_2' \\ \ell_3' \end{bmatrix} = \begin{bmatrix} \left(\bar{\bar{A}}_1^T P \bar{\bar{A}}_1\right)^{-1} \bar{\bar{A}}_1^T P \\ \left(\bar{A}_2^T P \bar{A}_2\right)^{-1} \bar{A}_2^T P \\ \left(B^T P B\right)^{-1} B^T P \end{bmatrix} \ell$$

Shortly

$$\ell' = S\ell$$

The matrix B fulfills $A^T PB=0$, and also $\bar{\bar{A}}_1^T PB=0$, $\bar{A}_2^T PB=0$. Our adjustment problem transforms into

$$\begin{aligned} \ell_1'+v_1' &= y_1 \\ \ell_2'+v_2' &= z_2 \\ \ell_3'+v_3' &= 0 \end{aligned} \quad , \quad \Sigma(\ell') = \begin{bmatrix} (\bar{\bar{A}}_1^T P \bar{\bar{A}}_1)^{-1} & 0 & 0 \\ 0 & (\bar{A}_2^T P \bar{A}_2)^{-1} & 0 \\ 0 & 0 & (B^T PB)^{-1} \end{bmatrix} \sigma^2$$

Already at this step we see that the best estimates and corrections are

$$\begin{aligned} y_1 &= \ell_1', & v_1' &= 0 \\ z_2 &= \ell_2', & v_2' &= 0 \\ & & v_3' &= -\ell_3' \end{aligned}$$

Remark: Recall the geometric interpretation of this transformation. A new basis is chosen in L. The new basis is the union of bases in $L_{\bar{\bar{A}}_1}$, $L_{\bar{A}_2}$, L_B. These 3 subspaces are orthogonal. The inner product in L was represented by P with respect to the old bases. With respect to the new bases it is represented by

$$P' = \begin{bmatrix} \bar{\bar{A}}_1^T P \bar{\bar{A}}_1 & 0 & 0 \\ 0 & \bar{A}_2^T P \bar{A}_2 & 0 \\ 0 & 0 & B^T PB \end{bmatrix}$$

The reproducing kernel is represented by

$$Q' = (P')^{-1}$$

Consequently

$$\Sigma(\ell') = Q'\sigma^2$$

Note also the validity of the error propagation law:

$$\ell' = S\ell, \quad \Sigma(\ell') = S\Sigma(\ell)S^T, \quad \Sigma(\ell) = Q\sigma^2$$

2.5. Orthonormalizing the bases of the subspaces.

Let V be a vector space; let $e_1,\ldots,e_n$ be a basis, and let the positive definite matrix G represent the inner product. As outlined in section A.4.7., a set of orthonormal vectors $e_1',\ldots,e_n'$ may be derived. If these vectors are chosen as new basis vectors, the coordinates of vectors transform as

$$x = R^{-1}x'$$

(The earlier notation used in section A.3.2., paragraph (6) was $x_{OLD}=Ax_{NEW}$.). The inner product with respect to the new basis is represented by the identity matrix

$$G' = \left(R^{-1}\right)^T GR^{-1} = I$$

We see that G is represented as

$$G = R^T R$$

Remark: There is a very old, very practical and well- known procedure to compute the matrix G. It is the method by Cholesky. R is the Cholesky- factor of G. R is an upper triangular matrix (cf. section D.3.).

The inner products in our 3 subspaces $L_{\bar{\bar{A}}_1}$, $L_{\bar{A}_2}$, L_B are represented by

$$G_1 = \bar{\bar{A}}_1^T P \bar{\bar{A}}_1$$
$$G_2 = \bar{A}_2^T P \bar{A}_2$$
$$G_3 = B^T P B$$

We factorize these matrices as

$$G_1 = R_1^T R_1$$
$$G_2 = R_2^T R_2$$
$$G_3 = R_3^T R_3$$

We accordingly transform the observations by

$$\ell_1'' = R_1 \ell_1'$$
$$\ell_2'' = R_2 \ell_2'$$
$$\ell_3'' = R_3 \ell_3'$$

Shortly

$$\ell'' = S'\ell'$$

Our adjustment problem transforms into

$$\begin{aligned} \ell_1''+v_1'' &= R_1 y_1 \\ \ell_2''+v_2'' &= R_2 z_2 \\ \ell_3''+v_3'' &= 0 \end{aligned}, \qquad \Sigma\begin{vmatrix}\ell_1''\\ \ell_2''\\ \ell_3''\end{vmatrix} = \begin{vmatrix} I & 0 & 0\\ 0 & I & 0\\ 0 & 0 & I\end{vmatrix}\sigma^2$$

Remark: The described transformations imply an alternative form of the final stage given by

$$\begin{aligned} E\{\ell_1''\} &= R_1 y_1 \\ E\{\ell_2''\} &= R_2 z_2 \\ E\{\ell_3''\} &= 0 \end{aligned}$$

with $\Sigma(\ell'')$ as given above.

The solution is

$$\begin{aligned} y_1 &= R_1^{-1}\ell_1'', & v_1'' &= 0 \\ z_2 &= R_2^{-1}\ell_2'', & v_2'' &= 0 \\ & & v_3'' &= -\ell_3'' \end{aligned}$$

The covariances of the various quantities are

$$\Sigma(y_1) = (R_1^{-1})(R_1^{-1})^T \sigma^2 = (R_1^T R_1)^{-1} \sigma^2 = G_1^{-1} \sigma^2$$
$$\Sigma(z_2) = (R_2^T R_2)^{-1} \sigma^2$$
$$\Sigma(v_3'') = I \sigma^2$$

Because a change of basis in L amounts to an isometric transformation (cf. section A.8.4.2.), we have

$$\|v\|^2 = v^T P v = \|v_3''\|^2 = \|\ell_3''\|^2 = (\ell_3'')^T \ell_3''$$

Because $y_1 = \Phi x$, we see that y_1 is the BLUE for the functionals $\Phi = \Phi x$. On the other hand, the BLUE for Φx can also be obtained conventionally by adjusting for the x's directly

$$x = (A^T P A)^{-1} A^T P \ell$$

with

$$\Sigma(x) = (A^T P A)^{-1} \sigma^2$$

and

$$y_1 = \Phi x$$

with

$$\Sigma(y_1) = \Sigma(\Phi x) = \Phi (A^T P A)^{-1} \Phi^T \sigma^2$$

Thus we see

$$G_1^{-1} = \left(R_1^T R_1\right)^{-1} = \phi\left(A^T P A\right)^{-1}\phi^T$$

This formula will help us to utilize the insight gained through canonical transformation without actually performing this transformations in practical calculations.

3. Distribution of various quantities resulting from a least squares adjustment.

3.1. The joint distribution of BLUE'S and of the residuals.

We now make the decisive assumption that the observations ℓ are normally distributed. Thus the vector ℓ is assumed to have an n- dimensional Gauss distribution with mean $E(\ell)=\lambda=Ax$ and covariance matrix $\Sigma(\ell)=Q\sigma^2$.

We see that the mean is specified in terms of m unknown parameters x. The covariance may also have an unknown parameter, namely σ^2. However, σ^2 may also be assumed to be known.

The best estimates of functionals $\phi=\phi x$ are given by

$$\tilde{\phi} = \phi\tilde{x} = \phi\left(A^TPA\right)^{-1}A^TP\ell$$

The residuals are given by

$$v = -\left(I - A\left(A^TPA\right)^{-1}A^TP\right)\ell$$

$\tilde{\phi}$ as well as v are linear functions of the observation ℓ. Hence they are also normally distributed (cf. section 1.2., paragraph (2)). In order to specify their multidimensional normal distribution, it suffices to specify the vector of expectations and the covariance matrix. Because the $\tilde{\phi}$ are BLUE's we have

$$E(\tilde{\phi}) = \phi = \phi x$$

This may also be verified directly. On the other hand

$$E(v) = 0$$

as one recognizes by applying the E- operator to the above expressions for v and by recalling $E(\ell)=Ax$.

A straightforward application of the error propagation law (for covariances) yields:

$$\Sigma\begin{bmatrix}\tilde{\Phi}\\ v\end{bmatrix} = \begin{bmatrix}\Phi\left(A^TPA\right)^{-1}\Phi^T & 0\\ 0 & Q-A\left(A^TPA\right)^{-1}A^T\end{bmatrix}\sigma^2$$

It is seen that the BLUE's $\tilde{\Phi}=\Phi\tilde{x}$ and the residuals v are uncorrelated and, since we deal with normal distributions, stochastically <u>independent.</u>

<u>3.2. Distribution of the "weighted sum of residuals".</u>

The quantity

$$v^TPv$$

is the quantity which is minimized during least squares adjustment. Its distribution will now be specified.

<u>Theorem:</u> The quantity

$$\frac{1}{\sigma^2} v^T P v = \frac{1}{\sigma^2} \sum_{i=1}^{n} \sum_{j=1}^{n} p_{ij} v_i v_j$$

has a χ^2_{n-m} distribution, i.e. a χ^2- distribution with n-m degrees of freedom.

Remark: n is the number of observations, m is the number of unknown parameters; hence

$$n-m$$

is the number of redundant observations.

Proof: We refer to section 2.5. There the equation

$$\left(v^T P v\right) = \left(l_3''\right)^T l_3''$$

was listed. The vector l_3'' has n-m components. The expectation of l_3'' is zero (this is stated in remark within section 2.5.).

The covariance of l_3'' is $I\sigma^2$. Hence the covariance of $(1/\sigma)l_3''$ is I. It follows that

$$\frac{1}{\sigma^2} \left(l_3''\right)^T l_3'' = \frac{1}{\sigma^2} v^T P v$$

is a sum of squares of n-m independent random variables having a normalized Gauss distribution. Thus the sum has the χ^2_{n-m} distribution. Confer the definition of χ^2 in section 1.3.

3.3. Expressions in $\phi\tilde{x}$ and v having the χ^2- or the F- distribution.

The following theorem is fundamental for statistical tests of linear hypotheses.

Theorem: Let $\phi\tilde{x}$ be the BLUE for ϕx, where ϕ is a p*m matrix. Denote

$$c = \phi x$$

Then

$$\chi^2_p = \frac{1}{\sigma^2} (\phi\tilde{x}-c)^T \left(\phi\left(A^TPA\right)^{-1}\phi^T\right)^{-1} (\phi\tilde{x}-c)$$

has a χ^2- distribution with p degrees of freedom. Furthermore

$$F_{p,n-m} = \frac{(\phi\tilde{x}-c)^T \left(\phi\left(A^TPA\right)^{-1}\phi^T\right)^{-1} (\phi\tilde{x}-c)/p}{\left(v^TPv\right)/(n-m)}$$

has an F- distribution with p and n-m degrees of freedom.

Remark: The quantity $c=\phi x$ is unknown because the x are unknown. Hence χ^2_p and $F_{p,n-m}$ involve the unknown quantity c. However the distribution of χ^2_p and $F_{p,n-m}$ is known. In later sections hypothetical values for c will be specified, and these hypotheses will be tested using the specified expressions and their distributions.

Proof: In section 2.5. we have seen that

$$\ell_1'' = R_1\tilde{y}_1 = R_1\phi\tilde{x}$$

$$E(\ell_1'') = R_1 y_1 = R_1\phi x = R_1 c$$

$$\Sigma(\ell_1'') = I\sigma^2$$

Hence

$$\frac{1}{\sigma^2}\left[\ell_1''-E(\ell_1'')\right]^T\left[\ell_1''-E(\ell_1'')\right]$$

is a sum of squares of p independent random variables (the components of $(1/\sigma)\left[\ell_1''-E(\ell_1'')\right]$), having the normalized Gauss distribution. By the definition of the χ^2- distribution, the above expression is a χ^2_p. Substituting for ℓ_1'' and $E(\ell_1'')$ we obtain

$$\chi^2_p = \frac{1}{\sigma^2}\left(R_1\phi\tilde{x}-R_1 c\right)^T\left(R_1\phi\tilde{x}-R_1 c\right) =$$

$$= \frac{1}{\sigma^2}(\phi\tilde{x}-c)^T R_1{}^T R_1\ (\phi\tilde{x}-c)$$

In section 2.5. it was noted that

$$R_1{}^T R_1 = G_1 = \left(\phi\left(A^T P A\right)^{-1}\phi^T\right)^{-1}$$

This proves the assertion on χ^2_p.

The assertion on $F_{p,n-m}$ is proved similarly by noting that $(1/\sigma)\left[\ell_1''-E(\ell_1'')\right]$ and $(1/\sigma)\ell_3''$ represent p + n-m random variables having the normalized Gauss distribution.

Hence

$$F_{p,n-m} = \frac{\left[\ell_1"-E\left(\ell_1"\right)\right]^T \left[\ell_1"-E\left(\ell_1"\right)\right]/p}{\left(\ell_3"^T\ell_3"\right)/(n-m)}$$

has the F- distribution with p and n-m degrees of freedom. Recalling that $\left(\ell_3"\right)^T\ell_3"=v^TPv$ makes the proof complete.

Theorem: (Alternative representation of the quantities χ^2_p and $F_{p,n-m}$ specified in the previous theorem.) Let v be the residuals of the adjustment problem

$$\ell+v = Ax, \qquad \Sigma(\ell) = Q\sigma^2$$

Let v_c be the residuals of this adjustment problem augmented by p additional constraints:

$$\ell+v_c = Ax, \qquad \Sigma(\ell) = Q\sigma^2$$
$$c = \Phi x$$

Here c is viewed as a vector of constants (the constraints may e.g. be used to reduce the number of parameters). Then

$$(\Phi\tilde{x}-c)^T \left[\Phi\left(A^TPA\right)^{-1}\Phi^T\right]^{-1} (\Phi\tilde{x}-c) = v_c{}^TPv_c - v^TPv$$

Hence

$$\chi^2_p = \frac{1}{\sigma^2}\left(v_c{}^TPv_c-v^TPv\right)$$

and

$$F_{p,n-m} = \frac{\left(v_c^T P v_c - v^T P v\right)/p}{\left(v^T P v\right)/(n-m)}$$

are the same quantities as those listed in the preceding theorem.

Remark: The newly specified expressions for χ^2_p and $F_{p,n-m}$ are frequently easier to calculate by means of available computer programs for least squares adjustment.

Proof: Just note from the canonical transformation that

$$v^T P v = \left(l_3''\right)^T l_3''$$

Next we must show that

$$v_c^T P v_c = \left(l_1''-R_1 c\right)^T \left(l_1''-R_1 c\right) + \left(l_3''\right)^T l_3''$$

This expression is verified by noting that the canonical transformation of the modified problem is

$$\begin{aligned} l_1''+v_1'' &= R_1 c \\ l_2''+v_2'' &= R_2 z_2 , \qquad \Sigma(l'') = I\sigma^2 \\ l_3''+v_3'' &= 0 \end{aligned}$$

The obvious solution of this problem is

$$v_1'' = -\left(\ell_1''-R_1c\right)$$
$$z_2 = R_2^{-1}\ell_2'', \quad v_2'' = 0$$
$$v_3'' = -\ell_3''$$

Due to isometry we have $v_c^T P v_c = \left(v''\right)^T v''$. Thus

$$v_c^T P v_c = \left(\ell_1''-R_1c\right)^T \left(\ell_1''-R_1c\right) + \left(\ell_3''\right)^T \ell_3''$$

and the proof is completed by plugging it into the proof at the preceding theorem.

3.4. Expressions in $\tilde{x}$ and v having the t- distribution.

A random variable $F_{1,n-m}$ having the F- distribution with 1 and n-m degrees of freedom can be seen as the square of a random variable t_{n-m} having the t-distribution with n-m degrees of freedom. Hence the following theorem is very closely related to the second assertion of the first theorem of the preceding section. In the following theorem we assume p=1, and we write ϕx as $\phi^T x$.

Theorem: Let $\phi=\phi^T x$ be a linear function of the unknowns and let $\tilde{\phi}=\phi^T\tilde{x}$ be its BLUE. Put

$$\phi^T x = c$$

Then

$$t_{n-m} = \frac{\left(\phi^T\tilde{x}-c\right)}{\sqrt{\phi^T\left(A^TPA\right)^{-1}\phi\ \left(v^TPv\right)/(n-m)}}$$

has the t- distribution with n-m degrees of freedom.

Proof: Replace in section 2 the quantity Φx by $\Phi^T x$ everywhere. Then ℓ_1'' has only one component. It follows that

$$t_{n-m} = \frac{\ell_1'' - E(\ell_1'')}{\sqrt{(\ell_3'')^T \ell_3''/(n-m)}}$$

is a t_{n-m}. Resubstituting for ℓ_1'' and ℓ_3'' gives the stated expression.

4. Confidence regions.

4.1. Confidence intervals for one- dimensional Gauss variables.

Suppose that X has the one- dimensional Gauss distribution with unknown expectation $\mu=E(X)$ and known standard deviation $\sigma=\sigma(X)$. As described in section 1.1., a transformation is made whose effect is a replacement of X by a random variable

$$\Xi = \frac{X-\mu}{\sigma}$$

having the normalized Gauss distribution. One specifies a certain probabiltiy α which is usually chosen close to 1. Values of $\alpha=0.9$, $\alpha=0.95$, $\alpha=0.99$ are common choices. The probability α is called confidence level. Using a table of the normalized Gauss distribution, one determines k_α such that

$$p\{-k_\alpha \leq \Xi \leq k_\alpha\} = \alpha$$

This is equivalent to

$$p\{-k_\alpha \leq \frac{X-\mu}{\sigma} \leq k_\alpha\} = \alpha$$

or

$$p\{X-k_\alpha\sigma \leq \mu \leq X+k_\alpha\sigma\} = \alpha$$

It is seen that an interval has been specified whose boundaries are random variables. The interval covers the unknown expectation μ with a prescribed probability α. The interval carries the name confidence interval.

4.2. Application to the Gauss- Markoff model with known unit weight error.

Consider the familiar Gauss- Markoff model

$$E(\ell) = Ax, \quad \Sigma(\ell) = Q\sigma^2, \quad P = Q^{-1}$$

Assume that σ^2 is known. Hence $\Sigma(\ell)$ is completely known. Consider a functional ϕ and its BLUE $\tilde{\phi}$

$$\phi = \phi^T x, \qquad \tilde{\phi} = \phi^T \tilde{x}$$

The distribution of $\tilde{\phi}$ is a Gauss distribution with unknown expectation

$$E(\tilde{\phi}) = \phi = \phi^T x$$

and with standard deviation

$$\sigma(\tilde{\phi}) = \sqrt{\phi^T \left(A^T P A\right)^{-1} \phi}\ \sigma$$

The previous subsection describes how to specify a confidence interval for ϕ: Choose a confidence level α, ask a table or a computer for k_α, and specify the interval

$$\tilde{\phi} - k_\alpha \sqrt{\phi^T \left(A^T P A\right)^{-1} \phi}\ \sigma \leq \phi \leq \tilde{\phi} + k_\alpha \sqrt{\phi^T \left(A^T P A\right)^{-1} \phi}\ \sigma$$

or

$$\tilde{\phi} - k_\alpha \sigma(\tilde{\phi}) \leq \phi \leq \tilde{\phi} + k_\alpha \sigma(\tilde{\phi})$$

It covers the unknown value ϕ with a probability α.

A special choice ϕ^T is the j-th row of the unit matrix:
In this case

$$\phi = \phi^T x = x_j$$

the j-th component of the vector of unknown parameters. It holds that

$$\sigma^2(\tilde{x}_j) = \phi^T (A^T P A)^{-1} \phi \, \sigma^2 = q_{x_j x_j} \sigma^2$$

where $q_{x_j x_j}$ is the j-th diagonal element of the inverse normal equation matrix

$$Q_{xx} = (A^T P A)^{-1}$$

The confidence interval is

$$\tilde{x}_j - k_\alpha \sqrt{q_{x_j x_j}} \, \sigma \leq x_j \leq \tilde{x}_j + k_\alpha \sqrt{q_{x_j x_j}} \, \sigma$$

or

$$\tilde{x}_j - k_\alpha \sigma(\tilde{x}_j) \leq x_j \leq \tilde{x}_j + k_\alpha \sigma(\tilde{x}_j)$$

Another special choice of ϕ^T is:

$$\phi^T = (a_{i1}, \cdots, a_{im})$$

i.e. the i-th row of the design matrix A. In this way, one can derive a confidence interval for the i-th component λ_i of the vector

$$\lambda = E(\ell)$$

Remark: Nobody can prevent a person from specifying confidence intervals for a multitude of functionals $\phi^T x$. For example confidence intervals for all parameters x_i, i=1,...,m and/or for all observables λ_i, i=1,...,n could be computed and displayed. One has to be careful not to interpret these confidence intervals and their associated confidence levels in a wrong way. In order to illustrate possible pitfalls, assume 2 confidence intervals, one for x_1, the first parameter, and one for x_2, the second parameter. We then have

$$p\{\tilde{x}_1 - k_\alpha \sigma(\tilde{x}_1) \leq x_1 \leq \tilde{x}_1 + k_\alpha \sigma(\tilde{x}_1)\} = \alpha$$

$$p\{\tilde{x}_2 - k_\alpha \sigma(\tilde{x}_2) \leq x_2 \leq \tilde{x}_2 + k_\alpha \sigma(\tilde{x}_2)\} = \alpha$$

The meaning of these equations is the following one. Suppose that the process of taking observations ℓ and computing estimates $\tilde{x}$ is repeated N times, where N is large. Then approximately in $N\alpha$ cases the first confidence interval covers x_1, and also in $N\alpha$ cases the second interval covers x_2. It is however not clear in how many cases both intervals cover appropriate values simultaneously. Thus the

probability of the joint event

$$\left\{\tilde{x}_1 - k_\alpha\sigma(\tilde{x}_1) \le x_1 \le \tilde{x}_1 + k_\alpha\sigma(\tilde{x}_1),\ \tilde{x}_2 - k_\alpha\sigma(\tilde{x}_2) \le x_2 \le \tilde{x}_2 + k_\alpha\sigma(\tilde{x}_2)\right\}$$

is unknown. This probability would be α^2 if $\tilde{x}_1$ and $\tilde{x}_2$ were independent. It would be α, if they were completely dependent (i.e. if $\tilde{x}_2$ were a function of $\tilde{x}_1$). Generally $\tilde{x}_1$ and $\tilde{x}_2$ are correlated. Hence the probability for the joint event is somewhere between the specified limits. We shall see later how ellipsoidal confidence regions can be specified which cover both x_1 and x_2 with a pre-specified probability α.

4.3. Studentization.

Consider now the Gauss- Markoff model

$$E(\ell) = Ax, \qquad \Sigma(\ell) = Q\sigma^2$$

where σ is now assumed as unknown. If $\phi=\phi^T x$ is a functional, then its BLUE is

$$\tilde{\phi} = \phi^T\tilde{x}$$

as before. The variance is

$$\sigma^2(\tilde{\phi}) = \phi^T\left(A^TPA\right)^{-1}\phi\ \sigma^2$$

It is unknown, because σ^2 is unknown. We are used to estimate σ^2 by

$$\frac{v^T P v}{n-m} = \tilde{\sigma}^2$$

The estimate $\tilde{\sigma}^2$ for σ is unbiased. From section 3.2. we know that $\left(1/\sigma^2\right)v^T P v$ is a χ^2_{n-m}. Hence $E\left(v^T P v\right)=(n-m)\sigma^2$. Thus

$$E\left(\tilde{\sigma}^2\right) = \sigma^2$$

We denote by $\tilde{\sigma}$ the square root of $\tilde{\sigma}^2$ (we cannot claim that $E(\tilde{\sigma})=\sigma$, but $\tilde{\sigma}$ appears to be a reasonable estimate for σ).

We also denote

$$\tilde{\sigma}(\tilde{\Phi}) = \sqrt{\phi^T\left(A^T P A\right)^{-1}\phi}\ \tilde{\sigma}$$

Thus $\tilde{\sigma}(\tilde{\Phi})$ is an estimate for $\sigma(\tilde{\Phi})$. The crucial point is now that

$$t_{n-m} = \frac{\tilde{\Phi}-\Phi}{\tilde{\sigma}(\tilde{\Phi})}$$

has Student's t- distribution with n-m degrees of freedom. The proof is given by the theorem of section 3.4. Just mind that $\tilde{\Phi} = \phi^T\tilde{x}$, that $\Phi=\phi^T x$ has been denoted by c, and that the denominator of the expression in the theorem of section 3.4. is nothing but $\tilde{\sigma}(\tilde{\Phi})$.

We are now able to specify a confidence interval for ϕ. After prescribing α, one uses a table of the t- distribution to find k_α such that

$$p\left\{-k_\alpha \leq t_{n-m} \leq k_\alpha\right\} = \alpha$$

Thus

$$p\left\{-k_\alpha \leq \frac{\tilde{\phi}-\phi}{\tilde{\sigma}(\tilde{\phi})} \leq k_\alpha\right\} = \alpha$$

or

$$p\left\{\tilde{\phi} - k_\alpha\tilde{\sigma}(\tilde{\phi}) \leq \phi \leq \tilde{\phi} + k_\alpha\tilde{\sigma}(\tilde{\phi})\right\} = \alpha$$

Thus a ramdom interval has been specified which covers the unknown value $\phi=\phi^T x$ with a prespecified probability.

4.4. Confidence regions for σ^2.

The quantity

$$\tilde{\sigma}^2 = \frac{v^T P v}{n-m}$$

was seen to be an unbiased estimator for σ^2.

Further more, in section 3.2. it was seen that

$$\chi^2_{n-m} = (n-m)\,\frac{\tilde{\sigma}^2}{\sigma^2} = \frac{v^T P v}{\sigma^2}$$

has the χ^2- distribution with n-m degrees of freedom. Confidence regions for σ^2 may be constructed as follows.

(1) Two- sided confidence interval. After specifying the confidence level α, choose q_α and r_α such that

$$p\{q_\alpha > \chi^2_{n-m}\} = p\{\chi^2_{n-m} > r_\alpha\} = \frac{1-\alpha}{2}$$

Then

$$p\{q_\alpha \leq \chi^2_{n-m} \leq r_\alpha\} = \alpha$$

or

$$p\{q_\alpha \leq (n-m)\frac{\tilde{\sigma}^2}{\sigma^2} \leq r_\alpha\} = \alpha$$

or

$$p\{(n-m)\frac{\tilde{\sigma}^2}{r_\alpha} \leq \sigma^2 \leq (n-m)\frac{\tilde{\sigma}^2}{q_\alpha}\} = \alpha$$

(2) One- sided confidence interval of finite size. Choose q_α such that

$$p\{q_\alpha \leq \chi^2_{n-m}\} = \alpha$$

Then

$$p\{\sigma^2 \leq (n-m)\frac{\tilde{\sigma}^2}{q_\alpha}\} = \alpha$$

(3) One- sided confidence interval of infinite size. Choose r_α such that

$$p\{\chi^2_{n-m} \leq r_\alpha\} = \alpha$$

Then

$$p\{\sigma^2 \geq (n-m)\frac{\tilde{\sigma}^2}{r_\alpha}\} = \alpha$$

The choice of a confidence interval of type (1), (2), (3) depends on the situation. If one is suspicious against σ^2 which are either too small or too

large, then (1) is chosen. If one is suspicious against σ^2 which are too large, then (2) is chosen. Similarly for (3).

4.5. Ellipsoidal confidence regions for sets of linear estimates.

We consider a set of p linear functions

$$\Phi = \phi x$$

together with their BLUE's

$$\tilde{\Phi} = \phi \tilde{x}$$

The matrix Φ is of size p*m. From section 3.3. we know that

$$\chi^2_p = \frac{1}{\sigma^2} (\phi\tilde{x}-\phi x)^T \left[\phi\left(A^TPA\right)^{-1}\phi^T\right]^{-1} (\phi\tilde{x}-\phi x)$$

$$= (\tilde{\Phi}-\Phi)^T \Sigma(\tilde{\Phi})^{-1} (\tilde{\Phi}-\Phi)$$

has the χ^2- distribution with p degrees of freedom.

We also know that

$$F_{p,n-m} = \frac{(\phi\tilde{x}-\phi x)^T \left[\phi\left(A^TPA\right)^{-1}\phi^T\right]^{-1} (\phi\tilde{x}-\phi x)/p}{\left(v^TPv\right)/(n-m)}$$

$$= (\tilde{\Phi}-\Phi)^T \tilde{\Sigma}(\tilde{\Phi})^{-1} (\tilde{\Phi}-\Phi)/p$$

has an F- distribution with p and n-m degrees of freedom. In these formulas we denote as usual

$$\Sigma(\tilde{\Phi}) = \Phi\left[A^T P A\right]^{-1}\Phi^T \sigma^2 \quad \ldots \text{ covariance of } \tilde{\Phi} = \Phi\tilde{x}$$
$$\tilde{\sigma}^2 = \left[v^T P v\right]/(n-m) \quad \ldots \text{ estimate for } \sigma^2$$
$$\tilde{\Sigma}(\tilde{\Phi}) = \Phi\left[A^T P A\right]^{-1}\Phi^T \tilde{\sigma}^2 \quad \ldots \text{ estimate for } \Sigma(\tilde{\Phi})$$

The above formulae put us into the position to specify ellipsoidal confidence regions when σ is either known or unknown.

(1) σ^2 known. Choose a confidence level α, find k_α such that

$$p\left[\chi^2_p \leq k_\alpha\right] = \alpha$$

Then

$$p\left[(\tilde{\Phi}-\Phi)^T \Sigma(\tilde{\Phi})^{-1} (\tilde{\Phi}-\Phi) \leq k_\alpha\right] = \alpha$$

The matrix $\Sigma(\tilde{\Phi})^{-1}$ is known and positive definite.

If M is any p*p positive definite matrix then all points (position vectors) x fulfilling

$$\left[x_0-x\right]^T M \left[x_0-x\right] = c_0$$

are situated on a p- dimensional ellipsoid. The ellipsoid has its center at x_0. The directions of the axes are the directions of the eigenvectors of M. The

lengths of the axes are given by the square roots of c_0 divided by the eigenvalues of M. The points x fulfilling

$$\left(x_0-x\right)^T M \left(x_0-x\right) \leq c_0$$

are situated in the interior and at the boundary of the above ellipsoid.

If we interprete $\tilde{\Phi}$ and Φ as position vectors of points, we see that all points Φ fulfilling

$$(\tilde{\Phi}-\Phi)^T \Sigma(\tilde{\Phi})^{-1} (\tilde{\Phi}-\Phi) \leq k_\alpha$$

are situated in the interior and at the boundary of a p- dimensional ellipsoid centered at $\tilde{\Phi}$. Thus we have specified an ellipsoidal region which covers the unknown p- dimensional point Φ with a prespecified probability α.

(2) σ^2 unknown. Choose α and k_α such that

$$p\left\{F_{p,n-m} \leq k_\alpha\right\} = \alpha$$

i.e.

$$p\left\{(\tilde{\Phi}-\Phi)^T \tilde{\Sigma}(\tilde{\Phi})^{-1} (\tilde{\Phi}-\Phi)/p \leq k_\alpha\right\} = \alpha$$

Hence an ellipsoidal region has been specified covering the unknown point Φ in p- dimensional space with prescribed probability.

5. Tests of linear hypotheses.

5.1. Linear hypotheses.

We again start from the familiar Gauss- Markoff model for n observations and m unknowns

$$E(\ell) = Ax, \quad \Sigma(\ell) = Q\sigma^2, \quad P = Q^{-1}$$

The unit weight error σ may be known or unknown. A linear hypotheses is a system of p linear equations of the form

$$\Phi x = c$$

The vector c is comprised of p pre- specified constants. One could say that a linear hypothesis assumes that p linear functionals Φx on L_A (the space of adjusted observations) have certain pre- specified values c. This assumption is usually called the "null hypothesis". The "alternative hypothesis" would be that $\Phi x \neq c$.

The usual procedure to test the null hypothesis is the following one. The BLUE $\tilde{x}$ for x is an m- dimensional random variable. It has the multidimensional Gauss distribution with (unknown) mean x and covariance matrix $\Sigma(\tilde{x})$. $\Sigma(\tilde{x})$ may be known or unknown. In the latter case $\tilde{\Sigma}(\tilde{x})$ is an estimate for $\Sigma(\tilde{x})$. $\left[\tilde{\Sigma}(\tilde{x})=\left(A^TPA\right)^{-1}\tilde{\sigma}^2,\ \tilde{\sigma}^2=v^TPv/(n-m)\right]$. The m- dimensional space R^m of realizations of $\tilde{x}$ is divided into two regions, a region of acceptance and a region of rejection. If the outcome of $\tilde{x}$ is in the region of acceptance, the null hypothesis is accepted (with some

reservations). If $\tilde{x}$ is in the region of rejection, the null hypothesis is rejected (definitely).

As regions of acceptance the confidence intervals of chapter 4 may be used with the roles of Φ and $\tilde{\Phi}$ interchanged. The region of acceptance is thus an ellipsoid centered at $\Phi=c$. The region of rejection is the complementary region. It is seen that a certain probability α is associated with a test. $1-\alpha$ is the probability of rejecting a true null hypothesis. In section 4 on confidence intervals, α was called "level of confidence". Now we call $1-\alpha$ the "level of significance".

Rejecting a true null hypothesis is called "error of the first kind". The probability of accepting a false hypothesis, i.e. the probability of an "error of the second kind", is more difficult to specify. It depends on "how wrong" the null hypothesis is, i.e. it depends on $\Phi x-c$. If $\Phi x-c$ is small, then the probability of an error of the second kind is near α, and is therefore quite large. This is the reason wy acceptance of the null hypothesis is done with some reservation. A subsequent, larger sample could lead to rejection of an earlier accepted hypothesis.

5.2. Tests of variances.

If σ^2 is unknown, an estimate $\tilde{\sigma}^2$ is available as

$$\tilde{\sigma}^2 = \frac{v^T P v}{n-m}$$

One can adopt the null hypothesis

$$\sigma^2 = \sigma_0^2$$

where σ_0^2 is a pre- specified value. The null hypothesis can be tested by means of the χ^2- distribution with n-m degrees of freedom. One specifies a level of significance 1-α which is small. Recall that α was called confidence level in section 4. As explained in section 4.4., one finds q_α and r_α such that

$$p\left\{q_\alpha \leq (n-m)\,\frac{\tilde{\sigma}^2}{\sigma_0^2} \leq r_\alpha\right\} = \alpha$$

The interval

$$q_\alpha \leq (n-m)\,\frac{\tilde{\sigma}^2}{\sigma_0^2} \leq r_\alpha \quad \text{or} \quad \frac{q_\alpha}{n-m}\,\sigma_0^2 \leq \tilde{\sigma}^2 \leq \frac{r_\alpha}{n-m}\,\sigma_0^2$$

is the region of acceptance. The complementary region is the region of rejection.

Remark: If one is suspicious that σ_0 might have been specified as too small, one is better advised to use a one sided region of acceptance. One finds r_α such that

$$p\left\{(n-m)\,\frac{\tilde{\sigma}^2}{\sigma_0^2} \leq r_\alpha\right\} = \alpha$$

and uses the region of acceptance

$$\tilde{\sigma}^2 \leq \frac{r_\alpha}{n-m}\,\sigma_0^2$$

5.3. A simple example.

A base line used for comparison measurements has a known length of 151.723 m. A newly delivered distance meter gives values ℓ_i, i=1,...,20 listed in table 5.1. The company specifies a standard deviation (root mean square error) of σ=5mm.

i	ℓ_i	v_i*10^4	i	ℓ_i	v_i*10^4
1	151.745	- 105	11	151.752	- 175
2	.743	- 85	12	.730	+ 45
3	.728	+ 65	13	.724	+ 105
4	.728	+ 65	14	.711	+ 235
5	.744	- 95	15	.745	- 105
6	.724	+ 105	16	.738	- 35
7	.739	- 45	17	.730	+ 45
8	.721	+ 135	18	.733	+ 15
9	.744	- 95	19	.738	- 35
10	.731	+ 35	20	.742	- 75

Table 5.1

Reading of a distance meter for a base line with known length of 151.723 m's.

We first check the hypothesis σ=5mm. We perform an adjustment whose Gauss-Markoff model is

$$E\{\ell_i\} = x \quad i = 1,\ldots,n \quad , \qquad \Sigma\{\ell_i\} = I\sigma^2$$

This is the model for direct observations of equal accuracy, the most elementary

model of least squares adjustment.

The BLUE $\tilde{x}$ for x is the arithmetic mean

$$x = \frac{1}{20} \sum_{i=1}^{20} \ell_i = 151.734_5 \text{ m}$$

We calculate corrections (residuals) v_i which are also shown in table 5.1. One computes

$$\tilde{\sigma}^2 = \frac{1}{19} v^T v = \frac{1975}{19} \text{ mm}^2 = 103.9_5 \text{ mm}^2$$

$$\sigma = 10.2 \text{ mm}$$

We are suspicious that σ_0 has been specified to small. Hence we perform a one-sided test as explained in the remark at the end of the previous subsection. Under the null hypothesis $\sigma=\sigma_0=5$mm the quantity

$$(n-m) \frac{\tilde{\sigma}^2}{\sigma_0^2} = 19 \frac{\tilde{\sigma}^2}{\sigma_0^2} = \frac{v^T v}{\sigma_0^2} = \chi^2_{19}$$

has a χ^2-distribution with n-m (=19) degrees of freedom. Taking $1-\alpha=0.05$, i.e. $\alpha=0.95$, and using a table for the χ^2 distribution, we find

$$P\{\chi^2_{19} \leq 30.1\} = 0.95$$

or

$$p\left\{19\,\frac{\tilde{\sigma}^2}{\sigma_0^2} \leq 30.1\right\} = 0.95$$

or

$$p\left\{\tilde{\sigma}^2 \geq \frac{30.1}{19}\,\sigma_0^2\right\} = 0.05$$

The region of rejection is

$$\tilde{\sigma}^2 \geq \frac{30.1}{19}\,\sigma_0^2$$

or

$$\tilde{\sigma} \geq 1.26\sigma_0 = 6.29 \text{ mm}$$

Our value σ=10.2mm is in the region of rejection. Thus the hypothesis $\sigma=\sigma_0$ is rejected. We use $\tilde{\sigma}$=10.2mm instead of σ_0=5mm in further tests.

Remark: Note that there is no reason to put all the blame on the company. It may have been that the instrument was not used according to the specifications.

Our next null hypothesis concerns the measured length of the base line. We postulate that

$$E(\tilde{x}) = 151.723 \text{ m} = c$$

Thus the null- hypothesis assumes that the baseline has been measured without any systematic error. We need information on the variance of $\tilde{x}$. Because σ_0=5mm has been rejected, we do not calculate

$$\sigma^2(\tilde{x}) = \frac{1}{20}\,\sigma_0^{\,2} = 1.25\ \text{mm}^2$$

but we rather estimate

$$\tilde{\sigma}^2(\tilde{x}) = \frac{1}{20}\,\tilde{\sigma}^2 = \frac{(10.2)^2}{20}\ \text{mm}^2 = 5.20\ \text{mm}^2$$

or

$$\tilde{\sigma}(\tilde{x}) = 2.3\ \text{mm}$$

Under the null hypothesis the quantity

$$\frac{\tilde{x}-c}{\tilde{\sigma}(\tilde{x})}$$

whose observed value is

$$\frac{151.734_5 - 151.723}{0.0023} = 5.0$$

has a t- distribution with 19 degrees of freedom. Confer sections 3.4. and 4.3. Using a table for the t- distribution we find

$$p\left\{-2.09 \leq t_{19} \leq 2.09\right\} = 0.95$$

It is seen that the null hypothesis is rejected.

The region of acceptance can also be displayed as:

$$151.723 - 2.09*0.0023 \leq \tilde{x} \leq 151.723 + 2.09*0.0023$$

or

$$151.718_1 \text{ m} \leq \tilde{x} \leq 151.727_8 \text{ m}$$

The value $\tilde{x}=151.734_5$m is outside this range. We conclude that either the instrument is wrong, or that the base line has changed its length, or that the measurement was biased for some reason.

5.4. A sophisticated example.

It shall be tested whether a dam is subsiding as time goes on. After finishing the construction of the dam the leveling network shown in fig.5.1 was measured. It will be called the time 1- network.

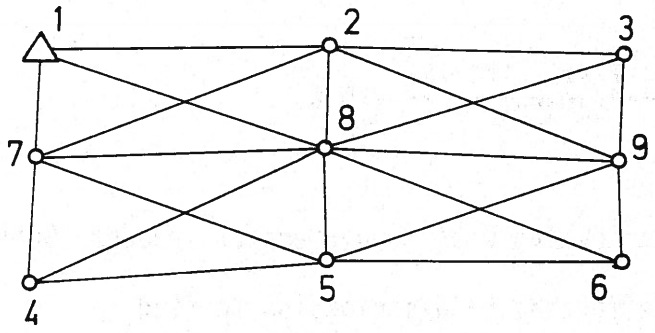


Fig.5.1 Time 1- network

The points 7,8,9 are situated on the dam. The other points 1,...,6 are located on firm ground. The height of point 1 is assumed to be known.

After the lapse of some time (one year maybe), new levelings were carried out. The same instrument and rods were used, the same observer as well. Also otherwise it was attempted to measure under the same conditions as they were given during the first measurement. This justifies the assumption that the unit weight error was the same for both time periods.

Not all height differences were releveled at the second time. The network for the second measurement (i.e. the time 2- network) looked as shown in fig.5.2.

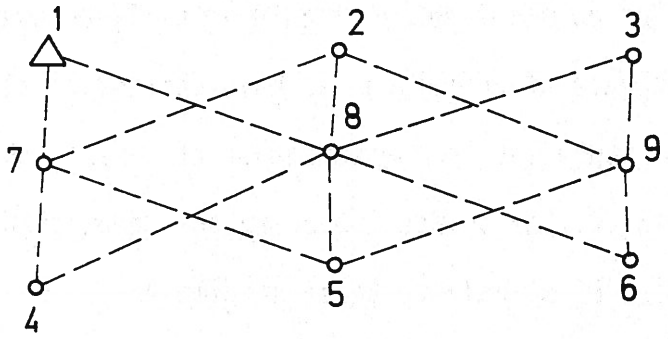


Fig.5.2 Time 2- network

The null hypothesis postulates that the heights of points 7,8,9 did not change. Rejection of the null hypothesis means that at least one height has changed.

The procedure for testing the null hypothesis goes on as follows. Both networks are combined, whereby the points 7,8,9 are duplicated by considering points 7',8',9' in addition. The unprimed points refer to time 1, the primed points refer to time 2. The combined network looks as shown in fig.5.3.

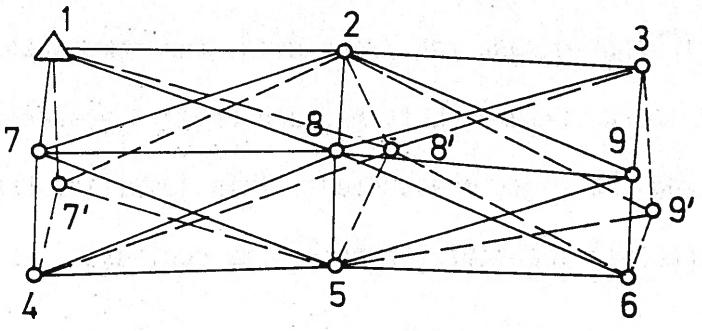


Fig.5.3 Combined network

Note that the points 7 and 7' should actually be drawn on top of each other. However for ease of comparison we have shifted 7' slightly away from 7. The solid lines represent the measurements at time 1, the dashed lines represent those at time 2. The points 7,8,9 are only connected to measurements at time 1, the points 7',8',9' only to those at time 2.

The combined network is now adjusted in agreement with the Gauss- Markoff model

$$E(\ell) = Ax, \qquad \Sigma(\ell) = I\sigma^2$$

We have n=34 and m=11, n-m=23. The vector of parameters x comprises the height of the points 2,3,4,5,6,7,8,9,7',8',9'

$$x = \left(H_2,\ H_3,\ H_4,\ H_5,\ H_6,\ H_7,\ H_8,\ H_9,\ H_{7'},\ H_{8'},\ H_{9'}\right)^T$$

From the vector of 34 residuals we calculate

$$v^T v = \sum_{i=1}^{34} v_i^2$$

We compute the estimate of the unit weight error

$$\tilde{\sigma} = \sqrt{\frac{v^T v}{34-11}} = \sqrt{\frac{v^T v}{23}}$$

The covariance of the adjusted heights is estimated as

$$\tilde{\Sigma}(\tilde{x}) = \left(A^T A\right)^{-1} \tilde{\sigma}^2$$

The null hypothesis is the following linear hypothesis

$$H_7 = H_{7'}$$
$$H_8 = H_{8'}$$
$$H_9 = H_{9'}$$

or

$$H_7 - H_{7'} = 0$$
$$H_8 - H_{8'} = 0$$
$$H_9 - H_{9'} = 0$$

or

$$\begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & -1 \end{bmatrix} x = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

This is denoted

$$\phi x = c = 0, \qquad \text{i.e. } c = 0$$

The best estimates for ϕx are

$$\phi\tilde{x}$$

whereby $\tilde{x}$ comprises the adjusted heights

$$\tilde{x} = \left(\tilde{H}_2,\ \tilde{H}_3,\ \tilde{H}_4,\ \tilde{H}_5,\ \tilde{H}_6,\ \tilde{H}_7,\ \tilde{H}_8,\ \tilde{H}_9,\ \tilde{H}_{7'},\ \tilde{H}_{8'},\ \tilde{H}_{9'}\right)^T$$

Under the null hypothesis the quantity

$$F_{3,23} = (\phi\tilde{x})^T\ \tilde{\Sigma}(\phi\tilde{x})^{-1}\ (\phi\tilde{x})/3$$

with

$$\tilde{\Sigma}(\phi\tilde{x}) = \phi\left(A^TA\right)^{-1}\phi^T\ \tilde{\sigma}^2$$

has the F- distribution with 3 and 23 degrees of freedom. Specifying a significance level of $1-\alpha=0.05$, and using a table for the F- distribution, one finds

$$p\left\{F_{3,23} \leq 3.03\right\} = 0.95$$

The region of acceptance is therefore given by

$$(\phi\tilde{x})^T \tilde{\Sigma}(\phi\tilde{x})^{-1} (\phi\tilde{x}) \leq 3 * 3.03$$

Remark: The above outlined testing procedure requires the calculation of $\phi(A^TA)^{-1}\phi^T$, which may not be easy if conventional software for least squares adjustment is used. The second theorem of section 3.3. offers another possibility to calculate $F_{3,23}$. From the above adjustment of the combined network one just notes v^Tv. The combined network is then adjusted a second time, whereby the pairs of points (7,7'), (8,8'), (9,9') are identified. This is equivalent to an adjustment of

$$E(\ell) = Ax$$
$$\phi x = 0$$

with ϕ as given above. The equations $\phi x = 0$ are used to eliminate $H_{7'}$, $H_{8'}$, $H_{9'}$, so to speak.

Thus, during the second adjustment, one has only 8 parameters instead of 11. These parameters are

$$(H_2, H_3, H_4, H_5, H_6, H_7, H_8, H_9)^T$$

From the residuals v_c of the second adjustment one calculates $v_c^Tv_c$.

According to the second theorem of section 3.3., the quantity

$$F_{3,23} = \frac{(v_c^T v_c - v^T v)/3}{(v^T v)/23}$$

is the same as that one used in the earlier procedure.

Problem: Our null hypothesis was: "The heights of the points 7,8,9 did not change". The alternative hypothesis was therefore "At least one height changed". A change of height is either a decrease or an increase. One may wish to exclude the possibility of increasing heights on a dam which is expected to subside. Hence the null hypothesis could be formulated differently as follows: "The heights of points 7,8,9 did not decrease". Can you imagine a test procedure for this modified null hypothesis?

D. SPECIAL TOPICS

1. Adjustment of Doppler observations.

1.1. The Transit system.

This short subsection cannot replace any solid background information on the Transit Doppler system. For more information the reader may consult e.g. D.E. Wells (1974). (See the reference at the end of this chapter.)

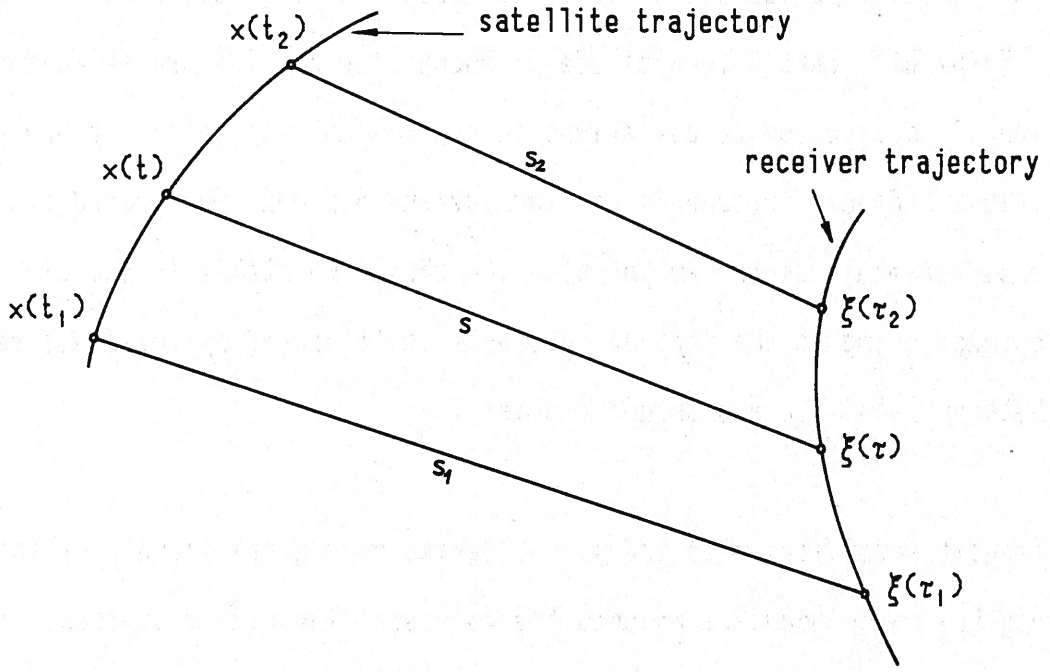
The Navy-Navigation satellite system uses 5 satellites in polar orbits (as of 1979/01/01). Satellite altitude is about 1100 km. The orbital planes are not equally spaced due to deviations in precession. Any satellite broadcasts at two stable harmonic frequencies of 150 MHz and 400 MHz. The ground station Doppler receiver measures the amount by which these two stable frequencies have been changed owing to the Doppler frequency shift caused by the relative velocity between satellite and ground station.

Transit satellites also transmit a series of digital signals by impressing digital phase modulations onto the carriers. The timing of these phase modulations is controlled by the satellite time standard so that they can be used as timing signals. The signals include parameters describing the satellite orbit. (Broadcast ephemerides.)

The two different frequencies allow a correction for errors introduced by the ionosphere. The influence of the troposphere is accounted for by a mathematical model. We do not discuss any relevant details.

There are 4 tracking stations in the United States responsible for monitoring the satellite orbits. About 20 cooperating stations distributed all over the globe contribute observations which are used to calculate and circulate (in a somewhat restricted way) improved a-posteriori ephemerides, called "precise ephemerides".

1.2. Observing a difference in light travel time.



Satellite and receiver are moving in inertial space. A geocentric coordinate system with non-rotating axes can be considered as inertial to a sufficient degree of accuracy. A signal emitted from the satellite at time t is received at the ground-station at time

$$\tau = t + \frac{s}{c}$$

$\tau(t)$...arrival time of signal emitted at time t

$s(t)$...distance traveled by light signal emitted at t and received at τ

c...speed of light.

We occasionally also denote (somewhat inconsistently):

$s(\tau)$...distance traveled by light signal received at time τ. The signal was emitted at time $t = \tau - \frac{s}{c}$.

Considering two signals emitted at t_1, t_2, we get

$$(\tau_2 - \tau_1) = (t_2 - t_1) + \frac{1}{c}(s_2 - s_1).$$

This equation immediately implies the following two relations between differentials:

$$d\tau = dt + \frac{1}{c}ds$$

This can be used in the following two ways:

$$\frac{d\tau}{dt} = 1 + \frac{1}{c}\frac{ds}{dt}$$

$$\frac{dt}{d\tau} = 1 - \frac{1}{c}\frac{ds}{d\tau}$$

1.3. The frequency shift.

Assume now that N oscillation periodes are emitted between t_1 and t_2. The satellite frequency is then

$$f_s = \frac{N}{t_2 - t_1}$$

The same N oscillation periods are received between τ_1 and τ_2. The averaged receiver frequency is then

$$f_r = \frac{N}{\tau_2 - \tau_1}$$

While f_s is constant, f_r is time varying. We see

$$\frac{f_r}{f_s} = \frac{t_2 - t_1}{\tau_2 - \tau_1}$$

In the limit

$$\frac{f_r}{f_s} = \frac{dt}{d\tau} = 1 - \frac{1}{c}\frac{ds}{d\tau} = \left[\frac{d\tau}{dt}\right]^{-1} = \left[1 + \frac{1}{c}\frac{ds}{dt}\right]^{-1}$$

Note that $\frac{ds}{dt}$, $\frac{ds}{d\tau}$ are not relative velocities of satellite and receiver. They are time rates of change of the light travel distance of a signal emitted at t and received at $\tau = t + \frac{s}{c}$.

1.4. Technique of cycle counting.

At the receiver a frequency

$$f_g = f_s + \Delta f$$

is generated. Superposition with the received signal allows to observe the slower cycles of the beat frequency

$$f_b = f_s + \Delta f - f_r$$

There are two basic alternatives

(1) The count is gated (i.e. initiated and terminated) by satellite time marks. Beats between $\tau_1 = t_1 + \frac{s_1}{c}$ and $\tau_2 = t_2 + \frac{s_2}{c}$ are counted.

(2) The count is gated by receiver time marks. Beats between τ_1 and τ_2 are counted, whereby $\tau_2 - \tau_1$ is a fixed time interval.

The most common receivers are Magnavox, JMR and Marconi. Magnavox and Marconi permit both modes. JMR uses (1), however a comparison between satellite time and receiver time takes place.

Corresponding to modes (2) and (1) we get two versions of the subsequent equation

$$D = \int_{\tau_1}^{\tau_2} f_b \, d\tau = \int_{\tau_1}^{\tau_2} (f_s + \Delta f) \, d\tau - \int_{\tau_1}^{\tau_2} f_r \, d\tau =$$

$$= (\tau_2 - \tau_1)(f_s + \Delta f) - \int_{\tau_1}^{\tau_2} f_s \left(1 - \frac{1}{c}\frac{ds}{d\tau}\right) d\tau =$$

$$\boxed{D = (\tau_2 - \tau_1)\,\Delta f + \frac{f_s}{c}(s_2 - s_1)} \qquad (2)$$

$$= \left(t_2 - t_1 + \frac{s_2 - s_1}{c}\right)\Delta f + \frac{f_s}{c}(s_2 - s_1) =$$

$$\boxed{D = (t_2 - t_1)\,\Delta f + \frac{f_g}{c}(s_2 - s_1)} \qquad (1)$$

The fly-by of a satellite allowing uninterrupted observation of its frequency is called a "pass". The duration of a pass is about 20 minutes. During a pass many counts may be observed. The typical duration of a count is between 4.6 seconds and 2 minutes. (This is $\tau_2 - \tau_1$; it is fixed in case of mode (2), and (slightly) variable in case of mode (1).)

1.5. Parameters accounting for receiver imperfections.

The parameters depend on the type of equipment. In particular they depend on the two alternative ways to gate the Doppler counts.

In case of mode (1), i.e. satellite gated counts, the socalled delay of time mark reception is important. Time marks are realized by phase modulations of the carrier frequency. They are processed by different circuits and delayed much stronger than the carrier signal itself. There is a constant delay communicated

by the manufacturer. It is about 500-1000 μs. Superimposed is a variable part of $\pm$30 μs. It is usually modelled by a pass internal parameter.

The delay of time mark reception causes the integration to start at $\tau_1 + \Delta$ and to end at $\tau_2 + \Delta$. Without delay, the starting and ending times would be τ_1 and τ_2. The effect on the observation equation (1) is visible only if arguments of s_1 and s_2 are revealed. We have

$$D = (t_2 + \Delta - t_1 - \Delta)\,\Delta f + \frac{f_g}{c}\,\{s(t_2 + \Delta) - s(t_1 + \Delta)\}$$

Not taking Δ into account would mean that wrong satellite positions are used, namely positions at t_1, t_2 instead of positions at $t_1 + \Delta$, $t_2 + \Delta$.

If a receiver clock is used to gate the counts, a receiver clock offset has a similar effect.

Another receiver error is given if the receivers reference frequency is imperfect. Instead of its nominal value $f_g = f_s + \Delta f$ the receiver frequency may be given by the following expression:

$$f_g + \delta f_g + \dot{\delta f_g}\tau$$

The error splits into a constant part and a drift rate. Both may be introduced as parameters into the observation equation. Reference frequency errors effect the above equations (1),(2).

Let us see what happens in case of equation (1), i.e. in case of satellite gated counts. Assume $\delta\dot{f}_g = 0$. The beat frequency is

$$f_b = f_g + \delta f_g - f_r$$

Since τ_1 and τ_2 are not falsified if gating is prompted by satellite time marks, we have

$$D = \int_{\tau_1}^{\tau_2} f_b \, d\tau$$

This leads us to equation (1) with f_g replaced by $f_g + \delta f_g$:

$$D = (t_2 - t_1)\,(f_g - f_s + \delta f_g) + \tfrac{1}{c}(f_g + \delta f_g)(s_2 - s_1) \qquad (1')$$

1.6. Transformation into an earth-fixed frame.

The rotation of the earth is described by the following rotation matrix

$$U(\tau) = \begin{bmatrix} \cos \omega t & -\sin \omega t & 0 \\ \sin \omega t & \cos \omega t & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

We write

$$\xi(\tau) = U(\tau)\bar{\xi}$$

$$x(t) = U(t)\,\bar{x}(t)$$

$\xi(\tau)$...station position in inertial frame

$\bar{\xi}$...station position in earth-fixed frame

$x(t)$.....satellite position in inertial frame

$\bar{x}(t)$...satellite position in earth-fixed frame.

We have

$$\begin{aligned} s(\tau) &= \|x(\tau - \tfrac{s}{c})\ \xi(\tau)\| = \\ &= \|U(\tau - \tfrac{s}{c})\ \bar{x}\ (\tau - \tfrac{s}{c}) - U(\tau)\ \bar{\xi}\| = \\ &= \|U(-\tfrac{s}{c})\ \bar{x}\ (\tau - \tfrac{s}{c}) - \bar{\xi}\| = \\ &= \|\bar{x}(\tau - \tfrac{s}{c}) - U(\tfrac{s}{c})\ \bar{\xi}\| \end{aligned}$$

(Mind $U(t_1 + t_2) = U(t_1)U(t_2)$, $U(-t) = U^{-1}(t) = U^{T}(t)$)

The above equation could also be written as

$$s(t) = \|\bar{x}(t) - U(\tfrac{s}{c})\ \bar{\xi}\|$$

(Recall the somewhat inconsistent notation $s(t) = s(\tau)$).

1.7. Parameters accounting for orbit corrections.

The following information on satellite orbits is available:

Broadcast ephemerides. They are transmitted by the satellites at 2 minutes intervals. They are based on a 36 hours period of observations at 4 tracking stations in the USA.

They are injected (uploaded, transmitted to satellites) at 12 hours intervals. Their errors may amount to 20-30 meters.

Precise ephemerides. They are post-computed by NWL-DMA. Positions at 1 min. intervals are communicated to users. The precise ephemerides are based on 20 tracking stations distributed over the globe. Their errors are estimated at 2-5 meters.

The orbit is represented in the computer by means of Chebyshev polynomials of degree 7 to 9. (One could imagine that spline interpolation would do somewhat better.) Corrective parameters to the orbits are assumed. Typical are 3 parameters accounting for deviation along track, across track and out of (orbital) plane. The parameters are not allowed to vary freely. They are viewed as pseudo observation with value zero and a prespecified variance. This corresponds to a mixed adjustment model, combining elements of conventional least squares adjustment with techniques of prediction and collocation.

1.8. Linearization of the observation equations.

It is apparent from the previous discussion that the Doppler observation equation may appear in various different shapes. In order to illustrate the principle, we take the nonlinear equation in the form (1') corresponding to satellite gated counts:

$$D = (t_2-t_1)(f_g - f_s + \delta f_g) + \frac{1}{c}(f_g + \delta f_g)[s(t_2+\Delta) - s(t_1+\Delta)]$$

It is seen that we assume a receiver delay Δ and a receiver frequency bias δf_g. We also assume orbital parameters. We denote the along-track, across-track and out-of-plane errors by a, b, c, respectively.

In agreement with section 1.6 we represent

$$s(t+\Delta) = \|\bar{x}(t+\Delta) - U(\tfrac{s}{c})\bar{\xi}\|$$

We assume that $\bar{X}(t)$ is the ephemeris satellite locus in the earth fixed frame. We also introduce unit vectors A(t), B(t), C(t) pointing into the tangential, (i.e along-track), the across-track and into the out-of-plane direction of the orbit. We introduce v(t), the scalar satellite velocity. The vectors A(t), B(t), C(t), and the scalar v(t) may be derived from the ephemeris. The satellite position at time t + Δ is then given by

$$\bar{x}(t+\Delta) = \bar{X}(t+\Delta) + A(t)a + B(t)b + C(t)c$$

Thus we have

$$s(t+\Delta) = \|\bar{X}(t+\Delta) + A(t)a + B(t)b + C(t)c - U(\tfrac{s}{c})\bar{\xi}\|$$

The quantities Δ, a, b, c, $\frac{s}{c}$ are small. Although s is unknown, a known approximate value may be used in $\frac{s}{c}$. We also represent the station coordinate vector as

$$\bar{\xi} = \bar{\xi}^{(0)} + \Delta\bar{\xi}.$$

Here $\bar{\xi}^{(0)}$ are approximate known values. Now comes the familiar linearization procedure. One gets to a sufficient degree of accuracy:

$$U(\tfrac{s}{c}) = \begin{bmatrix} 1 & -\omega\frac{s}{c} & 0 \\ \omega\frac{s}{c} & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

Hence

$$U(\tfrac{s}{c})\bar{\xi} = \begin{bmatrix} \bar{\xi}_1^{(0)} + \Delta\bar{\xi}_1 - \bar{\xi}_2^{(0)}\omega\frac{s}{c} \\ \bar{\xi}_2^{(0)} + \Delta\bar{\xi}_2 + \bar{\xi}_1^{(0)}\omega\frac{s}{c} \\ \bar{\xi}_3^{(0)} + \Delta\bar{\xi}_3 \end{bmatrix}$$

Introducing the direction cosines $\alpha_1(t)$, $\alpha_2(t)$, $\alpha_3(t)$ from $\bar{\xi}^{(0)}$ to $\bar{X}(t)$ one gets

$$s(t+\Delta) = \|\bar{X}(t) - \bar{\xi}^{(0)}\| +$$
$$+ \alpha_1(t)\ \{A_1(t)(v(t)\Delta+a) + B_1(t)b + C_1(t)c - \Delta\bar{\xi}_1 + \bar{\xi}_2^{(0)}\omega_c^s\}$$
$$+ \alpha_2(t)\ \{A_2(t)(v(t)\Delta+a) + B_2(t)b + C_2(t)c - \Delta\bar{\xi}_2 - \bar{\xi}_1^{(0)}\omega_c^s\}$$
$$+ \alpha_3(t)\ \{A_3(t)(v(t)\Delta+a) + B_3(t)b + C_3(t)c - \Delta\bar{\xi}_3\}$$

We abbreviate this as

$$s(t+\Delta) = \|\bar{X}(t) - \bar{\xi}^{(0)}\| + k_0(t) + k_\Delta(t)\Delta + k_a(t)a + k_b(t)b + k_c(t)c +$$
$$+ k_{\bar{\xi}_1}\ \Delta\bar{\xi}_1 + k_{\bar{\xi}_2}\ \Delta\bar{\xi}_2 + k_{\bar{\xi}_3}\ \Delta\bar{\xi}_3$$

Here we have put

$$k_0(t) = (\alpha_1(t)\bar{\xi}_2^{(0)} - \alpha_2(t)\bar{\xi}_1^{(0)})\ \omega_c^s$$

As mentioned above, in the small corrective term ω_c^s an approximate known value for s may be inserted.

We have thus linearized the worst term occurring in the Doppler observation equation.

We use the abbreviation

$$r(t) = \|\bar{X}(t) - \bar{\xi}^{(0)}\|$$

and obtain :

$$
\begin{aligned}
D = {} & (t_2 - t_1)(f_g - f_s) + \frac{f_g}{c}[r(t_2) - r(t_1) + k_0(t_2) - k_0(t_1)] \\
& + \{(t_2 - t_1) + \tfrac{1}{c}[r(t_2) - r(t_1)]\}\, \delta f_g \\
& + \frac{f_g}{c}\, \{(k_\Delta(t_2) - k_\Delta(t_1))\, \Delta \\
& \quad + (k_a(t_2) - k_a(t_1))\, a \\
& \quad + (k_b(t_2) - k_b(t_1))\, b \\
& \quad + (k_c(t_2) - k_c(t_1))\, c \\
& \quad + (k_{\bar{\xi}_1}(t_2) - k_{\bar{\xi}_1}(t_1))\, \Delta\bar{\xi}_1 \\
& \quad + (k_{\bar{\xi}_2}(t_2) - k_{\bar{\xi}_2}(t_1))\, \Delta\bar{\xi}_2 \\
& \quad + (k_{\bar{\xi}_3}(t_2) - k_{\bar{\xi}_3}(t_1))\, \Delta\bar{\xi}_3\}
\end{aligned}
$$

Remark: Because the effect of an along track error and a time delay are nearly equal, the two parameters a and Δ are practically inseparable. The normal equations are near singular unless a-priori pseudo observations for either a or Δ, or both, are faked. The pseudo observations have a value of zero, and a certain weight is given to them.

The typical structure of a Doppler observation equation is

$$
l + v_D = \varphi^T \Delta e + \psi^T \Delta o + \chi^T \Delta\bar{\xi}
$$

Here Δe refers to Δf_g, Δ, and, in general, to parameters resulting from receiver imperfections. Δo refers to a, b, c and, in general, to parameters resulting from orbit corrections. $\Delta\bar{\xi}$ refers to station coordinate increments. l stands for

the difference of D and the constant terms in the above equation. Newly introduced is v_D, the correction to D.

1.9. Single station adjustment.

At a single station a large number of satellite passes is observed. Any pass gives many observations, i.e. integrated Doppler counts D over short periods of time. (Typically, a pass lasts 20 minutes, while the periods for Doppler counts are 4.6 seconds to 2 minutes.)

A suitable adjustment model is one of phased observations. Any pass gives rise to a phase. Common parameters are the station coordinates. Pass-internal parameters are all others, i.e. parameters due to short periodic receiver imperfections and parameters for orbit corrections. Sometimes also meteorological parameters are included.

Precise ephemerides are almost obligatory in order to get meaningful results for single station positioning.

1.10. Multi-station adjustment.

If a satellite is co-observed during a pass from several stations, meaningful results can be obtained also on the basis of broadcast ephemerides. The reason is that orbit uncertainties affect all station locations in nearly the same way, causing, so to speak, a common translation and rotation of the co-observing stations. This mode of observations is sometimes called translocation. It may give reasonably good relative position vectors.

For a group of co-observing stations a phased approach may be used again. Common parameters are the station coordinates. Pass-internal parameters are the orbital corrections as before, the corrective parameters for the receivers, however, multiply. There are as many sets of receiver parameters as passes are observed by individual receivers.

A set of partially reduced normal equations is obtained for a group of co-observing stations. Other groups of stations may be treated similarly. There may be and should be overlaps between the groups. The partially reduced normals of the groups are added and a set of normals for all participating stations is finally obtained and solved.

References and bibliography.

BROWN, D.C. (1976): Doppler positioning by the short arc method. Paper presented at Satellite Doppler positioning International Symposium, Las Cruces, N.M, Oct.1976.

CHEN, J.Y. (1981): Geodetic Datum and Doppler Positioning. Dissertation, Technical University Graz.

JENKINS, R.E.; B.D. Merrit; D.R. Messent; J.R. Lucas (1979): Refinement of positioning software. (DOPPLR). Proceedings of 2nd International Symposium on Satellite Doppler positioning, Austin, Texas, Jan. 1979.

KOUBA, J. (1979): Improvements of Canadian geodetic Doppler programs. Proceedings of 2nd International Symposium on Satellite Doppler positioning, Austin, Texas, Jan. 1979.

SEEBER, G. (1980): Satelliten-Dopplerverfahren. In Pelzer (ed.) Geodaetische Netze in Landes- und Ingenieurvermessung. Konrad Wittwer, Stuttgart, p. 145-162. (In German).

WELLS, D.E. (1974): Doppler Satellite control. Technical Report No. 29, Dep. of Surveying Engineering, Univ. of New Brunswick, Fredericton, N.B., Canada.

2. Geodetic data bases.

2.1. Storage media.

The central memory of a computer allows very rapid processing of data. Searches through tables, matching of data, computations, can be done very fast once the data are in central memory. Data can be accessed there in fractions of 10^{-6} seconds. Central memory is expensive, hence its size is limited. Microcomputers or desktop computers may offer about $32 * 10^3$ to $128 * 10^3$ bytes of central memory. One byte contains 8 bits of information and is usually taken to encode one digit, one alphabetic character or one special symbol. On microcomputers we may have four times as much. On large computers, $4 * 10^6$ to $8 * 10^6$ bytes of central memory may be available.

None of these numbers is sufficient to store the data associated with a large network, nor is it desirable to do this. Data should be in central memory if they are being processed. Otherwise they reside on secondary storage such as disks or tapes. Tapes allow the sequential storage of data. A few multiples of 10^7 bytes may be put on a single reel of tape. Data on tapes can only be accessed sequentially. This makes tapes useless for many applications. However, sequential reading of data from a tape into central memory is fast, typically at a rate of a few 10^5 bytes per second. Tapes offer a very cheap way to store information.

Disks are most suitable for data bases where data must be accessed instantly. The amount of data which can be stored on the disks of one single disk drive is comparable to that of a tape. ($3 * 10^8$ bytes may be stored on some disk drives.)

One may imagine that data are stored on disks in sections. Sequential reading of a section is fast. However, locating the beginning of a new section may require one millisecond to 1/10 of a second.

It is not necessary to acquire a much deeper understanding of computer hardware. The few pieces of information given above shall serve to create a feeling for the difficulties encountered during the design of a data base.

2.2. Requirements for geodetic data bases.

The requirements depend on the type of application. We consider two applications, namely

(1) a data base for the automatically recorded observations of a field project

(2) a data base for a large national network

Let us first discuss the common features of these two types of data bases. In both cases, information on points and on observations is stored. These data are stored in a structured way. There will be indexes serving the rapid access of certain data, e.g. point coordinates, according to specified key values, e.g. point numbers. There will also be cross references (pointers) between the data, allowing e.g. quick access to all observations taken at a certain point. In a different situation, it may be desirable to find all observations whose forepoint (i.e. target point) is the station under consideration.

All situations requiring a linkage of the data must be foreseen when the data base is designed. If unexpected situations arise later on, there is frequently no other way than performing an exhausting (and expensive) search through the entire data base.

The most convenient type of linkage of data is that one of contiguous locations on segments of external storage. Recall that segments of disk storage can be loaded into central memory very quickly. One can imagine that such segments hold points which are located in a certain area. If one point is needed in that area, it is very likely that other points in the vicinity are needed, too. One can also imagine that all observations taken at a certain station are located in one physical segment of external storage.

Unfortunately, it is not possible to realize all types of linkages between data in the convenient way of placing them together. If observations for one standpoint are located together, then the observations having one and the same target point are scattered over various segments.

Let us now briefly elaborate on the differences between the two types of data bases. The purpose of data base (1) is the calculation of coordinates from redundant observations. Quick access to all types of data is necessary. On the other hand, the total amount of data is not large, usually less than 10^6 bytes. Such data can be accommodated on minidisks or diskettes of small desktop computers. Many computer manufacturers offer general purpose data base software that can be used. A self-made solution can be faster and more economical. Such a

system is described in Bartelme, Hofmann-Wellenhof, and Meissl (1981).

Not all data stored on a data base of type (2), i.e. one for a large national network, must be accessed instantly. It may be convenient to get point coordinates quickly. It is also useful to get quick information about the connectivity of points and about the availability of data in a certain region. Hence only a small portion of data should be stored on disk. The mass of observational data, of station descriptions and of historical data can be put onto magnetic tape. It will be sufficient to answer requests for such data within a few days. In the subsequent section we briefly describe the data base that has been established at the U.S. National Geodetic Survey.

2.3. The data base of NGS. [cf. Schwarz (1975)].

As part of the activity related to the readjustment of the North American network, NGS is placing all its horizontal positions, observations and descriptions into a data base. The storage of data is station-oriented. For any station, the following types of information are stored:

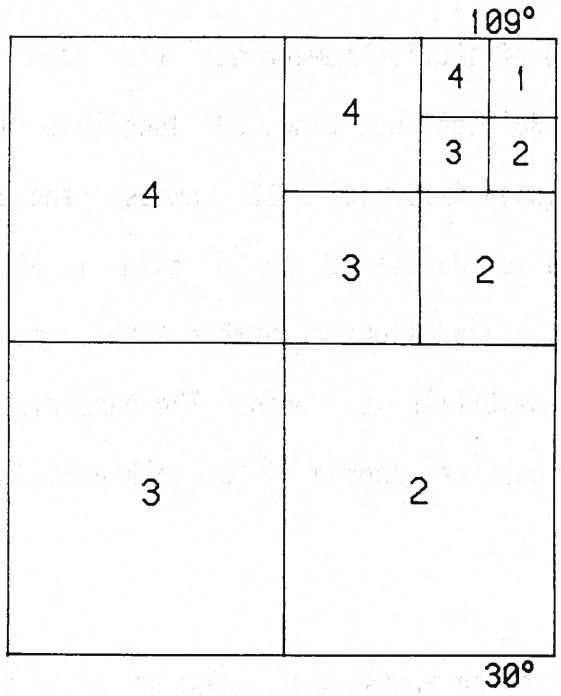
Position, and, if needed, elevation
Terrestrial observations taken at this station
Descriptions
Doppler observations
Astronomical observations
Cross references
Historical data

For one station, all these data occupy contiguous portions on magnetic tape.

For the purpose of station numbering, the area of the U.S. is subdivided into quadrangles of 30' latitude times 30' longitude. In areas of dense control, a further subdivision into 15' * 15' quads, and even into 7.5' * 7.5' quads, may occur. Any quad is identified by a quad identifier (QID). Within a quad, any station carries a quad sequence number (QSN). Thus the concatenation of QID and QSN uniquely identifies a station. The numbering system automatically implies a grouping of stations according to geographical regions. The details are explained by fig. 2.1.

The data base can be accessed by means of a query language. This language offers menues to the user, allowing him to quickly access the index part of the data base, and to submit batch jobs for detailed investigations. Figures 2.2, 2.3 illustrate the separation into an index part and a part with the detailed records.

Indexing of stations



φ λ

QID = 30 109 1 1 1

30′ indicator

15′ indicator

$7\frac{1}{2}$′ indicator

QSN = quad station number
(within 30′, 15′, or $7\frac{1}{2}$′ quads)

Fig. 2.1.

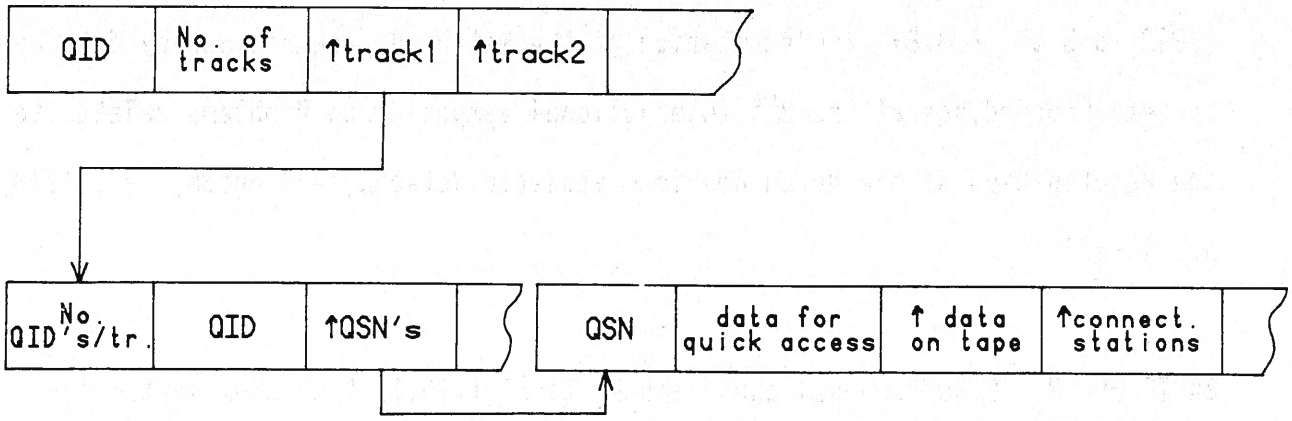
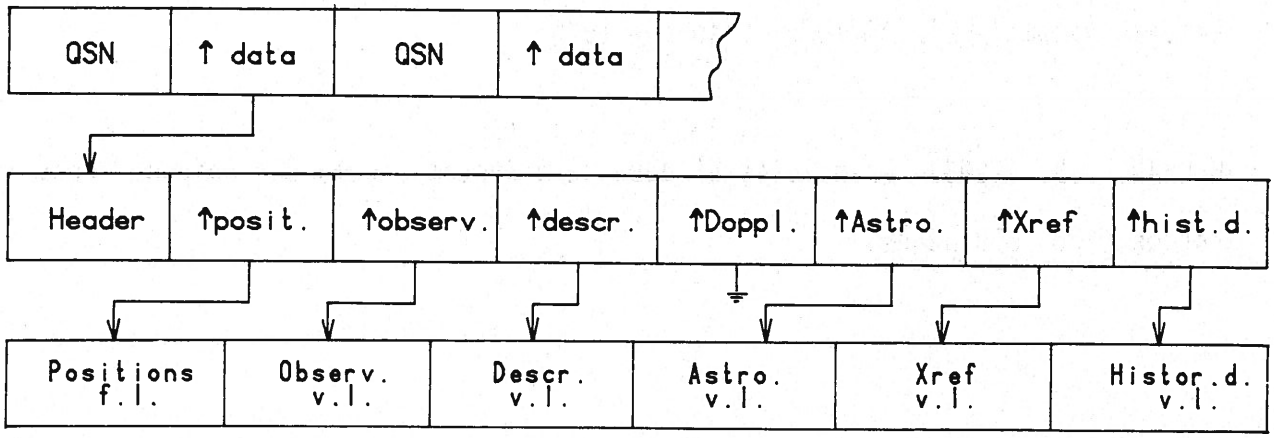


Fig. 2.2. Data stored on disk.



f.l. ... fixed length
v.l. ... variable length

Fig. 2.3. Data stored on tape.

REFERENCES.

ALGER, D.E.Jr. (1978): Implementation of the National Geodetic Survey data base system. Proceedings of the 2nd International Symposium on Problems related to the Redefinition of the North American Geodetic Network, Arlington, Va., 1978, pp.247-257.

BARTELME, N., B.Hofmann-Wellenhof and P.Meissl (1981): A program system for interactive processing of automatically recorded geodetic measurement data. (English translation of an article written in German). ZfV, Jg. 107, Heft 4, pp.144-154.

SCHWARZ, C.R. (1975): The geodetic data base of NGS. Paper presented at XVI General Assembly of IAG/IUGG, Grenoble 1975.

ULLMAN, J.D. (1980): Principles of Data Base Systems. Computer Science Press, Potomac, Md. 379 pages.

3. Cholesky's algorithm applied to the normal equations of geodetic networks.

3.1 Cholesky's algorithm for a general symmetric positive definite system.

Suppose that the system is written in matrix form as

$$A\,x = b$$

Cholesky's algorithm relies on a decomposition of the positive definite matrix A as

$$A = R^T R$$

where R is an upper triangular matrix. During the first or so-called "triangular decomposition phase" of the algorithm, the system is, in effect, multiplied by $(R^T)^{-1}$. The result is the following triangular system:

$$R\,x = s$$

with

$$s = (R^T)^{-1} b$$

During the second or "back-substitution phase" of Cholesky's algorithm, the triangular system is solved for x recursively, starting with the last component of x and proceeding to the first.

The details of Cholesky's algorithm can be best described by switching to indices notation. The original system then reads

$$\sum_{j=1}^{n} a_{ij}x_j = b_i, \quad i = 1, \ldots, n$$

The triangularized system is

$$\sum_{j=i}^{n} r_{ij}x_j = s_i, \quad i = 1, \ldots, n$$

which is calculated from the original system by

$$\left.\begin{aligned} r_{ii} &= \Big(a_{ii} - \sum_{k=1}^{i-1} r_{ki}^2 \Big)^{1/2} \\ r_{ij} &= \Big(a_{ij} - \sum_{k=1}^{i-1} r_{ki}r_{kj} \Big)/r_{ii}, \quad j = i+1, \ldots, n \\ s_i &= \Big(b_i - \sum_{k=1}^{i-1} r_{ki}s_k \Big)/r_{ii} \end{aligned}\right\} \quad i = 1, \ldots, n$$

During the back substitution phase, the triangular system is solved by

$$x_i = \Big(s_i - \sum_{j=i+1}^{n} r_{ij}x_j \Big)/r_{ii}, \quad i = n, \ldots, 1$$

3.2 Partial reduction by Cholesky's algorithm.

Split the original system as

$$A_{11}x_1 + A_{12}x_2 = b_1$$
$$A_{21}x_1 + A_{22}x_2 = b_2$$

Split R accordingly:

$$R = \begin{bmatrix} R_{11} & R_{12} \\ 0 & R_{22} \end{bmatrix}$$

From the equation $R^T R = A$ we deduce the following identities

$$R_{11}^T R_{11} = A_{11}$$

$$R_{11}^T R_{12} = A_{12}$$

$$R_{12}^T R_{12} + R_{22}^T R_{22} = A_{22}$$

Multiply the first set of the original normals by $(R_{11}^T)^{-1}$ and then eliminate the unknowns x_1 from the second set by subtracting proper multiples of the equations of the first set, the multiplying matrix factor being $A_{21}R_{11}^{-1}=R_{12}^T$. The resulting system is

$$R_{11}x_1 + R_{12}x_2 = s_1$$

$$A_{22}^{(p)}x_2 = b_2^{(p)}$$

The second set of these equations is called the "partially reduced" set of normal equations. Explicit and equivalent expressions for the quantities involved are

$$A_{22}^{(p)} = A_{22} - R_{12}^T R_{12} = R_{22}^T R_{22} = A_{22} - A_{21}A_{11}^{-1}A_{12}$$

$$b_2^{(p)} = b_2 - R_{12}^T s_1 = b_2 - A_{21}A_{11}^{-1}b_1$$

These expressions are easily checked by the identities exhibited above. The last expression in any of the two lines reveals that the reduced normal equations do not depend in any way on the peculiarities of Cholesky's algorithm. In fact, any method of elimination that removes the unknowns x_1 from the second set by subtracting proper multiples of the first set must uniquely arrive at the partially reduced normals exhibited above.

In indices notation, the partial Cholesky reduction is

$$\left.\begin{aligned} r_{ii} &= \Big(a_{ii} - \sum_{k=1}^{i-1} r_{ki}^2 \Big)^{1/2} \\ r_{ij} &= \Big(a_{ij} - \sum_{k=1}^{i-1} r_{ki}r_{kj} \Big)/r_{ii}, \quad j = i+1, \ldots, n \\ s_i &= \Big(b_i - \sum_{k=1}^{i-1} r_{ki}s_k \Big)/r_{ii} \end{aligned}\right\} \; i = 1, \ldots, p$$

$$\left.\begin{aligned} a_{ij}^{(p)} &= a_{ij} - \sum_{k=1}^{p} r_{ki}r_{kj}, \quad j = i+1, \ldots, n \\ b_i^{(p)} &= b_i - \sum_{k=1}^{p} r_{ki}s_k \end{aligned}\right\} \; i = p+1, \ldots, n$$

Cholesky's algorithm can be organized in many different ways. The programs used by the NGS, writen by R.H. Hanson and based on earlier work of Poder and

Tscherning (1973) at the Danish Geodetic Institut, execute Cholesky's algorithm in the following manner;

```
FOR j=1 TO n+1
  FOR i=1 TO MIN(n,j)
    SUM = 0
    FOR k=1 TO MIN(p,i-1)
      SUM = SUM + A(k,i)*A(k,j)
    NEXT k
    A(i,j) = A(i,j) - SUM
    IF (i≤MIN(j-1,p)) A(i,j) = A(i,j)/A(i,i)
  NEXT i
  IF (j≤p) A(j,j) = SQRT(A(j,j))
NEXT j
```

In this algorithm, the A(i,j) are place holders. They denote storage locations for a number of quantities. In detail,

* the original coefficients a_{ij} are stored at A(i,j);
  the original coefficients b_i are stored at A(i,n+1);
* the $a_{ij}^{(p)}$ are stored at A(i,j), the $b_i^{(p)}$ at A(i,n+1);
* the r_{ij} are stored at A(i,j), the s_i at A(i,n+1).

It should be stressed that the above algorithm is still a simplification of the actual NGS programs. First, these programs make use of a more complicated data

structure which allows storage and retrieval of coefficients A(i,j) columnwise to and from mass storage (disks). Second, the programs allow for exploiting the sparsity of the normal equations to some extent. The normals have many zero coefficients. If the equations are ordered in a sensitive way, many of the zeros are retained throughout the reduction. This results in a great saving of computer storage und computation time. The NGS programs store only a section of each column, excluding coefficients that will never become nonzero during the execution of the algorithm. We shall come back to the problem of ordering in section 3.5.

Remark: We briefly mention another way to execute Cholesky's algorithm, which amounts to a series of partial reductions for p proceeding from 1 to n. In this fashion Cholesky's algorithm becomes very similar to Gauss' algorithm. Denote $a_{ij}^{(0)}=a_{ij}$ and $b_i^{(0)}=b_i$. We then have

$$
\left.
\begin{aligned}
r_{pp} &= \left(a_{pp}^{(p-1)} \right)^{1/2} \\
r_{pj} &= a_{pj}^{(p-1)}/r_{pp}, \quad j = p+1, \ldots, n \\
s_p &= b_p^{(p-1)}/r_{pp} \\
a_{ij}^{(p)} &= a_{ij}^{(p-1)} - r_{pi}r_{pj} \qquad \left. \begin{aligned} & i = p+1, \ldots, n \\ & j = i, \ldots, n \end{aligned} \right\} \\
b_i^{(p)} &= b_i^{(p-1)} - r_{pi}s_i
\end{aligned}
\right\} \quad p = 1, \ldots, n
$$

If the algorithm stops at any p<n, a partially reduced system results. It adds insight into the problem of equation ordering, discussed later in this section, that any equation is modified by either dividing it by the square root of the diagonal element or by subtracting proper multiples of preceding equations.

Remark. Common to all versions of Cholesky's algorithm is that they operate only on the portion above and including the main diagonal of the matrix A, as well as on the right-hand side. Hence only the upper triangular portion of the matrix A needs to be stored in computer memory. Substantially more storage is saved if the sparse structure of A is exploited, which is typical for matrices associated with network problems.

3.3 Geodetic normal equations.

Our system of normal equations results from a geodetic ground control network. Adjustment is done on a spheroidal rotational ellipsoid. We assume that the reader is familiar with the principles of network adjustment. Our outline will mainly serve to point out peculiarities and to specify the terminology and notation used in the sequel.

The network will be adjusted by variation of parameters. The parameters, or unknowns, are the ellipsoidal coordinates of the stations (points,nodes). Any station has two parameters, namely ellipsoidal latitude and longitude. The so-called orientation unknowns of direction bundles will be eliminated before the normal equations are assembled and will not appear in the final set of

equations.

Approximate coordinates must be known a priori. Denote these coordinates by the vector $p^{(0)}$. The observations ℓ, comprising distances, azimuths, bundles of directions, and Doppler posiotions, will not fit the approximate coordinates. There will be discrepancies $\Delta\ell$, i.e. only the set of observations $\ell-\Delta\ell$ will fit the approximate coordinates. An adjustment applies corrections v to the observations, so that they become the corrected observations $\ell+v$. It also applies shifts Δp to the approximate coordinates so that they become the adjusted coordinates $p=p^{(0)}+\Delta p$. The functional relation between the corrected observations and the adjusted coordinates is (after elimination of the orientation unknowns) in linearized form written as:

$$\Delta\ell + v = B\,\Delta p$$

Weights are prescribed for the individual observations. They are arranged along the diagonal of the weight matrix P which has zero off-diagonal coefficients. Gauss' minimum principles, i.e.,

$$v^T P v = \text{Minimum}$$

is used to uniquely determine v and Δp satisfying the side constraints $\Delta\ell+v=B\Delta p$. The extremum problem leads to the normal equations

$$B^T P B\Delta p = B^T P\Delta\ell$$

which for brevity are written as

$$A\,x = b$$

Note that the unknowns x are actually small shifts leading from the approximate coordinates to the adjusted coordinates.

An important feature of geodetic network adjustment is the local nature of the observations. Any observation involves only a small number of stations which are located close together. For distance and direction observations, direct visibility between two stations must be given. This limits the spacings between stations connected by such a line of vision to 30 km or less in most cases. The normal equation matrix will have only nonzero off-diagonal elements $a_{ij} \neq 0$, if i,j refer either to the two coordinates of one station or to coordinates of two stations connected by a measurement. Such a connection is established either by a direction, a distance, or an azimuth between the two stations, or is due to the preelimination of the orientation unknowns in case of a directional co-observation of the two stations from a third station. The Doppler position observations refer to the two coordinates of one station and will not cause any a_{ij}, $i \neq j$, to be nonzero. While the network covers a large portion of a continent and extends over several thousands of kilometers, there will only be nonzero coefficients a_{ij} if the involved stations are not farther apart than 60 km (in most cases).

Remark. In the literature on numerical linear algebra it is frequently argued that formation and solution of a normal equation system is not a good procedure for doing a least squares adjustment. Instead one should go along with the observation equation system, subjecting it to orthogonalization, singular value decomposition, or other procedures. The argument is based on the condition number of a matrix. The condition number of the normal equation matrix is inferior to that of the observation equations. This is certainly true. On the other hand, it has been proven that storage requirement and computational labor is much less for a geodetic network if it is adjusted by the direct solution of a normal equation system as compared to any other procedure. Refer to the discussion in Avila et al. (1978,p.16). Singular value decomposition or orthogonalization appears to be very efficient for moderately large linear systems that are very ill-conditioned. In the case of very large sparse geodetic network systems which are not extremely ill-conditioned, storage requirement and computational labor are the decisive criteria for selecting a solution method. The observation equation matrix for the U.S. network is of size 3,000,000 x 350,000. To my knowledge no technique is known that preserves sparsity during orthogonalization or singular value decomposition as efficiently as that method which applies direct elimination to the normal equation system, as will be shown later in this chapter.

3.4 Geodetic interpretation of the partial Cholesky-reduced system.

The geodetic meaning of the quantities appearing in the system that has undergone a partial reduction by Cholesky's method is perhaps best understood in

terms of a parameter transformation. The original normals are written as

$$A_{11}x_1 + A_{12}x_2 = b_1$$
$$A_{21}x_1 + A_{22}x_2 = b_2$$

and consider a parameter transformation which changes x_1 into y_1 leaving x_2 unchanged:

$$y_1 = R_{11}x_1 + R_{12}x_2$$
$$x_2 = \qquad\qquad x_2$$

The inverse transformation is

$$x_1 = R_{11}^{-1}y_1 - R_{11}^{-1}R_{12}x_2$$
$$x_2 = \qquad\qquad\qquad x_2$$

The normal equations for the new parameters are

$$y_1 \qquad\qquad = s_1$$
$$A_{22}^{(p)}x_2 = b_2^{(p)}$$

If we substitute for y_1, we get

$$R_{11}x_1 + R_{12}x_2 = s_1$$
$$A_{22}^{(p)}x_2 = b_2^{(p)}$$

This is precisely what we get after partial Cholesky reduction. We see that hidden behind these equations is the system of normal equations involving y_1, x_2. This system completely decomposes into two separate systems for y_1 and x_2. It follows that the adjusted values for y_1, x_2 will be uncorrelated. The covariance matrix for x_2 will be

$$\Sigma(x_2) = (A_{22}^{(p)})^{-1}$$

Let us go back to the original normal equations:

$$A\,x = b$$

If a certain subset of the components of x are forced to fixed values, which amounts to fixing the corresponding coordinates at the values $p^{(0)}+x$, then the normal equations for the remaining unknowns are obtained as follows: Noting that any equation belongs to a certain coordinate, disregard all equations belonging to the fixed components. In the remaining equations, insert the prescribed values for the x's to be fixed, and move these terms toward the right. The desired system results. Note that the same procedure may be applied to the partially reduced Cholesky system

$$R_{11}x_1 + R_{12}x_2 = s_1$$
$$A_{22}^{(p)}x_2 = b_2^{(p)}$$

provided that the fixing is restricted to coordinates out of set x_2. This observation allows us to give the coefficients r_{ij}, s_i, $a_{ij}^{(p)}$, $b_i^{(p)}$ the following geodetic interpretation.

* $a_{ii}^{(p)}$, $i>p$, is the reciprocal of the variance of coordinate i, povided that the coordinates k, $p<k\leq n$, $k\neq i$ are fixed, while the coordinates k, $1\leq k\leq p$, as well as coordinate i itself, are allowed to vary freely.

* $-a_{ij}^{(p)}/a_{ii}^{(p)}$, $i,j>p$, $i\neq j$ is the shift, with respect to the adjusted position, suffered by coordinate i if coordinate j is displaced by one unit from the adjusted position, and if coordinates k, $p<k\leq n$, $k\neq i,j$ are fixed to their adjusted position, while coordinates k, $1\leq k\leq p$ as well as coordinate i itself, are allowed to vary freely.

* $b_i^{(p)}/a_{ii}^{(p)}$, $i>p$ is the shift, with respect to the approximate position, suffered by coordinate i if coordinates k, $p<k\leq n$, $k\neq i$ are fixed to their approximate positions, while coordinates k, $1\leq k\leq p$, as well as coordinate i itself, are allowed to vary freely.

* r_{ii}, $i\leq p$, is the standard deviation of coordinate i, if coordinates k, $i<k\leq n$, are fixed, while coordinates k, $1\leq k\leq i$ are allowed to vary freely.

* $-r_{ij}/r_{ii}$, $i\leq p$, $j>i$ is the shift, with respect to the adjusted position, suffered by coordinate i, provided that coordinate j is displaced by one unit from its adjusted position, that coordinates k, $i<k\leq n$, $k\neq j$ are fixed to their adjusted positions while coordinates k, $1\leq k\leq i$ are allowed to vary freely.

* s_i/r_{ii}, $i\leq p$ is the shift, with respect to the approximate position, suffered by coordinate i, provided that coordinate k, $i<k\leq n$ are fixed to their approximate positions, while coordinates k, $1\leq k\leq i$ can vary freely.

The last three statements require an additional argument because coordinates k, $k \leq p$ are also held fixed, while earlier we said that fixing is restricted to the second set of unknowns, i.e., those with $k > p$.

The three last statements should be clear if we set $i=p$, because then only coordinates $k>p$ are fixed. On the other hand, the r_{ij}'s are no longer subject to any change, as p moves on from i to higher values. Hence the argument also applies for $i<p$.

Remark. (Elastostatic interpretation of normal equations before and after partial reduction.) To the structural engineer the normal equations $Ax=b$ appear as equilibrium equations of an elastic system. The matrix A is called the stiffness matrix, x are coordinate shifts of the nodes, and b are external forces acting at the nodes. The coefficients of the stiffness matrix have the following physical meaning: Suppose that the system is in equilibrium with $x=0$, $b=0$. Displace coordinate j by one unit from its equilibrium position, keeping all other coordinates fixed to their equilibrium position. An elastic force will then be acting on coordinate i. This force is precisely a_{ij}. This holds also for $i=j$. The partially reduced normals $A_{22}^{(p)}x_2=b_2^{(p)}$ refer to a so-called statically reduced system. $A_{22}^{(p)}$ is still a stiffness matrix. $a_{ij}^{(p)}$, $p<i$, $j \leq n$ is the force acting in coordinate i when coordinate j is displaced by one unit from its equilibrium position, when coordinates k, $p<k \leq n$ are fixed, while coordinates k, $1 \leq k \leq n$ are allowed to adjust freely. The right-hand coefficients $b_i^{(p)}$ have the meaning of forces. The original $b_i=b_i^{(0)}$ are nodal forces due to inconsistencies

in the network. As nodes are freed during elimination, different forces $b_i^{(p)}$ must be applied to the remaining nodes such that the equilibrium position of the remaining nodes remains the same. The forces of the eliminated nodes must be transported to the uneliminated ones. Occasionally it is also advantageous to consider external forces. If the vector b is chosen as the j-th column of the unit matrix, the solution x of the system becomes the j-th column f_{*j} of the inverse F of the stiffness matrix A. Hence f_{ij} is the shift of coordinate i if a unit force is applied to coordinate j. Thereby it is assumed that prior to application of the unit force a free equilibrium state had been reached. In particular, f_{ii} is the shift of coordinate i with respect to its adjusted position, if (after adjustment) a unit force is applied to coordinate i. A more lucid interpretation of the variance f_{ii} of the adjusted coordinate i can hardly be given. The elastostatic interpretation is thus somewhat simpler and of great physical significance. I personally prefer to think in terms of elastostatics, where the $a_{ij}^{(p)}$, $b_i^{(p)}$ themselves have a most simple interpretation, whereas in geodetic reasoning the ratios $a_{ij}^{(p)}/a_{ii}^{(p)}$, $b_i^{(p)}/a_{ii}^{(p)}$ are most easily understood. However, since this publication is addressed to the geodesist, elastostatic language will very rarely be used in the sequel. For further details the reader is referred to Rubinstein and Rosen (1970).

Remark. (On the near vanishing of row sums.) Another property of geodetic normal equations is concerned with the row sums

$$\sum_{j=1}^{n} a_{ij}^{(p)}$$

of the original as well as the partially reduced normals. If i is a coordinate whose station - call it P - is involved only in relative measurements, i.e. in measurements other than absolute positioning by Doppler, then the above row sum nearly vanishes for any p. The row sum vanishes precisely if the network is plane. On the ellipsoid it vanishes only approximately. The proof, for the plane network, goes back to the observational equations $Bx=\Delta l+v$. All observational equations involving station P can be thought of as being formulated in terms of differences of coordinate increments. This implies that the row sums pertaining to the station P vanish. The property of station P's vanishing row sums carries over from the observational equations matrix B to the original normal equation matrix $A=B^TPB$. Note that station P's normal equations can be formed by considering only the observations that involve this station. If station P is involved in a Doppler measurement, the row sum of equation i will not vanish, even if the network is plane. However, since the Doppler observations have weights much smaller than those of the relativ measurements (directions, distances, azimuths), the row sum will be appreciably smaller than the larger coefficients in the i-th row of A. Hence, we conclude that all row sums of the normals are small. The remark at the end of section 3.2 tells us that Cholesky's algorithm is a successions of subtractions of multiples of rows from others. Hence the property of near vanishing of row sums is retained throughout reduction and carries over to the partially reduced normal equation matrix $A_{22}^{(p)}$.

3.5 Problem of station ordering.

Coordinate i is associated with row and column i of the normal equations. Ordering the coordinates in a different way leads to a system of normal equations with rows and columns simultaneously permuted, i.e., with diagonal elements permuted rows and columns arranged accordingly. Mathematically, the two systems are equivalent, numerically they are not. Widely recognized in recent literature are the great differences in storage requirement and computation time that result from different orderings and when algorithms are used that take into account the sparseness of A.

In geodetic networks, nonzero off-diagonal elements result from observations between stations rather than between coordinates. The problem of ordering the unknowns becomes a problem of ordering the stations. The two coordinates of one station will always be placed together.

We will refrain from giving a thorough discussion of ordering schemes currently in fashion. We shall briefly review three ordering strategies. The first serves as an introduction to the problem, the other two will be relevant to the readjustment of the U.S. network.

3.5.1 Ordering for small bandwidth.

A supposed geodetic network is depicted in figure 3.1. The solid lines indicate directions observed at both end points. Additional distances and azimuths (measured along some of the solid lines) as well as some Doppler positional

observations may be available. Recall that two stations are connected by nonzero off-diagonal coefficients in the normal equations if there is a direction-, distance-, or azimuth-observation between these two points, or if the two points are directionally co-observed from a third station. In this way, station 1 is connected to stations 2,3,5,6,8,9. Station 8 is connected to 1,2,5,6,9,10,12, 13,14,17,18,19. For any station i we can specify the highest numbered station s_i connected to station i. Thus $s_1=9$, $s_8=19$. We may calculate the number

$$w = 2 * MAX(s_i-i+1)$$

which is called the bandwidth of the system. The factor 2 has been introduced to account for the fact that we have two coordinates per station. In our above example we would have $w=2(s_8-8+1)=24$.

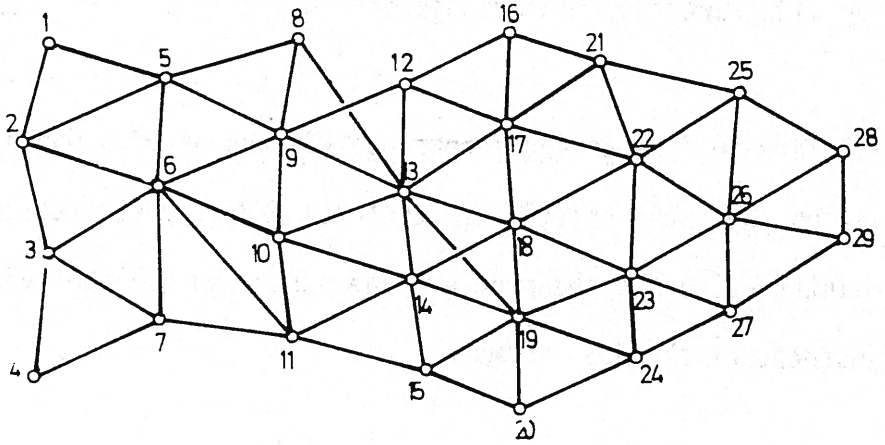


Figure 3.1. - Sample network.

It turns out that the normal equation matrix A will have nonzero coefficients restricted to a band of width w as indicated in figure 3.2. Note that w counts

only lines of coefficients above and including the main diagonal. The coefficients below the main diagonal are never used.

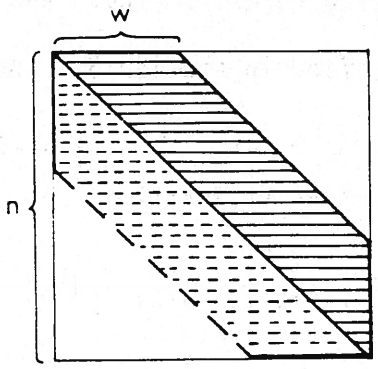


Figure 3.2. - Banded normal equations.

In general, the band will not be completely filled with nonzero coefficients of A. It will also contain some zeroes. The important thing to note, however, is that nonzero coefficients r_{ij}, $a_{ij}^{(p)}$, arising during (partial) Cholesky reduction, are also confined within the band. Some of these will appear at places where A also had nonzero coefficients, and others will take the place of original zeroes. The latter ones are called "fill-in" coefficients.

The proof that fill-in is confined to the band is most easily derived from the next to the last remark in section 3.2. There we saw that any row of any of the Cholesky reduced states results by subtracting multiples of preceding rows from it (and by dividing the row by a factor, if $i \leq p$). However, preceding rows k, $k<i$ can never have nonzero coefficients to the right of the rightmost eligible position for a nonzero coefficient of a row i.

A consequence of the banded structure of A is that any one of the inner products, i.e., the sums of products appearing in Cholesky's algorithm, will have, at most, w-1 nonzero terms. In fact, the first version of the full Cholesky algorithm specified in section 3.1 can be respecified as follows:

$$\left.\begin{aligned} r_{ii} &= \Big(a_{ii} - \sum_{k=MAX(1,i-w+1)}^{i-1} r_{ki}^2 \Big)^{1/2} \\ r_{ij} &= \Big(a_{ij} - \sum_{k=MAX(1,j-w+1)}^{i-1} r_{ki} r_{kj} \Big) / r_{ii} \\ & \quad j = i+1, \ \ldots \ , MIN(n, i+w-1) \\ s_i &= \Big(b_i - \sum_{k=MAX(1,i-w+1)}^{i-1} r_{ki} s_k \Big) / r_{ii} \end{aligned}\right\} \quad i=1, \ \ldots \ ,n$$

and

$$x_i = \Big(s_i - \sum_{j=i+1}^{MIN(n,i+w-1)} r_{ij} x_j \Big) / r_{ii} \qquad i=n, \ \ldots \ ,1$$

On the one hand, a computer program for this algorithm would be more complicated; on the other hand, for w<<n, it would be much faster. It would save much storage if the coefficients within the band were stored in a compacted way, for example, as the columns of an array of size w*n.

A different numbering of the stations would generally result in a different bandwidth w. One could try to minimize w over all possible permutations; however, this is not economical. There are computer algorithms that find near optimal orderings in a short time. Frequently, a good ordering is found by inspection. If a network is elongated, as in the example above, then numbering along the lines that cross the network at the shorter distances often leads to a good ordering. I believe the ordering specified in the figure 3.2 is near optimal.

3.5.2 Ordering for small profile.

As pointed out in section 3.5.1, the saving of computer time and storage comes from the reduced number of product accumulations in Cholesky's algorithm. It is plausible, therefore, that methods have been designed which aim at minimizing the number of product accumulations in the first place rather than doing this indirectly by minimizing the bandwidth. One such method, which is still a compromise between simplicity and efficiency, is ordering for small profile.

The profile of a symmetric matrix includes all elements of a column that are located between the topmost nonzero element and the main diagonal, inclusively. Hence, an element a_{ij}, $i \leq j$ is within the profile, if there is an element $a_{kj} \neq 0$ for a certain $k \leq i$. A typical profile is shown in figure 3.3.

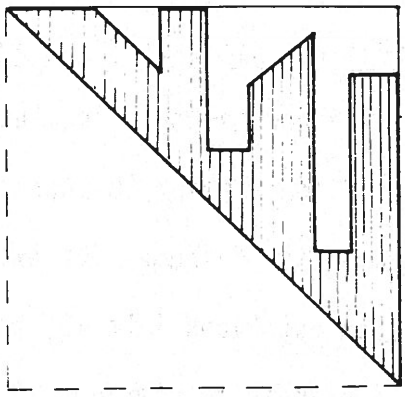


Figure 3.3. - Profiled normal equations.

The profile may include zero coefficients. Again it is important to note that fill-in is restricted to the profile. The proof relies on a similar argument as previously given for the banded structure. Subtracting a multiple of a row from a subsequent row will never cause any nonzero entry outside the profile. The

Cholesky factorization $A=R^TR$ will result in a matrix R which has nonzero coefficients only within the profile. R, being upper triangular, will have zeroes below the main diagonal, whereas A will have coefficients implied by the symmetry there.

NGS computer programs which are currently being used to adjust moderately small networks (up to about 2,500 stations) rely on ordering for a small profile. The ordering algorithm, designed and described by Snay (1976), is heuristic and does not yield a minimal profile in the strict sence. It will, however, establish a fairly small profile in a short time. As will be clear later on, the algorithm will also contribute to the adjustment of the entire U.S. network.

3.5.3 Identifying nonzero coefficients for a certain reduction state.

Before we proceed to still another ordering technique, we pause briefly and reflect on the problem of identifying the nonzero coefficients of A associated with a certain reduction state. Assume, for example, that the partial Cholesky reduction has "eliminated" stations 1 to 12, also marked by black circles in figure 3.4. White circles indicate stations 13 to 29 that participate in the partially reduced system $A_{22}^{(p)}x_2=b_2^{(p)}$. The network is the same as that one in section 3.5, except that the station numbering now conforms with a changed sequence of elimination steps. From section 3.4, dealing with the geodetic interpretation of a Cholesky-reduced system, we infer that the pattern of nonzero coefficients after partial Cholesky reduction up to station p=12, inclusively is shown in figure 3.5.

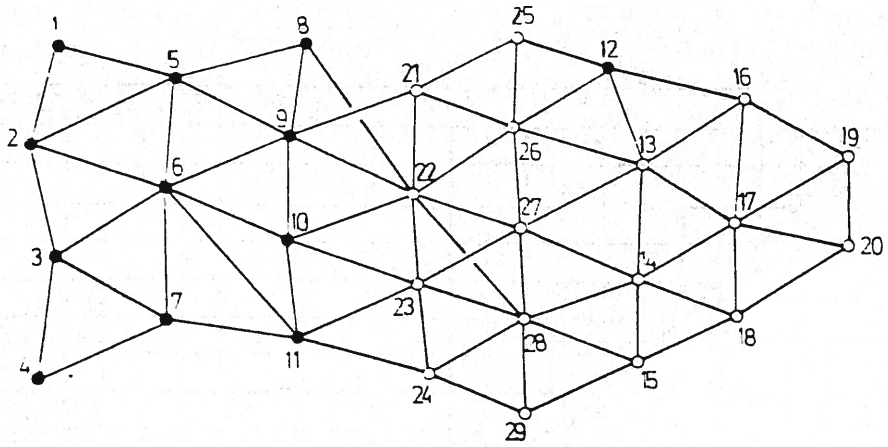


Figure 3.4. - Sample network with stations 1 through 12 eliminated from normal equations.

The numbering of rows and columns of the matrix A in figure 3.5 refers to nodes rather than to coordinates. Hence the individual entries represent actually 2x2 matrices. Heavily shaded entries represent nonzero elements of the original normals. Lightly shaded areas indicate the fill-in which occurs during partial Cholesky reduction up to and including station p=12. Let us give the appropriate argument for a few entries.

* Entry (14,21). The shading indicates fill-in. Why are nodes 14 and 21 connected at this time? According to section 3.4 (cf., the explanation of the expression $-a_{ij}^{(p)}/a_{ii}^{(p)}$ there), we assume that nodes 1 to 12(=p) are free, as well as node 14. We assume the other nodes fixed to their adjusted position, except for node 21, which is displaced from its adjusted position. The displacement of node 21 will cause the direction bundles at the neigboring nodes 25,26 to rotate. As a consequence, the free node 12 will move away from its

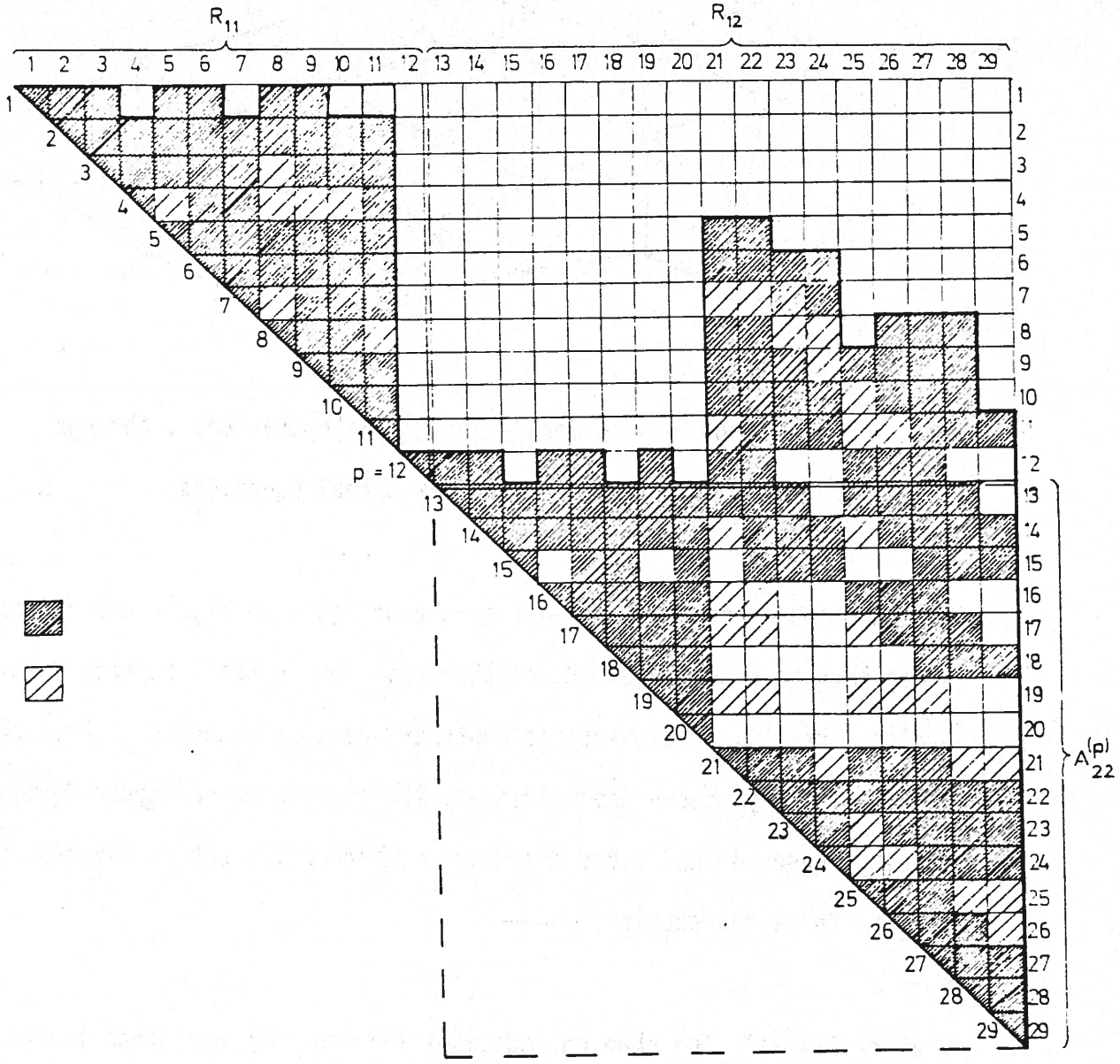


Figure 3.5. - Structure of normal equation when stations 1 through 12 are eliminated.

adjusted position, causing in turn the bundle in 13 to be displaced rotationally. This bundle finally will displace station 14. Hence $a_{14,21}^{(p)}$ will be nonzero, as was to be shown. (The possibility that the resulting movement of 14

is the zero movement is neglected here, as it is in all treatises of sparse matrices.)

* Entry (3,8). The shading indicates fill-in again. This time we refer to the rule for $-r_{ij}/r_{ii}$ given in section 3.4. We pretend that only nodes 1,2,3 have been eliminated, i.e., we temporarily assume p=3. We further assume nodes 4 to 29 fixed to their adjusted positions, except for node 8 which is displaced. This causes the bundle in 5 to deviate from its adjusted position, which in turn displaces nodes 1,2. The displacement of 1 and 2 will finally displace node 3. Hence $r_{3,8}$ must be nonzero, in general.

* Entry (10,13). We may put p=10. Displacing node 13 causes movements of the bundles connected to node 13. No movement takes place to the left of the barrier formed by the double line of nodes 21 to 29. Hence the coefficient must be zero. In fact, coefficients (i,j), $i \leq 11$, $12 \leq j \leq 20$, must be zero. We see that a barrier of a double line of nodes crossing the network can effectively keep down the fill-in. This observation leads us to the ordering scheme considered in the next subsection.

3.5.4 Nested dissection.

We have just seen that by appropriately ordering the stations we may establish barriers which divide the network into parts such that the interior stations of one part will never become connected to interior stations of another part. The numerical analyst George (1973) fully exploited this idea. He calls his ordering

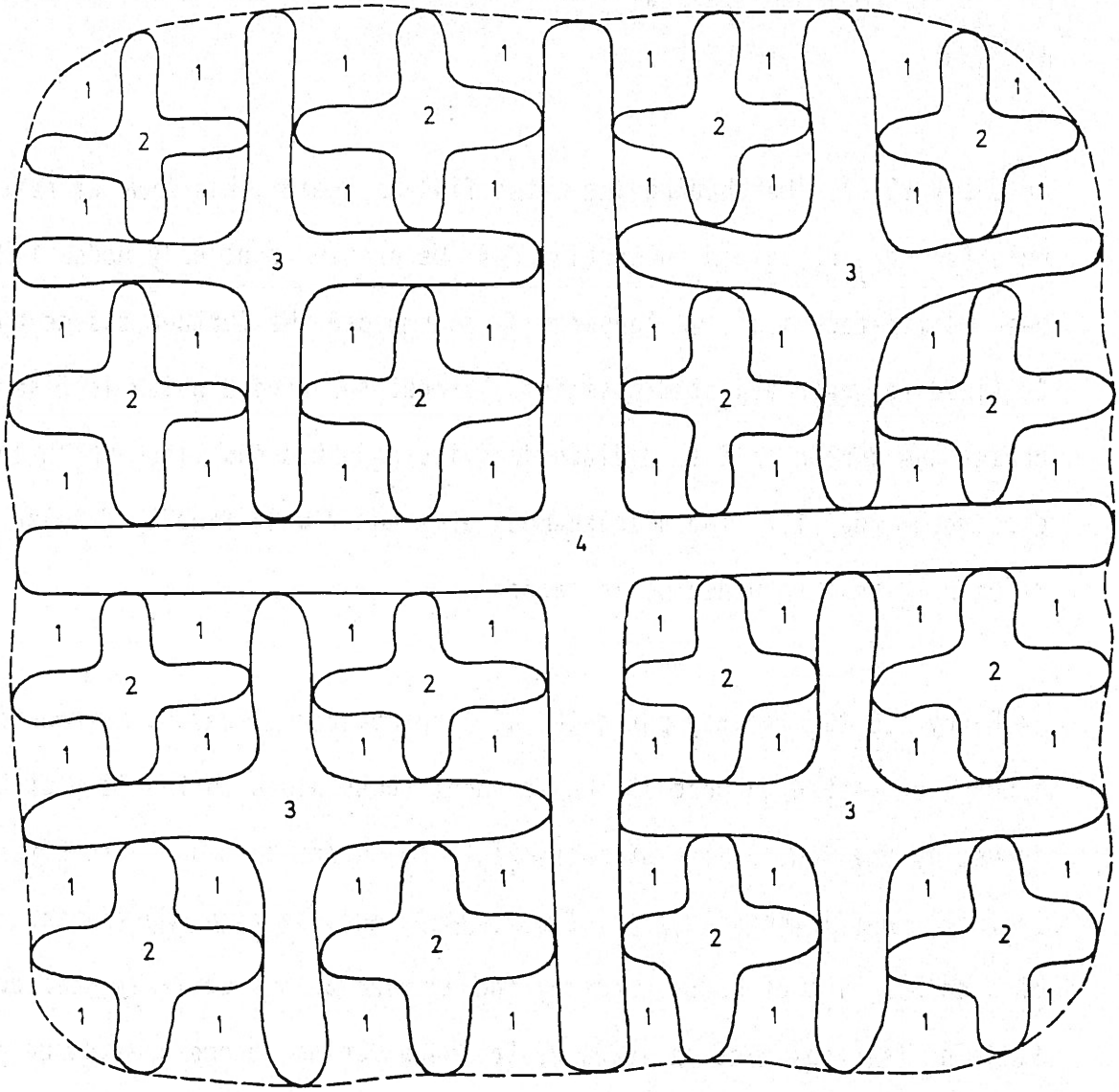


Figure 3.6. - Nested dissection

scheme "nested dissection". As we shall see later, this is anticipated to some extent by what is known among geodesists as "Helmert blocking".

Figure 3.6 exemplifies the idea of nested dissection. The individual stations are not shown here. Instead, we see subsets of stations carrying labels 1 to 4.

We imagine that these labels are attached to all nodes of a particular subset. Nodes carrying label 1 are eliminated first. The sequence in which this is done is not of much importance as long as the number of stations in one connected subset is small. Should this number be larger, we may imagine that an ordering for small profile is done in each individual subset. At the next step we eliminate nodes labeled 2, then 3, and finally 4.

Let us now take a look at the connections a certain node labeled i may encounter to nodes that come later in the ordering sequence. Such nodes carry either the label i or a label j>i. Connections to label i nodes are possible only if the other node is in the same connected label i subset. This is true, because all other label i subsets are separated by barrier subsets of higher labels. Connections of a label i node to nodes of higher labels are only possible if the higher label nodes are located at a barrier surrounding the subset of node i.

Any node will be connected to only a few nodes that come later in the ordering sequence. This is particularly true at the lower levels. It follows that matrix A will be quite sparse, although the pattern of zeroes is now rather complicated.

In order to see the power of nested dissection, we imagine a fairly homogeneous network of n stations covering a region which is shaped somewhat like a square. George (1973) shows that the number of nonzero coefficients (original A plus fill-in) is bounded by

$$const_1 \; n \log n$$

If, in contrast to this, we subject the network to ordering for small bandwidth, we can bound the nonzeroes only by

$$const_2 \; n^{3/2}$$

Also ordering for small profile could not achieve anything much better. Assuming an efficient storage scheme, the storage requirement grows roughly proportional to the number of nonzeroes. However, the factor of proportionality is different from method to method. Nested dissection, in particular, has a more complicated pattern of zeroes that necessitates the storage of additional pointers to keep track of the nonzero elements.

Despite the different proportionality factors and also the difference between $const_1$ and $const_2$ in the above formulas, it becomes clear that asymptotically, i.e., as n grows on and on, nested dissection is superior. In fact, as $n \to \infty$, the ratio of storage requirement for nested dissection and bandwidth tends to zero as const log n / $n^{1/2}$. In this context it is interesting to note that no ordering scheme can improve upon nested dissection asymtotically by more than a constant factor.

We have argued that the number of nonzeroes is directly related to storage requirement. It is also indirectly related to the amount of computational labor. Let us take a look at the number of product accumulations necessary for the

triangular decomposition of A. As it turns out, these product accumulations account for most of the computation time needed to solve the normal equations by Cholesky's method. George (1973) shows that this number is bounded by

$$\text{const}_3\ n^{3/2}$$

if nested dissection is done. Bandwidth ordering, on the other hand, requires

$$\text{const}_4\ n^2$$

for a homogeneous network of the type mentioned. Again the asymptotic superiority of nested dissection becomes evident.

We conclude this subsection with a few remarks.

Remark. Asymptotic superiority of a method does not necessarily mean superiority for moderately small networks. As already indicated, the exploitation of a complicated pattern of zeroes can cause an overhead of storage and computation time. In addition to nonzero coefficients, overhead storage is needed for addressing information which must be stored and for storing a more complicated program.

Remark. Faced with a given network, the subdivision of nodes into categories of different labels is not always immediate. The network will not always be rectangularly shaped, and it will not always be possible to identify a number of

first level sets equal to a power of 4. In pratice, it will be necessary to compromise. Occasionally, the connected subsets of stations of the same label will deviate in number and shape from the ideal case shown in figure 3.6.

Remark. To avoid pitfalls, one must be sure that the barriers dividing the network, as indicated in figure 3.6, are virtually impenetrable. For the types of networks considered, i.e., those involving bundles of directions, distances, azimuths, and absolute positions, the following rule applies. From and to a node of label i there may be lines of vision only to and from; (1) nodes of an adjacent lower label set, (2) nodes of label i which are in the same label i subset, (3) nodes of higher labeled adjacent sets. Otherwise one will try to keep the barriers as thin as possible. Roughly one will arrive at barrier sets composed of double rows of points, as already encountered in the example of figure 3.4. However, there will be exceptions, particularly in the presence of very long lines of vision.

3.5.5 Helmert blocking.

Let us briefly review the basic idea of Helmert blocking for the small network shown in figure 3.4. We reproduce the network in figure 3.7. The dashed line separates two blocks. The nodes marked by simple circles are interior to the relevant block. The nodes marked by double circles are junction nodes, forming a barrier between the two blocks. The normal equations are assembled separately for each block:

$$\text{Block 1:} \begin{bmatrix} A_{11} & B_{13} \\ B_{31} & B_{33} \end{bmatrix} \begin{bmatrix} x_1 \\ x_3 \end{bmatrix} = \begin{bmatrix} a_1 \\ b_3 \end{bmatrix}$$

$$\text{Block 2:} \begin{bmatrix} A_{22} & C_{23} \\ C_{32} & C_{33} \end{bmatrix} \begin{bmatrix} x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} a_2 \\ c_3 \end{bmatrix}$$

Here x_1, x_2 denote the coordinates of stations interior to blocks 1, 2, and x_3 denotes the junction station coordinates. Observations between interior stations of block 1 contribute to block 1 equations. Observations between stations interior to block 1 and junction stations also contribute to it. A similar statement can be made for block 2. Observations between junction stations contribute to the block in which the instrument was positioned. In this context note that the dashed line attributes uniquely a block to any station.

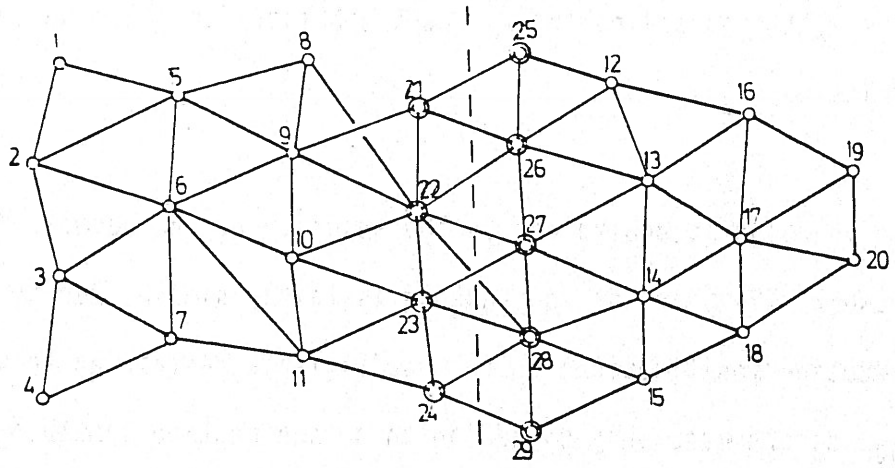


Figure 3.7. - Sample network decomposed into two Helmert blocks.

Adding the two systems of normal equations would result in the conventional normals for the entire network. However, elimination starts for each block

separately. The unknowns x_1, x_2 are eliminated from the two systems by partial Cholesky reduction:

$$\text{Block 1:} \begin{bmatrix} R_{11} & R_{13} \\ & B_{33}^{(p)} \end{bmatrix} \begin{bmatrix} x_1 \\ x_3 \end{bmatrix} = \begin{bmatrix} s_1 \\ b_3^{(p)} \end{bmatrix}$$

$$\text{Block 2:} \begin{bmatrix} Q_{22} & Q_{23} \\ & C_{33}^{(q)} \end{bmatrix} \begin{bmatrix} x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} t_2 \\ c_3^{(q)} \end{bmatrix}$$

The two partially reduced systems for the unknowns x_3 are taken out and added:

$$\left\{ B_{33}^{(p)} + C_{33}^{(q)} \right\} x_3 = \left\{ b_3^{(p)} + c_3^{(q)} \right\}$$

This system is solved for x_3. Back substitution into the two above systems yields x_1, x_2.

The solution is equivalent to the solution of the normals for the entire network. The proof of equivalence is fairly simple. During the partial Cholesky reduction modifications to the coefficients pertaining to x_3, i.e., to B_{33}, b_3, C_{33}, c_3 are made only by adding to or subtracting something from them. Because the quantities added or subtracted are same as they would be if the entire system were partially reduced, it is irrelevant whether the equations for x_3 are added before or after the partial reduction.

A larger network will be partioned into more than two blocks. A hierarchy of blocks can be established that is similar to the nested dissection procedure. In

fact, one can view figure 3.6 as a Helmert blocking scheme. There are as many first-level blocks as there are sets labeled 1, i.e., the number is 64. The normal equations are formed for each first level block separately. Higher labeled nodes situated in adjacent barrier sets take part in the normal equations as junction nodes. The dashed lines separating the first-level blocks have to be imagined as bisecting the barrier sets between the sets labeled 1. All observations must be used in forming the normals, and any observation must be used only once.

The interior nodes are eliminated from the first-level blocks. The partially reduced normals for the junction nodes of four adjacent earlier first-level blocks are added to form the normals of a second-level block. In such a second-level block the nodes labeled 2 now play the role of interior nodes. The junction nodes have labels higher than two. There are 16 second-level blocks. The number of blocks has been reduced by the factor of one-fourth. The interior nodes are eliminated from the second-level blocks, etc. Finally at the fourth and last level we deal with a system for the coordinates of these stations. Back substitution cascades down through the previous levels and successively yields the coordinates of the lower labeled stations.

What is the difference now between Helmert blocking as described here and nested dissection? Not much. In fact, Helmert blocking is slightly more sophisticated because the normals are not fully formed before reduction starts. Instead, the normals are formed separately for each first-level block. After partial reduction at any level, normals of a number of blocks are merged by adding them.

These operation have to be viewed as part of the formation of the normals rather than part of the solution process. George (1973) pointed out that substantial savings are realized in computer time and storage associated with the peculiar way of combining four i-level blocks to form one i+1-level block. Although Helmert blocking has been widely used by geodesists, I do not know of any reference where it has been done by nested dissection. Instead, in most cases, only two levels have been considered. Helmert or his geodetic followers did not appear to anticipate George's logarithmic law.

The U.S. network will be adjusted by the Helmert blocking technique. Partial reduction at the intermediate block level, as well as the reduction of the last level system will be done by Cholesky's method. First-level blocks will be ordered individually for small profile. Higher level blocks will also be ordered to some extent, but ordering becomes less significant as the systems tend to become less and less sparse.

References

AVILA, J., Malloy, B., and Tomlin, J., 1978: Use of the ILLIAC IV for the readjustment of the North American Datum. T.M. 5732, Institute for Advanced Computation, Sunnyvale, Calif.94085, 87 pp.

FORSYTHE, G., and Moler, C.B., 1967: Computer Solution of Linear Algebraic Systems. Prentice-Hall, 148 pp.

GEORGE, A., 1973: Nested dissection of a regular finite element mesh. SIAM Journal of Numerical Analysis, 10(2), 345-363.

MEISSL, P., 1980: A Priori Prediction of Roundoff Error Accumulation in the Solution of a Super-Large Geodetic Normal Equation System. NGS, NOS, Rockville, Md. 20852, 128 pp.

PODER, K., and Tscherning, C.C., 1973: Cholesky's method on a computer. Internal Report No. 8 of the Danish Geodetic Institute, Copenhagen, 22 pp.

RUBINSTEIN, M., and Rosen, R., 1970: Error analysis in structural computation. Journal of the Franklin Institute, 290, 37-48.

SNAY, R.A., 1976: Reducing the profile of sparse symmetric matrices. NOAA Tech. Memorandum NOS NGS-4, Nat. Oceanic and Atmospheric Administration, Rockville, Md.,24 pp.(Available from NTIS, Sprinfield, Va., accession No. PB 258476.)

4. One dimensional cubic spline interpolation.

4.1. Introduction.

We start with the familiar problem of interpolation. We are given a finite number of abscissas $x_0, x_1, \ldots, x_n$, and a corresponding set of function values $y_0, y_1, \ldots, y_n$. The problem is to specify a smooth curve interpolating these data. Without loss of generality we may assume

$$x_0 < x_1 < \ldots < x_n$$

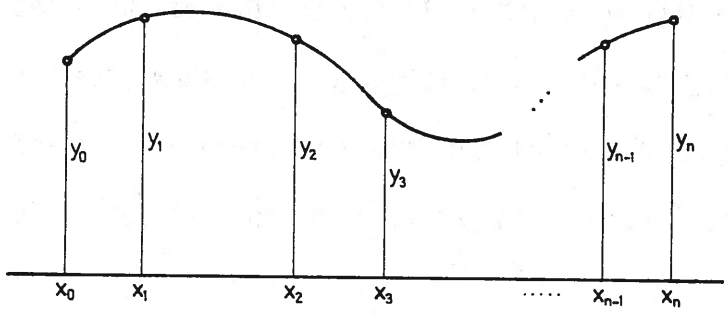


Fig. 4.1.

The abscissas need not be equally spaced.

The classical solution to this problem is polynomial interpolation. The polynomials are either algebraic, i.e.

$$p(x) = a_0 + a_1 x + a_2 x^2 + \ldots + a_n x^n$$

or trigonometric

$$t(x) = \frac{a_0}{2} + a_1 \cos x + b_1 \sin x + a_2 \cos 2x + b_2 \sin 2x + \ldots$$

The polynomials are fairly easy to set up by means of methods specified by Newton, Lagrange and others. They are also quickly evaluated, in particular if they are algebraic.

A disadvantage, often noted in practice, and strongly supported by theory, is a tendency toward instabilities if the number of specified nodes increases. If the polynomial is algebraic, its degree equals the number of locations minus 1. A polynomial of high degree, which is forced to interpolate a set of specified data, has a tendency to oscillate. It can even be shown, that a sequence of interpolating polynomials resulting from a sequence of more and more dense data, will diverge in "most cases". In practice, polynomials of degree exceeding 5 are rarely used.

In a number of fundamental papers, I.I. Schoenberg proposed a different interpolation scheme which is based on the use of piece-wise polynomials. Instead of using a single polynomial interpolating all data $y_0, y_1, \ldots, y_n$ at $x_0, x_1, \ldots, x_n$, Schoenberg proposes to use different polynomials in the successive intervals

$$[x_i, x_{i+1}], \quad i = 0, \ldots, n-1$$

The polynomials are of low degree. Cubic polynomials have proved themselves to be very useful. At the interval boundaries, i.e. at the abscissas

$x_0, x_1, \ldots, x_n$, the polynomials are forced to attain the prescribed values $y_0, y_1, \ldots, y_n$.

This obvious interpolation requirement makes the resulting function s(x) continuous everywhere. In addition one postulates continuity of a number of derivatives s'(x), s''(x), ... at $x_1, \ldots, x_{n-1}$.

The interpolating function s(x) is called a spline functior.. In the case of cubic polynomials we call it a cubic spline. Cubic splines are required to be continuous together with their first and second derivatives. Cubic spline curves are very smooth. Their name "splines" is derived from elastic rules used by Dutch shipbuilders as an aid to draw smooth curves which are constrained to pass through prespecified points.

4.2. Parameterizing a cubic polynomial.

A cubic polynomial is represented as

$$p(x) = a_0 + a_1 x + a_2 x^2 + a_3 x^3$$

A slightly different representation is obtained if the origin is shifted to $x = x_0$:

$$p(x) = a_0 + a_1 (x - x_0) + a_2 (x - x_0)^2 + a_3 (x - x_0)^3$$

In any case, the polynomial has 4 coefficients which serve to parameterize it,

i.e. to identify it. Other ways to parameterize the polynomial are frequently more useful. Suppose that at two locations x_a, x_b, corresponding function values $y_a = p(x_a)$, $y_b = p(x_b)$ are prescribed. Assume that also the derivatives $y'_a = p'(x_a)$, $y'_b = p'(x_b)$ are prescribed. Our aim is to parameterize the polynomial in terms of these 4 values.

We imagine p(x) to be written as

$$p(x) = a_0 + a_1(x - x_a) + a_2(x - x_a)^2 + a_3(x - x_a)^3$$

We require p(x) to attain the prescribed values at $x = x_a$, x_b. Four equations result:

$$y_a = a_0$$

$$y'_a = a_1$$

$$y_b = a_0 + a_1(x_b - x_a) + a_2(x_b - x_a)^2 + a_3(x_b - x_a)^3$$

$$y'_b = a_1 + 2a_2(x_b - x_a) + 3a_3(x_b - x_a)^2$$

This system of equations expresses the new parameters $y_a, \ldots, y'_b$ in terms of the old ones $a_0, \ldots, a_3$. In order to obtain the expressions of $a_0, \ldots, a_3$ in terms of $y_a, \ldots, y'_b$, we solve the linear system for $a_0, \ldots, a_3$. The result is

$$a_0 = y_a$$

$$a_1 = y'_a$$

$$a_2 = \frac{3(y_b - y_a) - (2y'_a + y'_b)(x_b - x_a)}{(x_b - x_a)^2}$$

$$a_3 = \frac{-2(y_b - y_a) + (y'_a + y'_b)(x_b - x_a)}{(x_b - x_a)^3}$$

For later use we express the second derivatives of the polynomial at $x = x_a, x_b$ in terms of the new parameters $y_a, \ldots, y'_b$. Using $p''(x) = 2a_2 + 6a_3(x - x_a)$, we obtain:

$$p''(x_a) = -\frac{4y'_a}{x_b - x_a} - \frac{2y'_b}{x_b - x_a} + \frac{6(y_b - y_a)}{(x_b - x_a)^2} = y''_a$$

$$p''(x_b) = \frac{2y'_a}{x_b - x_a} + \frac{4y'_b}{x_b - x_a} - \frac{6(y_b - y_a)}{(x_b - x_a)^2} = y''_b$$

Remark: We mention in passing that another useful parameterization of a cubic polynomial relies on the parameters y_a, y_b, y''_a, y''_b. Such a parameterization is occasionally used in the literature. However we prefer the one described earlier.

4.3. Condition at the inner nodes.

Our interpolating cubic spline is now represented as

$$s(x) = p_{i,i+1}(x), \quad x_i \leq x \leq x_{i+1}, \quad i = 0, \ldots, n-1$$

The cubic polynomial $p_{i,i+1}(x)$ refers to the interval $[x_i, x_{i+1}]$. The formulas at the previous section apply if we identify

$$x_i = x_a; \quad x_{i+1} = x_b; \quad y_i = y_a; \quad y_{i+1} = y_b$$

Let the polynomial $p_{i,i+1}(x)$ be represented in terms of coefficients as

$$p_{i,i+1}(x) = a_0^{i,i+1} + a_1^{i,i+1}(x - x_i) +$$

$$+ a_2^{i,i+1}(x - x_i)^2 + a_3^{i,i+1}(x - x_i)^3$$

One could parameterize the whole spline s(x) by the set of parameters $a_0^{i,i+1}, \ldots, a_3^{i,i+1}$, $i = 0, \ldots, n-1$. However this set of parameters is redundant. The interpolation requirement $s(x_i) = y_i$, $i = 0, \ldots, n$, and the continuity requirements for s(x), s'(x), s''(x) at x_i, $i = 1, \ldots, n-1$, impose conditions on the parameters $a_j^{i,i+1}$, $i = 0, \ldots, n-1$, $j = 0, \ldots, 3$.

The polynomial $p_{i,i+1}(x)$ is alternatively parameterized by y_i, y_{i+1}, y'_i, y'_{i+1}. This implies a parameterization of the spline s(x) in terms of y_i, y'_i, $i=0,\ldots,n$. This parameterization automatically guarantees that

(1) s(x) interpolates the values y_i at x_i, $i = 0, \ldots, n$

(2) s(x) has a continuous derivative.

A third condition, namely that

(3) s(x) has continuous second derivatives

must be enforced. It result in the set of n-1 equations:

$$p''_{i-1,i}(x_i) = p''_{i,i+1}(x_i), \qquad i = 1, \ldots, n-1$$

Using the expressions for the second derivatives given in the previous section we obtain after division by 2:

$$\frac{1}{x_i - x_{i-1}} y'_{i-1} + 2\left(\frac{1}{x_i - x_{i-1}} + \frac{1}{x_{i+1} - x_i}\right) y'_i + \frac{1}{x_{i+1} - x_i} y'_{i+1} =$$

$$= \frac{3(y_i - y_{i-1})}{(x_i - x_{i-1})^2} + \frac{3(y_{i+1} - y_i)}{(x_{i+1} - x_i)^2}, \qquad i = 1, \ldots, n-1$$

These are n-1 equations for n+1 unknowns $y'_0, \ldots, y'_n$. Two equations are missing. They will be specified in the next section.

4.4. Boundary conditions.

In the previous section we have seen that a cubic spline is not uniquely determined by the interpolation requirement. Two additional conditions must be imposed. They are usually formulated as boundary conditions. We consider the following 3 types of boundary conditions:

(1) Constrained spline. The values of y'_0 and y'_n are prescribed.

(2) Free spline. The conditions are

$$y''_0 = y''_n = 0$$

Explicitely (using the last equations of section 4.3):

$$\frac{2}{x_1 - x_0} y'_0 + \frac{1}{x_1 - x_0} y'_1 = \frac{3(y_1 - y_0)}{(x_1 - x_0)^2}$$

$$\frac{1}{x_n - x_{n-1}} y'_{n-1} + \frac{2}{x_n - x_{n-1}} y'_n = \frac{3(y_n - y_{n-1})}{(x_n - x_{n-1})^2}$$

(3) The periodic spline. It relies on the periodicity of the data, i.e.

$$y_0 = y_n$$

Continuity of the first derivative implies the assumption

$$y'_0 = y'_n$$

Continuity of the second derivative must be enforced also at x_0, x_n respectively:

$$p''_{0,1}(x_0) = p''_{n-1,n}(x_n)$$

i.e.

$$\frac{1}{x_n - x_{n-1}} y'_{n-1} + 2\left(\frac{1}{x_n - x_{n-1}} + \frac{1}{x_1 - x_0}\right) y'_0 + \frac{1}{x_1 - x_0} y'_1 =$$

$$= \frac{3(y_n - y_{n-1})}{(x_n - x_{n-1})^2} + \frac{3(y_1 - y_0)}{(x_1 - x_0)^2}$$

Other types of boundary conditions are possible, but will not be considered here.

4.5. Tridiagonal linear system.

Boundary conditions (1) and (2) listed in the previous section lead , together with the continuity conditions of s''(x) at the inner nodes, to a linear system which is tridiagonal

$$b_0 y'_0 + c_0 y'_1 = d_0$$

$$a_i y'_{i-1} + b_i y'_i + c_i y'_{i+1} = d_i, \quad i = 1, \ldots, n-1$$

$$a_n y'_{n-1} + b_n y'_n = d_n$$

The system looks like

$$\begin{bmatrix} b_0 & c_0 & & & & & & \\ a_1 & b_1 & c_1 & & & & & \\ & a_2 & b_2 & c_2 & & & & \\ & & a_3 & b_3 & c_3 & & & \\ & & & \ddots & \ddots & \ddots & & \\ & & & & \ddots & \ddots & \ddots & \\ & & & & & a_{n-1} & b_{n-1} & c_{n-1} \\ & & & & & & a_n & b_n \end{bmatrix} \begin{bmatrix} y'_0 \\ y'_1 \\ y'_2 \\ y'_3 \\ \vdots \\ \\ \\ y'_{n-1} \\ y'_n \end{bmatrix} = \begin{bmatrix} d_0 \\ d_1 \\ d_2 \\ d_3 \\ \vdots \\ \\ \\ d_{n-1} \\ d_n \end{bmatrix}$$

By straightforward Gauss elimination we transform the subdiagonal elements a_i into zeros. This is called the triangular decomposition phase:

Put

$$\bar{b}_0 = b_0 \,, \quad \bar{d}_0 = d_0$$

Compute:

$$\bar{b}_k = b_k - \frac{a_k}{\bar{b}_{k-1}} c_{k-1} \,, \quad \bar{d}_k = d_k - \frac{a_k}{\bar{b}_{k-1}} \bar{d}_{k-1} \,, \quad k = 1, \ldots, n$$

The triangularized system is

$$\begin{bmatrix} \bar{b}_0 & c_0 & & & & & \\ & \bar{b}_1 & c_1 & & & & \\ & & \bar{b}_2 & c_2 & & & \\ & & & \ddots & \ddots & & \\ & & & & \ddots & \ddots & \\ & & & & & \bar{b}_{n-1} & c_{n-1} \\ & & & & & & \bar{b}_n \end{bmatrix} \begin{bmatrix} y'_0 \\ y'_1 \\ y'_2 \\ \vdots \\ \vdots \\ y'_{n-1} \\ y'_n \end{bmatrix} = \begin{bmatrix} \bar{d}_0 \\ \bar{d}_1 \\ \bar{d}_2 \\ \vdots \\ \vdots \\ \bar{d}_{n-1} \\ \bar{d}_n \end{bmatrix}$$

Back-substitution gives the solution:

$$y'_n = \frac{\bar{d}_n}{\bar{b}_n}$$

$$y'_k = \frac{\bar{d}_k - c_k y'_{k+1}}{\bar{b}_k} \quad , \quad k = n-1, n-2, \ldots, 2, 1, 0$$

These calculations can be done very fast in a computer. Only four vectors whose components are a_k, b_k, c_k, d_k, $k = 0, \ldots, n$ need to be stored.

4.6. Modification for the periodic case.

In the periodic case the linear system looks as follows

$$\begin{bmatrix} b_0 & c_0 & & & & & a_0 \\ a_1 & b_1 & c_1 & & & & \\ & a_2 & b_2 & c_2 & & & \\ & & \ddots & \ddots & \ddots & & \\ & & & \ddots & \ddots & \ddots & \\ & & & & a_{n-2} & b_{n-2} & c_{n-2} \\ c_{n-1} & & & & & a_{n-1} & b_{n-1} \end{bmatrix} \begin{bmatrix} y'_0 \\ y'_1 \\ y'_2 \\ \cdot \\ \cdot \\ \cdot \\ y'_{n-2} \\ y'_{n-1} \end{bmatrix} = \begin{bmatrix} d_0 \\ d_1 \\ d_2 \\ \cdot \\ \cdot \\ \cdot \\ d_{n-2} \\ d_{n-1} \end{bmatrix}$$

The system is not completely tridiagonal. Hence the formulas for Gauss elimination must be modified. The triangular decomposition phase starts by treating only the first n-2 equations as follows:

Put

$$\bar{b}_0 = b_0; \quad e_0 = \bar{e}_0 = a_0; \quad e_k = 0; \quad k = 2, \ldots, n-3; \quad e_{n-2} = c_{n-2}$$

Thus e_k, $k = 1, \ldots, n-2$ gives the first n-2 elements of the last column of the matrix.

Compute:

$$\bar{b}_k = b_k - \frac{a_k}{\bar{b}_{k-1}} c_{k-1}$$

$$\bar{d}_k = d_k - \frac{a_k}{\bar{b}_{k-1}} \bar{d}_{k-1} \quad , \qquad k = 1, 2, \ldots, n-2$$

$$\bar{e}_k = e_k - \frac{a_k}{\bar{b}_{k-1}} \bar{e}_{k-1}$$

After these operations the system looks as

$$\begin{bmatrix} \bar{b}_0 & c_0 & & & & \bar{e}_0 \\ & \bar{b}_1 & c_1 & & & \bar{e}_1 \\ & & \bar{b}_2 & c_2 & & \bar{e}_2 \\ & & & \ddots & \ddots & \vdots \\ & & & & \ddots & \vdots \\ & & & & \bar{b}_{n-2} & \bar{e}_{n-2} \\ c_{n-1} & & & & a_{n-1} & \bar{b}_{n-1} \end{bmatrix} \begin{bmatrix} y'_0 \\ y'_1 \\ y'_2 \\ \vdots \\ \vdots \\ y'_{n-2} \\ y'_{n-1} \end{bmatrix} = \begin{bmatrix} \bar{d}_0 \\ \bar{d}_1 \\ \bar{d}_2 \\ \vdots \\ \vdots \\ \bar{d}_{n-2} \\ \bar{d}_{n-1} \end{bmatrix}$$

The triangular decomposition phase is completed by the following set of formulas.

Put

$$f_0 = c_{n-1}$$

$$f_k = 0 \qquad k = 2, \ldots, n-3$$

$$f_{n-2} = a_{n-1}$$

Thus $f_0, \ldots, f_{n-1}$ represent the first n-2 elements in the last row. Also put

$$\bar{f}_0 = f_0 .$$

Compute

$$\bar{f}_k = f_k - \frac{f_{k-1}}{\bar{b}_{k-1}} c_{k-1} \quad , \qquad k = 1, \ldots, n-2$$

and:

$$\bar{\bar{b}}_{n-1} = \bar{b}_{n-1} - \sum_{k=1}^{n-1} \frac{f_{k-1}}{\bar{b}_{k-1}} \bar{e}_{k-1}$$

$$\bar{\bar{d}}_{n-1} = \bar{d}_{n-1} - \sum_{k=1}^{n-1} \frac{f_{k-1}}{\bar{b}_{k-1}} \bar{d}_{k-1}$$

The system is now triangular:

$$\begin{bmatrix} \bar{b}_0 & c_0 & & & & \bar{e}_0 \\ & \bar{b}_1 & c_1 & & & \bar{e}_1 \\ & & \bar{b}_2 & c_2 & & \bar{e}_2 \\ & & & \ddots & \ddots & \vdots \\ & & & & \bar{b}_{n-2} & \bar{e}_{n-2} \\ & & & & & \bar{\bar{b}}_{n-1} \end{bmatrix} \begin{bmatrix} y'_0 \\ y'_1 \\ y'_2 \\ \vdots \\ y'_{n-2} \\ y'_{n-1} \end{bmatrix} = \begin{bmatrix} \bar{d}_0 \\ \bar{d}_1 \\ \bar{d}_2 \\ \vdots \\ \bar{d}_{n-2} \\ \bar{\bar{d}}_{n-1} \end{bmatrix}$$

Back-substitution is done by

$$y'_{n-1} = \frac{\bar{\bar{d}}_{n-1}}{\bar{\bar{b}}_{n-1}}$$

$$y'_{n-2} = \frac{\bar{d}_{n-2} - \bar{e}_{n-2} y'_{n-1}}{\bar{b}_{n-2}}$$

$$y'_k = \frac{\bar{d}_k - c_k y'_{k+1} - \bar{e}_k y'_{n-1}}{\bar{b}_k} , \qquad k = n-3, \ldots, 0$$

4.7. Interpolation of curves in the plane.

The parameter representation of a curve in the plane is given by two functions

$$x = x(t) , \quad y = y(t).$$

There is some redundancy in the parameter representation. It defines not only the shape of the curve; it supplies in addition a mapping of an interval of the real line onto the curve. One can require that the parameter t equals s, the arc length along the curve. In our subsequent formulas t will be close to s, but not quite identical.

Let a discrete set of points (x_i, y_i), $i = 0, \ldots, n$ be given. The requirement is to interpolate a smooth curve through those points. Cf. fig. 4.2.

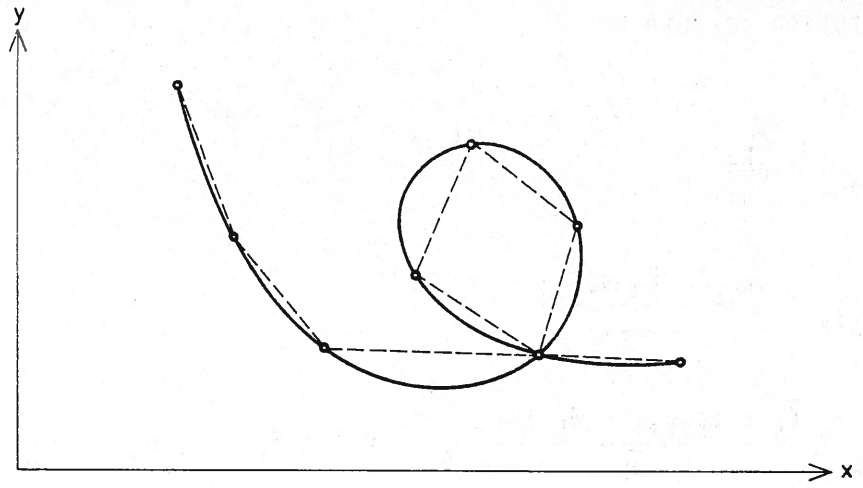


Fig. 4.2.

Note that the curve of fig. 4.2 could not be represented as

$$y = f(x) \quad \text{or} \quad x = g(y).$$

The functions f(x), g(y) would not be single valued.

We consider the polygon of chords also shown in fig. 4.2. Our first choice of the parameter t will be the arc length along this polygon. We arrive at two conventional interpolation problems:

Interpolate

$$x(t_i) = x_i, \quad i = 0, \ldots, n$$

and

$$y(t_i) = y_i, \quad i = 0, \ldots, n$$

for t_i being the length from point 0 to point i measured along the polygon of chords. We use the apparatus developed in the previous section to obtain two spline curves x(t), y(t). The boundary conditions are chosen according to the given situation. A closed curve would require periodic boundary conditions. The resulting curve has continous curvature, (because x(t) and y(t) are twice differentiable functions).

A slight flaw is that t is not the arc length. One can improve upon this by computing the arc length at the points 0, ..., n along the interpolated curve by means of numerical integration. One obtains value $s_0 = 0, s_1, \ldots, s_n$. Replacing $t_0 = 0, t_2, \ldots, t_n$ by these values, one could recompute the spline. The procedure could be iterated a few times. Frequently, however, one is satisfied with t being the arc length along the polygon.

4.8. Splines viewed as a vector space.

Given a fixed partition $x_0 < x_1 < \ldots < x_n$, we consider the set of all spline functions for all possible ordinates $y_0, \ldots, y_n$, and for all possible boundary conditions. One readily verifies that this set forms a vector space. Because n+3 parameters are necessary to uniquely specify a particular spline (a possible choice for those parameters is a set of values $y_0, \ldots, y_n$ together with y'_0, y'_n!), the dimension of the vector space is n+3. We can construct a basis as follows.

Let $a_i(x)$ be the spline fulfilling

$$a_i(x_j) = \delta_{ij} \qquad i,j=0,\ldots,n$$

and

$$a_i'(x_0) = a_i'(x_n) = 0$$

Let $\alpha_0(x)$, $\alpha_n(x)$ be two splines fulfilling

$$\alpha_0(x_j) = \alpha_n(x_j) = 0 \qquad j=0,\ldots,n$$
$$\alpha_0'(x_0) = 1, \quad \alpha_0'(x_n) = 0$$
$$\alpha_n'(x_0) = 0, \quad \alpha_n'(x_n) = 1$$

The splines $a_0(x),\ldots,a_n(x)$, $\alpha_0(x)$, $\alpha_n(x)$ form a basis. In fact the function

$$s(x) = \sum_{i=0}^{n} y_i a_i(x) + y_0'\alpha_0(x) + y_n'\alpha_n(x)$$

gives precisely the spline interpolating $y_0,\ldots,y_n$ and having boundary derivatives $s'(x_0) = y_0'$, $s'(x_n) = y_n'$. Fig. 4.3 shows some of the basis functions, assuming n=8 and equidistant data $x_i = i$, $i=0,\ldots,n$.

Another basis which is practically more important will be discussed in chapter 6.

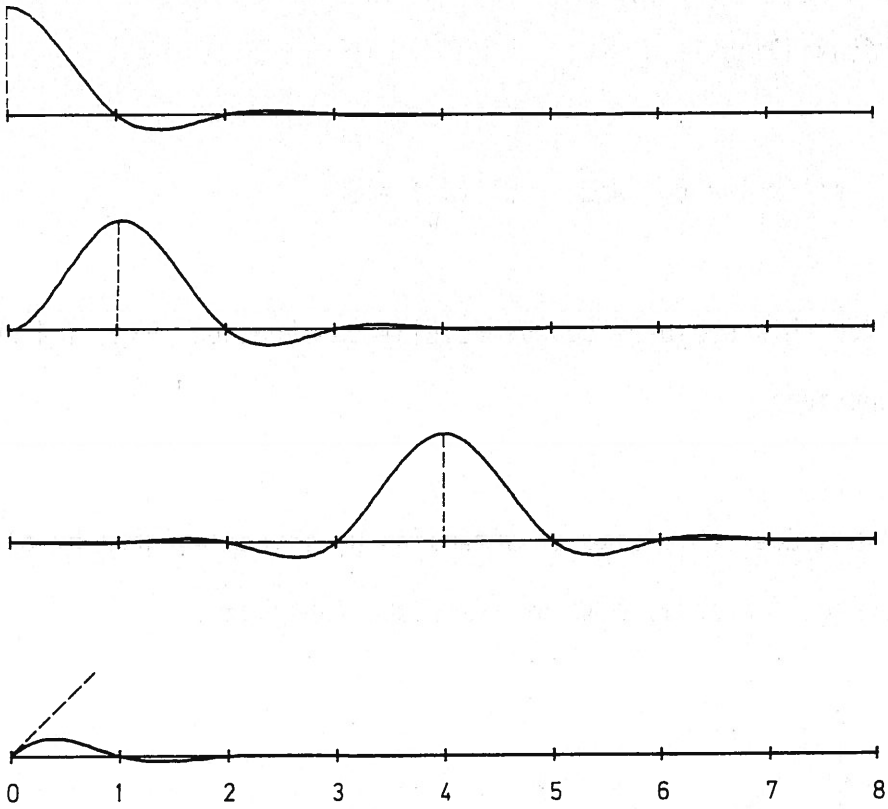


Fig. 4.3

A sample of basis splines for n=8

The splines s(x) fulfilling

$$s'(x_0) = 0 \quad \text{and} \quad s'(x_n) = 0$$

form a subspace of dimension n+1. A basis for this subspace is given by $a_i(x)$, $i=0,\ldots,n$.

Another subspace of dimension n+1 is given by the set of all free splines. i.e. splines fulfilling

$$s''(x_0) = 0 \quad \text{and} \quad s''(x_n) = 0$$

A basis for this subspace can be readily constructed. Fig. 4.4 shows some of the basis functions.

A third subspace, this time of dimension n, is represented by all periodic splines. Fig. 4.5 shows some of its basis functions.

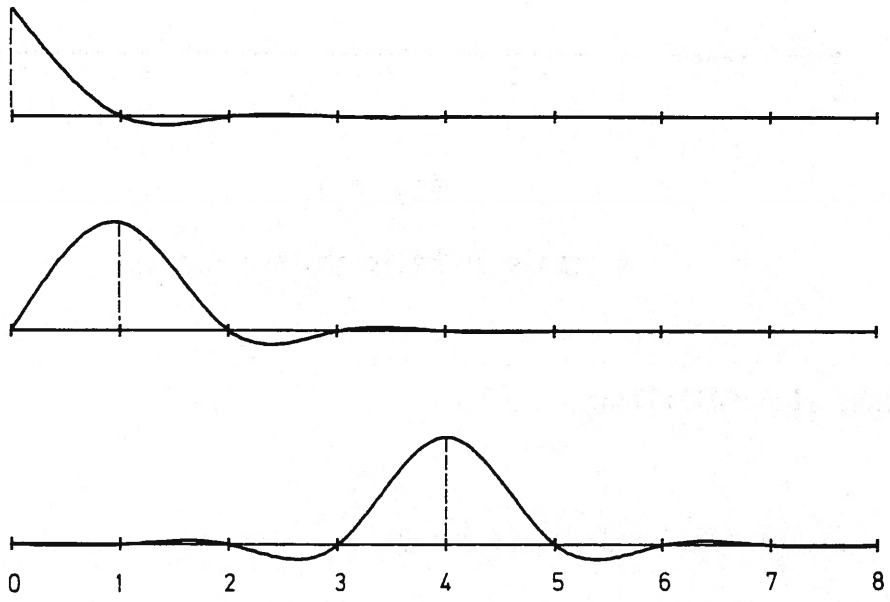


Fig. 4.4

A sample of free basis splines for n=8

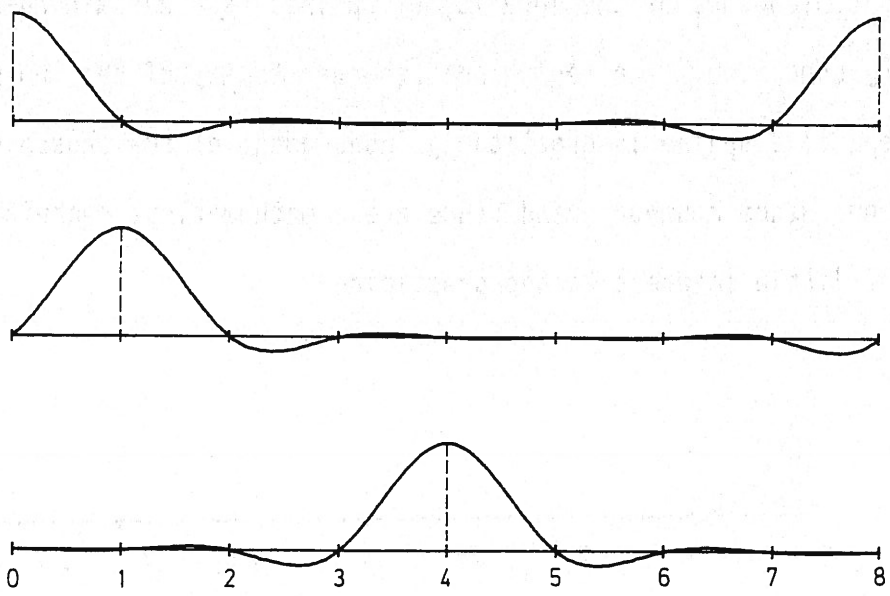


Fig. 4.5

A sample of periodic basis splines for n=8

4.9. The locality of splines.

A look at fig.'s 4.3 - 5 is very instructive. Any basis spline is appreciably different from zero only in the vicinity of the node to which it is associated. The more we go away from this node, the more the amplitudes are dampened. One can show that the dampening is exponential. The practical implication of this phenomenon is very important. The shape of the spline interpolated in a small region is only influenced by data near this region. If data are changed at a location far away the shape of the spline will not be noticeably changed anywhere else. The spline is as smooth as the data in a close vicinity imply. This is not so with polynomial interpolation. Look at fig. 4.6 showing basis functions for polynomial interpolation.

Another consequence of the locality of splines is a certain de-emphasis of the boundary condition. In a region not too near to any of the two boundary nodes the shape of a spline is practically independent of the chosen boundary condition. Hence boundary conditions are a mathematical technicality which is often of little interest to the practician.

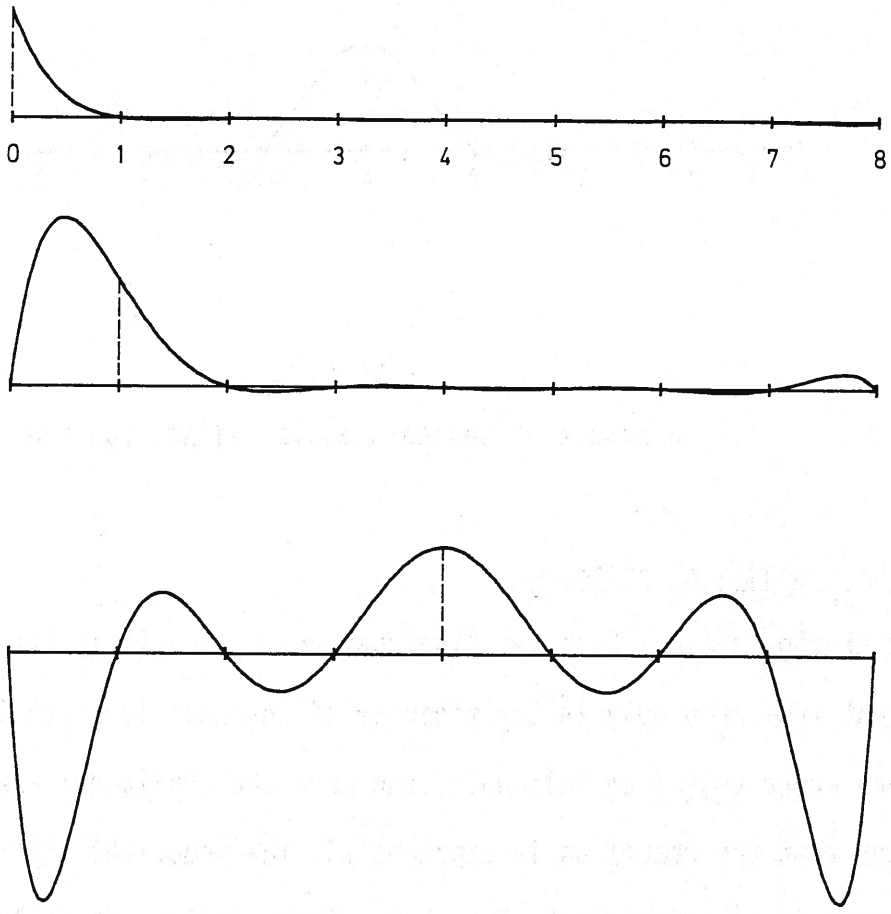


Fig. 4.6

A sample of Lagrange interpolation polynomials for n=8

References

AHLBERG, J.H.; E.N. Nilson, and J.L. Walsh (1967): The theory of splines and their applications. Academic Press, New York and London. XI + 284 pages.

SCHOENBERG, I.J. (1946): Contributions to the problem of approximation of equidistant data by analytic functions. Quart. Appl. Math., vol. 4, pp. 45-99, 112-141.

SPAETH, H. (1973): Spline-Algorithmen zur Konstruktion glatter Kurven und Flaechen. R. Oldenburg Verlag, Muenchen, Wien. 134 Seiten. (In German. Fortran algorithms for computing spline curves are included.)

5. Two-dimensional spline interpolation.

5.1. Introduction.

Assume a Cartesian coordinate system in the plane. Define a grid by means of lines

$$x = x_0, \quad x = x_1, \ \ldots, \quad x = x_m, \quad x_0 < x_1 < \ldots < x_m$$

and

$$y = y_0, \quad y = y_1, \ \ldots, \quad y = y_n, \quad y_0 < y_1 < \ldots < y_n$$

The grid covers the rectangle

$$x_0 \leq x \leq x_m$$
$$y_0 \leq y \leq y_n$$

The intersections of the grid lines, i.e. the points (x_i,y_j), $i = 1, \ldots, m$, $j = 1, \ldots, n$ are called grid points or nodes. Assume that function values

$$z_{ij} = z(x_i,y_j), \quad i = 0, \ldots, m, \quad j = 0, \ldots, n$$

are defined at the grid points. Our purpose is to interpolate these function values by means of a smooth function defined everywhere inside the area covered by the grid.

5.2. Bicubic polynomials.

Focus attention on a particular sub-rectangle

$$x_i \leq x \leq x_{i+1}, \quad y_j \leq y \leq y_{j+1}$$

Consider there a bicubic polynomial

$$p(x,y) = \sum_{k=0}^{3} \sum_{\ell=0}^{3} a_{k\ell}(x-x_i)^k (y-y_j)^\ell$$

The polynomial is currently parameterized by its 16 coefficients $a_{k\ell}$, $k,\ell = 0, \ldots, 3$. We will re-parameterize it in terms of "nodal parameters". These are the values of the function $z(x,y)$ to be interpolated together with some of its derivatives. The nodal parameters are:

$$z(x,y), \quad z_x(x,y), \quad z_y(x,y), \quad z_{xy}(x,y)$$

evaluated at the four corners of the subrectangle. We introduce the notation

$$z_{ij} = z(x_i,y_j) \qquad u_{ij} = z_x(x_i,y_j)$$
$$v_{ij} = z_y(x_i,y_j) \qquad w_{ij} = z_{xy}(x_i,y_j)$$

The 16 nodal parameters are then

$$z_{i+r,j+s}, \quad u_{i+r,j+s}, \quad v_{i+r,j+s}, \quad w_{i+r,j+s}, \quad r, s = 0, 1$$

We introduce auxiliary variables ξ, η by putting

$$x = x_i + (x_{i+1}-x_i)\xi, \qquad y = y_j + (y_{j+1}-y_j)\eta$$

Then the polynomial $p(x,y)$ transforms into

$$\pi(\xi,\eta) = \sum_{k=0}^{3} \sum_{l=0}^{3} \alpha_{kl}\, \xi^k\, \eta^l, \qquad 0 \le \xi,\ \eta \le 1$$

with

$$\alpha_{kl} = a_{kl}(x_{i+1}-x_i)^k\, (y_{j+1}-y_j)^l$$

It also follows that

$$\begin{aligned}
\zeta_{rs} &= \pi(r,s) = p(x_{i+r},y_{j+s}) = z_{i+r,j+s} \\
\varphi_{rs} &= \pi_\xi(r,s) = (x_{i+1}-x_i)\, p_x(x_{i+r},y_{j+s}) = (x_{i+1}-x_i)\, u_{i+r,j+s} \\
\psi_{rs} &= \pi_\eta(r,s) = (y_{j+1}-y_j)\, p_y(x_{i+r},y_{j+s}) = (y_{j+1}-y_j)\, v_{i+r,j+s} \\
\omega_{rs} &= \pi_{\xi\eta}(r,s) = (x_{i+1}-x_i)\,(y_{j+1}-y_j)\, p_{xy}(x_{i+r},y_{j+s}) = \\
&= (x_{i+1}-x_i)\,(y_{j+1}-y_j)\, w_{i+r,j+s}, \qquad r,\ s = 0,\ 1
\end{aligned}$$

The change of variables has transformed our problem into one for a square with unit sides: $0 \le \xi \le 1$, $0 \le \eta \le 1$. Note that

$$\pi(\xi,\eta) = [1 \quad \xi \quad \xi^2 \quad \xi^3] \begin{bmatrix} \alpha_{00} & \alpha_{01} & \alpha_{02} & \alpha_{03} \\ \alpha_{10} & \alpha_{11} & \alpha_{12} & \alpha_{13} \\ \alpha_{20} & \alpha_{21} & \alpha_{22} & \alpha_{23} \\ \alpha_{30} & \alpha_{31} & \alpha_{32} & \alpha_{33} \end{bmatrix} \cdot \begin{bmatrix} 1 \\ \eta \\ \eta^2 \\ \eta^3 \end{bmatrix}$$

Thus it holds that

$$\begin{bmatrix} s_{00} & \psi_{00} & s_{01} & \psi_{01} \\ \varphi_{00} & \omega_{00} & \varphi_{01} & \omega_{01} \\ s_{10} & \psi_{10} & s_{11} & \psi_{11} \\ \varphi_{10} & \omega_{10} & \varphi_{11} & \omega_{11} \end{bmatrix} =$$

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 \\ 0 & 1 & 2 & 3 \end{bmatrix} \cdot \begin{bmatrix} \alpha_{00} & \cdots & \alpha_{03} \\ . & & . \\ . & & . \\ \alpha_{30} & \cdots & \alpha_{33} \end{bmatrix} \cdot \begin{bmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 2 \\ 0 & 0 & 1 & 3 \end{bmatrix}$$

We write this as

$$K = H^T A H$$

It follows that

$$A = (H^T)^{-1} K H^{-1}$$

one verifies:

$$H^{-1} = \begin{bmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 2 \\ 0 & 0 & 1 & 3 \end{bmatrix}^{-1} = \begin{bmatrix} 1 & 0 & -3 & 2 \\ 0 & 1 & -2 & 1 \\ 0 & 0 & 3 & -2 \\ 0 & 0 & -1 & 1 \end{bmatrix}$$

Consequently the desired matrix of the coefficients α_{kl} is

$$\begin{bmatrix} \alpha_{00} & \alpha_{01} & \alpha_{02} & \alpha_{03} \\ \alpha_{10} & \alpha_{11} & \alpha_{12} & \alpha_{13} \\ \alpha_{20} & \alpha_{21} & \alpha_{22} & \alpha_{23} \\ \alpha_{30} & \alpha_{31} & \alpha_{32} & \alpha_{33} \end{bmatrix} =$$

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ -3 & -2 & 3 & -1 \\ 2 & 1 & -2 & 1 \end{bmatrix} \begin{bmatrix} \varsigma_{00} & \psi_{00} & \varsigma_{01} & \psi_{01} \\ \varphi_{00} & \omega_{00} & \varphi_{01} & \omega_{01} \\ \varsigma_{10} & \psi_{10} & \varsigma_{11} & \psi_{11} \\ \varphi_{10} & \omega_{10} & \varphi_{11} & \omega_{11} \end{bmatrix} \begin{bmatrix} 1 & 0 & -3 & 2 \\ 0 & 1 & -2 & 1 \\ 0 & 0 & 3 & -2 \\ 0 & 0 & -1 & 1 \end{bmatrix}$$

From the α_{kl}'s the a_{kl}'s follow by

$$a_{kl} = \alpha_{kl} \; / \; \left\{ (x_{i+1}-x_i)^k \; (y_{j+1}-y_j)^l \right\}$$

Remark: A bicubic polynomial may be parameterized by different nodal values. For example, also the values of

$$z(x,y), \quad z_{xx}(x,y), \quad z_{yy}(x,y), \quad z_{xxyy}(x,y)$$

at the four corners of a subrectangle could be used.

5.3. Hermite bicubic interpolation.

Suppose that $z_{ij} = z(x_i,y_j)$, $u_{ij} = z_x(x_i,y_j)$, $v_{ij} = z_y(x_i,y_j)$, $w_{ij} = z_{xy}(x_i,y_j)$ are available at all $(m+1)(n+1)$ grid points (i,j), $i = 0, \ldots, m$, $j = 0, \ldots, n$. Interpolate a bicubic polynomial

$$p^{ij}(x,y) = \sum_{k,l=0}^{3} a_{kl}^{ij} (x-x_i)^k (y-y_j)^l$$

in any of the $m \cdot n$ rectangles $x_i \leq x \leq x_{i+1}$, $y_j \leq y \leq y_{j+1}$, $i = 0, \ldots, m-1$, $j = 0, \ldots, n-1$. The resulting function

$$h(x,y) = p^{ij}(x,y) \quad \ldots \quad \text{for} \quad x_i \leq x \leq x_{i+1}, \quad y_j \leq y \leq y_{j+1}$$

is defined for the whole domain $x_0 \leq x \leq x_m$, $y_0 \leq y \leq y_n$. It is called a Hermite bicubic interpolation function.

Theorem: $h(x,y)$ is continuous and has continuous derivatives $h_x(x,y)$, $h_y(x,y)$ and $h_{xy}(x,y)$.

Proof: Consider $h(x,y)$ as a function of x, viewing y as a parameter. In $x_i \leq x \leq x_{i+1}$ we have

$$\lim_{y \to y_j + 0} h(x,y) = p^{ij}(x,y_j) = f_+(x), \text{ say}$$
$$\lim_{y \to y_j - 0} h(x,y) = p^{i,j-1}(x,y_j) = f_-(x), \text{ say}$$

The functions $f_+(x)$, $f_-(x)$ are one-dimensional cubic polynomials in x. We have

$$f_+(x_i) = f_-(x_i) = z_{ij}, \quad f'_+(x_i) = f'_-(x_i) = u_{ij}$$
$$f_+(x_{i+1}) = f_-(x_{i+1}) = z_{i+1,j}, \quad f'_+(x_{i+1}) = f'_-(x_{i+1}) = u_{i+1,j}$$

Because a one-dimensional cubic polynomial is uniquely determined by these values we have

$$f_+(x) \equiv f_-(x)$$

It follows that $h(x,y)$, $h_x(x,y)$ are continuous across x-grid lines. (Such grid lines are parallel to the x-axis; they are straight lines of constant $y = y_j$). Likewise it follows that $h(x,y)$, $h_y(x,y)$ are continuous across y-grid lines. Hence the continuity of $h(x,y)$ is already established.

Next we show that $h_x(x,y)$ is continuous across y-grid lines, and also that $h_y(x,y)$ is continuous across x-grid lines. Consider $h_y(x,y)$ as a function of x while y plays the role of a parameter. Call

$$\lim_{y \to y_j + 0} h_y(x,y) = p_x^{i,j}(x,y_j) = g_+(x)$$
$$\lim_{y \to y_j - 0} h_y(x,y) = p_x^{i,j-1}(x,y_j) = g_-(x)$$

The functions $g_+(x)$, $g_-(x)$ are cubic polynomials in x. We have

$$g_+(x_i) = g_-(x_i) = u_{ij}, \qquad g'_+(x_i) = g'_-(x_i) = w_{ij}$$
$$g_+(x_{i+1}) = g_-(x_{i+1}) = u_{i+1,j}, \qquad g'_+(x_{i+1}) = g'_-(x_{i+1}) = w_{i+1,j}$$

Similarly as above, one concludes that

$$g_+(x) \equiv g_-(x)$$

This proves the continuity of h_y across x-grid lines. By symmetry the continuity of h_x across y-grid lines follows too.

Finally, the continuity of $h_{xy}(x,y)$ will be shown. We have shown that $h_y(x,y)$ is continuous. For fixed y, $h_y(x,y)$ is a cubic in x for each subinterval $x_i \leq x \leq x_{i+1}$. Thus $h_{yx}(x,y)$ may be formed. At $y = y_j$, the derivatives $h_{yx}(x_i,y_j) = w_{ij}$ and $h_{yx}(x_{i+1},y_j) = w_{i+1,j}$ are prescribed. One infers that $h_{yx}(x,y)$ is continuous in the strip $x_i \leq x \leq x_{i+1}$ across all the x-grid lines. Interchanging the roles of x, y, noting that $h_{yx} = h_{xy}$, one infers that in the strip $y_j \leq y \leq y_{j+1}$, h_{xy} is continuous across the y-grid lines. Thus h_{xy} is continuous everywhere.

5.4. Bicubic splines.

5.4.1. Definition.

We want interpolating functions that are smoother than the Hermite bicubic interpolators. At least the continuity of all second derivatives will be required.

We take a one-dimensional spline A(x) with nodes $x_0, \ldots, x_m$, and we take

another one B(y) with nodes $y_0, \ldots, y_n$. We form the product

$$s(x,y) = A(x)\, B(y)$$

It is a function having the following continuous derivatives

$$s(x,y),\ s_x(x,y),\ s_y(x,y),\ s_{xx}(x,y),\ s_{xy}(x,y),\ s_{yy}(x,y)$$
$$s_{xxy}(x,y),\ s_{xyy}(x,y),\ s_{xxyy}(x,y)$$

We consider finite sets of one-dimensional splines $A_k(x)$, $k = 1, \ldots, M$ and $B_\ell(y)$, $\ell = 1, \ldots, N$. We form

$$s(x,y) = \sum_{k=1}^{M} \sum_{\ell=1}^{N} A_k(x)\, B_\ell(y)$$

The same statement about the continuity of the above specified derivatives can be made. The set of all functions obtained in this way is a vector space. We call it the space of bicubic splines for the grid $x_0, \ldots, x_m,\ y_0, \ldots, y_n$.

5.4.2. The constrained bicubic spline.

Theorem: Specify

$$z_{ij} = z(x_i, y_j),\quad i = 0, \ldots, m,\quad j = 0, \ldots, n.$$

Also specify

$$u_{0j} = z_x(x_0, y_j),\quad u_{mj} = z_x(x_m, y_j),\quad j = 0, \ldots, n$$
$$v_{i0} = z_y(x_i, y_0),\quad v_{in} = z_y(x_i, y_n),\quad i = 0, \ldots, m$$

Finally specify

$$w_{00} = z_{xy}(x_0,y_0), \quad w_{m0} = z_{xy}(x_m,y_0), \quad w_{0n} = z_{xy}(x_0,y_n),$$
$$w_{mn} = z_{xy}(x_m,y_n)$$

It is asserted that a unique bicubic spline matching these values exists.

Proof: We first prove the existence of the spline. We take the basis $a_i(x)$, $i = 0, \ldots, m$, $\alpha_0(x)$, $\alpha_m(x)$ for the one-dimensional splines A(x). This basis was constructed in section 4.8. We take a similar basis $b_j(y)$, $j = 0, \ldots, n$, $\beta_0(y)$, $\beta_n(y)$ for the splines B(y).

We consider the bicubic spline:

$$\begin{aligned} s(x,y) = & \sum_{i=0}^{m} \sum_{j=0}^{n} z_{ij}\, a_i(x)\, b_j(y) + \\ & + \sum_{j=0}^{n} [u_{0j}\, \alpha_0(x)\, b_j(y) + u_{mj}\, \alpha_m(x)\, b_j(y)] + \\ & + \sum_{i=0}^{m} [v_{i0}\, a_i(x)\, \beta_0(y) + v_{in}\, a_i(x)\, \beta_n(y)] + \\ & + w_{00}\, \alpha_0(x)\, \beta_0(y) + w_{m0}\, \alpha_m(x)\, \beta_0(y) \\ & + w_{0n}\, \alpha_0(x)\, \beta_n(y) + w_{mn}\, \alpha_m(x)\, \beta_n(y) \end{aligned}$$

The bicubic spline s(x,y) completely solves the stated interpolation problem. This proves existence.

We now prove uniqueness. We show that no other function s(x,y) exists which has the following properties.

(1) $s(x,y)$ is a bicubic polynomial in each subrectangle.

(2) $s(x,y)$ has continuous derivatives s, s_x, s_y, ..., s_{xxyy} as specified earlier.

(3) $s(x,y)$ interpolates the data specified in the theorem.

It suffices to show that a function satisfying (1) and (2), having vanishing function values at the nodes and having vanishing boundary conditions, is necessarily the zero function. (In other words, a function satisfying (1) and (2), and interpolating data that are all zero, must be the zero function). Such a function is a one-dimensional bicubic spline along any grid line. Along a grid line, e.g. that one for $y = y_{j_0}$, we have with $f(x) = s(x,y_{j_0})$: $f(x_i) = z_{i,j_0} = 0$ and $f'(x_0) = u_{0,j_0} = f'(x_m) = u_{m,j_0} = 0$. Hence this one-dimensional spline is the zero spline. It follows that all u_{ij} vanish. Likewise one concludes that all v_{ij} are zero. The function $g(x) = s_y(x,y_0)$ is a cubic in each subinterval $x_i \leq x \leq x_{i+1}$. It must be a spline, because the derivatives s_{yx} and s_{yxx} are required to be continuous. One has $g(x_i) = v_{i0} = 0$. Also $g'(x_0) = w_{00} = g'(x_m) = = w_{m0} = 0$. Thus $g(x) = 0$, and consequently $w_{i0} = 0$, $i = 0, \ldots, m$. Similarly one infers that $w_{in} = 0$, $i = 0, \ldots, m$. Now the splines $h_j(x) = s_y(x,y_j)$ can be interpolated. We have $h_j(x_i) = v_{ij_0} = 0$, $h'_j(x_0) = w_{0j} = h'_j(x_m) = w_{mj} = 0$. Thus $h_j(x) = 0$ and $w_{ij} = 0$ for all i, j. A Hermite bicubic function having $z_{ij} = u_{ij} = v_{ij} = w_{ij} = 0$ must be the zero function. This concludes the proof.

During our uniqueness proof we have in effect proved more than we originally wanted. Without any further argument we can state the following theorem.

Theorem: The space of bicubic splines $s(x,y)$ coincides with the space of Hermite bicubic interpolation functions upon which the requirement is imposed that in addition to $s(x,y)$, $s_x(x,y)$, $s_y(x,y)$, $s_{xy}(x,y)$, also the derivatives $s_{xx}(x,y)$, $s_{xy}(x,y)$, $s_{yy}(x,y)$, $s_{xxy}(x,y)$, $s_{xyy}(x,y)$ and $s_{xxyy}(x,y)$ are continuous. The space of bicubic splines over the grid $x_0, \ldots, x_m, y_0, \ldots, y_n$ has dimension $(m+3)(n+3)$. A basis has been exhibited above. It consists of all products $A(x)\,B(y)$ with

$$A(x) \in \{a_0(x), \ldots, a_m(x), \alpha_0(x), \alpha_m(x)\}$$

and

$$B(y) \in \{b_0(y), \ldots, b_n(y), \beta_0(y), \beta_n(y)\}$$

Thus it consists of all products of basis functions for the one-dimensional splines over $x_0, \ldots, x_m$ and $y_0, \ldots, y_n$.

5.4.3. The free bicubic spline.

It has boundary conditions

$$z_{xx}(x_0,y_j) = z_{xx}(x_m,y_j) = 0, \quad j = 0, \ldots, n$$
$$z_{yy}(x_i,y_0) = z_{yy}(x_i,y_n) = 0, \quad i = 0, \ldots, n$$
$$z_{xxyy}(x_0,y_0) = z_{xxyy}(x_m,y_0) = z_{xxyy}(x_0,y_n) = z_{xxyy}(x_m,y_n) = 0$$

A basis is readily obtained by using bases for the one-dimensional free splines

and performing the (tensor) products, as it was done in the case of the constrained spline.

5.4.4. The double periodic spline.

The requirements are that s(x,y) is periodic in x as well as in y. A basis can be specified in an obvious way.

6. Geometry of exact spline interpolation.

6.1. Formulation of the problem.

Let V, W be Hilbert spaces. We consider two linear and continuous mappings.

(1) The operator Λ maps V onto R^n. In case of finite-dimensional V, we let Λ be represented by the matrix A. This matrix has more columns than rows, i.e. V is of higher dimension than n, the dimension of R^n.

(2) The operator Θ maps V into W. The inner product in W shall be represented by the matrix Q. In case of finite-dimensional V, W, we let the matrix B represent the operator Θ. B is then also a matrix having more columns than rows, i.e. V is of higher dimension than W.

We require that the nullspaces of Λ and Θ have only the zero vector in common

$$N(\Lambda) \cap N(\Theta) = 0$$

and pose the following problem.

Given $y \in R^n$, find $x \in V$ such that

(a) $\Lambda(x) = y$ ... the interpolation requirement

(b) $\|\Theta(x)\|_W$ = minimum ... the minimum norm requirement.

Example 1: Let V be the Hilbert space of functions $F(\xi)$, $\alpha \leq \xi \leq \beta$ with inner

product

$$(f_1, f_2) = \int_\alpha^\beta \left\{ f_1(\xi)\, f_2(\xi) + f_1'(\xi)\, f_2'(\xi) + f_1''(\xi)\, f_2''(\xi) \right\} d\xi$$

Let W be the Hilbert space of functions $g(\xi)$, $\alpha \leq \xi \leq \beta$ with inner products

$$(g_1, g_2) = \int_\alpha^\beta g_1(\xi)\, g_2(\xi)\, d\xi$$

Let $\xi_1, \ldots, \xi_n$ be distinct points in the interval $[\alpha,\beta]$. Let Λ be the operator mapping $f \in V$ onto the vector

$$\Lambda(f) = \begin{bmatrix} f(\xi_1) \\ \vdots \\ f(\xi_n) \end{bmatrix} \in R^n$$

Let θ be the operator mapping $f \in V$ onto $f'' \in W$. Thus θ maps a function f onto its second derivative

$$\theta(f) = f'' = \frac{d^2}{d\xi^2} f$$

We then pose the problem

Given

$$y = \begin{bmatrix} \eta_1 \\ \vdots \\ \eta_n \end{bmatrix} \in R^n$$

find a function $f \in V$ such that

(a) $f(\xi_i) = \eta_i$, $i = 1, \ldots, n$, i.e. f interpolates the prescribed function values η_i at the locations ξ_i

(b) $\|f''\|^2 = \int_\alpha^\beta (f''(x))^2 dx$ = minimum, i.e. f is in a sense the smoothest interpolating function.

Example 2: Instead of a continuous interval of arguments ξ, $\alpha \leq \xi \leq \beta$, we consider a large, but discrete set of equidistant values

$$\alpha = \xi_0 < \xi_1 < \ldots < \xi_m = \beta$$

We denote the step size

$$h = \xi_{i+1} - \xi_i, \quad i = 0, \ldots, m-1$$

We consider a corresponding set of discrete function values

$$f_i = f(\xi_i), \quad i = 0, \ldots, m$$

Out of the set $\{\xi_i\}$ we select a subset of n elements $(\xi_{i_1}, \ldots, \xi_{i_n})$. The $\xi_{i_1}, \ldots, \xi_{i_n}$ need not be equally spaced. At these selected locations we prescribe function values $\eta_1, \ldots, \eta_n$.

We pose the problem:

Given $\eta_1, \ldots, \eta_n$ find a vector $(f_0, f_1, \ldots, f_m)$ such that

(a) $f_{i_j} = \eta_j, \quad j = 1, \ldots, n \quad \ldots$ interpolation requirement

(b) $\frac{1}{h^2} \sum_{i=1}^{m-1} (-f_{i+1} + 2f_i - f_{i-1})^2 = \text{minimum} \quad \ldots$ smoothness requirement.

Obviously this is a discrete version of the problem in example 1. The operator Λ maps $f = (f_0, \ldots, f_m)$ onto the subset of components $\eta_1 = f_{i_1}, \ldots, \eta_n = f_{i_n}$. The operator θ maps the vector f onto the set of second difference quotients

$$\Delta^2 f_i = \frac{1}{h^2} (-f_{i+1} + 2f_i - f_{i-1}) = \frac{\frac{1}{h}(f_{i+1} - f_i) - \frac{1}{h}(f_i - f_{i-1})}{h}$$

6.2. Definition of splines.

In V we consider the nullspace $N = N(\Lambda)$ of the operator Λ. Referring to example 1, N_Λ consists of functions vanishing at the locations of interpolation ξ_i, $i = 1, \ldots, n$. The operator θ maps functions $f \in N(\Lambda)$ onto functions g of a certain subspace $U = U(\Lambda, \theta)$ of W:

$$U = U(\Lambda, \theta) = \{z \in W \mid \text{there is } x \in N(\Lambda) \text{ such that } \theta(x) = z\}$$

Referring to example 1, U is the set of second derivatives of functions vanishing at the locations of interpolation.

We consider the orthocomplement $U^\perp$ of U in W. We consider the pre-images

of elements in $U^{\perp}$ under θ. This is a subspace $S = S(\Lambda,\theta)$ of V.

$$S = S(\Lambda,\theta) = \{x \in V \mid \theta(x) \in U^{\perp}\}$$

S is called the subspace of splines in V. Its elements are called splines. Later it will be shown that they are the solutions of the extremum problem formulated in section 6.1.

Example 1a: (Continuation of example 1)
As we have noticed, $N = N(\Lambda)$ is the set of functions vanishing at the locations of interpolation. $U = U(\Lambda,\theta)$ is the set of second derivatives of such functions. What is the set $U^{\perp}$? What is the set $S = S(\Lambda,\theta)$?

Theorem: For the spaces and operators of example 1 the set S is the set of piecewise cubic polynomials $p(\xi)$ on $\alpha \leq \xi \leq \beta$ having the following properties

$p(\xi)$ is linear in $\alpha \leq \xi \leq \xi_1$, $\xi_n \leq \xi \leq \beta$
$p(\xi)$ is cubic in any of the intervals $\xi_i \leq \xi \leq \xi_{i+1}$, $i = 1, \ldots, n-1$
$p(\xi)$ is continuous together with its first and second derivatives.

Such functions are called cubic splines.

Remark: An equivalent formulation of the above theorem would be: The set $U^{\perp}$ consists of piecewise linear functions vanishing in $\alpha \leq \xi \leq \xi_1$, $\xi_n \leq \xi \leq \beta$, and whose slope is constant in any of the intervals $\xi_i \leq \xi \leq \xi_{i+1}$, $i = 1, \ldots, n-1$.

Proof: It suffices to prove the version given in the preceding remark. We decompose the proof into two steps, showing

(1) Any piecewise linear function $q(\xi)$ with possible discontinuities of $q'(\xi)$ only at ξ_i, $i = 1, \ldots, n$ and vanishing in $\alpha \leqslant \xi \leqslant \xi_1$, $\xi_n \leqslant \xi \leqslant \beta$, is orthogonal to any second derivative $f''(\xi)$ of a function vanishing at ξ_i, $i = 1, \ldots, n$.

To show this compute $(f'',q)_W =$

$$\int_\alpha^\beta f''(\xi)\, q(\xi)\, d\xi = \int_\alpha^{\xi_1} \ldots + \sum_{i=1}^{n-1} \int_{\xi_i}^{\xi_{i+1}} \ldots + \int_{\xi_n}^{\beta} \ldots$$

Apply partial integration twice, obtaining

$$\begin{aligned} & f'(\beta)\, q(\beta-0) - f'(\alpha)\, q(\alpha+0) - f(\beta)\, q'(\beta-0) + f(\alpha)\, q'(\alpha+0) \\ & - \sum_{i=1}^{n-1} \left\{ f(\xi_{i+1})\, q'(\xi_{i+1}-0) - f(\xi_i)\, q'(\xi_i+0) \right\} + \int_\alpha^{\xi_1} f(\xi)\, q''(\xi)\, d\xi \\ & + \sum_{i=1}^{n-1} \int_{\xi_i}^{\xi_{i+1}} f(\xi)\, q''(\xi)\, d\xi + \int_{\xi_n}^{\beta} f(\xi)\, q''(\xi)\, d\xi \end{aligned}$$

All terms vanish because of the properties of $f(\xi)$ and $q(\xi)$.

(2) Let $q(\xi)$ have properties required in (1). Such functions form a linear subspace in W. We show: Any function $g \in W$ orthogonal to this subspace can be viewed as the second derivative $g(\xi) = f''(\xi)$ of an $f \in V$, vanishing

at ξ_i, $i = 1, \ldots, n$. Put

$$f(\xi) = \int_{\alpha}^{\xi} d\eta \int_{\alpha}^{\eta} g(\theta)\, d\theta + c\,x + d$$

Then

$$g(\xi) = f''(\xi)$$

Adjust the constants of integration c, d such that

$$f(\xi_1) = f(\xi_2) = 0.$$

From

$$\int_{\alpha}^{\beta} g(\xi)\, q(\xi)\, d\xi = \int_{\alpha}^{\beta} f''(\xi)\, q(\xi)\, d\xi = 0$$

conclude by two fold partial integration (cf. the above formula under (1)) that:

$$\sum_{i=1}^{n-1} [f(\xi_{i+1})\, q'(\xi_{i+1}-0) - f(\xi_i)\, q'(\xi_i+0)] = 0$$

Choose piecewise linear functions $q_i(\xi)$, $i = 2, \ldots, n-1$ such that

$$q'_i(\xi_i-0) = \frac{1}{\xi_i - \xi_{i-1}} = \varkappa_i, \text{ say}$$

$$q'_i(\xi_i+0) = -\frac{1}{\xi_{i+1} - \xi_i} = \lambda_i, \text{ say}$$

Figure 6.1 shows such a function. The derivatives of $q_i(\xi)$ outside of

$\xi_{i-1} \leq x \leq \xi_{i+1}$ are required to vanish.

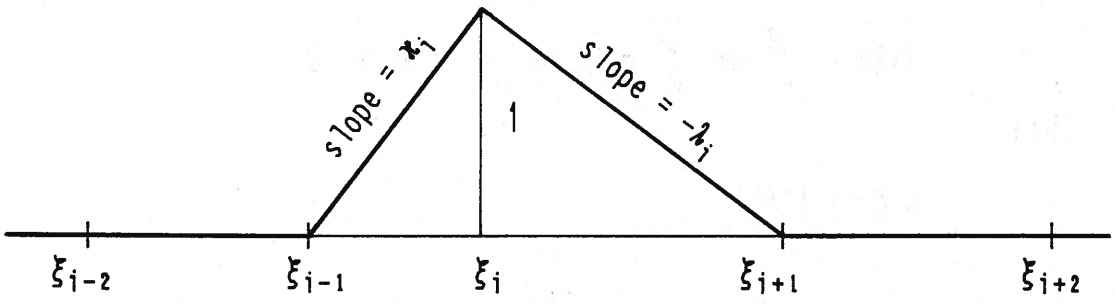


Fig. 6.1. The function $q_i(\xi)$

Inserting these functions into the above equation gives a linear system

$$\begin{bmatrix} -\lambda_2 & & & \\ (\alpha_3+\lambda_3), & -\lambda_3 & & \\ -\alpha_4, & (\alpha_4+\lambda_4), & -\lambda_4 & \\ & \cdot & \cdot & \cdot \\ & & -\alpha_{n-1}, & (\alpha_{n-1}+\lambda_{n-1}), & -\lambda_{n-1} \end{bmatrix} \cdot \begin{bmatrix} f(\xi_3) \\ f(\xi_4) \\ \cdot \\ \cdot \\ f(\xi_n) \end{bmatrix} = 0$$

This system is regular and homogeneous. Hence the solution must be zero. Q.e.d.

6.3. Existence and uniqueness of splines.

Recall the decomposition of the space W into U and $U^{\perp}$. $U = U(\Lambda,\theta)$ contains the images of $N = N(\Lambda)$ under the operator θ. Recall that the space of splines $S = S(\Lambda,\theta)$ was defined as the set of pre-images of vectors in $U^{\perp}$.

Theorem: The space V is the direct sum of $N(\Lambda)$ and $S(\Lambda,\theta)$.

Remark: This means that any vector $x \in V$ is uniquely represented as $x = x_N + x_S$, $x_N \in N$, $x_S \in S$. Confer section A.5.1.

Proof: We start by noting that any pre-image x of a vector $y \in U$ may be uniquely decomposed as $x = x_a + x_b$, $x_a \in N(\Lambda)$, $x_b \in N(\theta)$. [If $\theta(x) = y$ with $y \in U$, then y is also the image of some vector x_a in $N(\Lambda)$: $y = \theta(x_a)$. Put $x_b = x - x_a$, then $\theta(x_b) = \theta(x) - \theta(x_a) = y - y = 0$. Thus $x_b \in N(\theta)$]. Next we show that any $x \in V$ may be decomposed as

$$x = x_N + x_S, \quad x_N \in N, \quad x_S \in S$$

Let $y = \theta(x)$. Split $y = y_1 + y_2$, $y_1 \in U$, $y_2 \in U^{\perp}$. Let x_1, x_2 be pre-images of y_1, y_2, respectively. Put $x_3 = x - x_1 - x_2$. Then $\theta(x_3) = y - y_1 - y_2 = 0$. Thus $x = x_1 + x_2 + x_3$ with x_1 being a pre-image of a vector in U, x_2 being in S, and x_3 in $N(\theta)$. As shown earlier, we may split x_1 into x_4 and x_5, $x_4 \in N(\Lambda)$, $x_5 \in N(\theta)$. Vectors in $N(\theta)$ have zero images. They may therefore be viewed as vectors of S. Calling $x_N = x_4$, $x_S = x_2 + x_3 + x_5$, we get the desired decomposition.

Finally, we show that only the zero vector is common to N and S. If $x \in N \cap S$, then $y = \theta(x)$ is in U as well as in $U^{\perp}$; hence $y = 0$. Thus $x \in N(\theta)$. However it

was postulated in section 6.1 that the only vector common to $N = N(\Lambda)$ and $N(\theta)$ is the zero vector.

We have shown that any vector $x \in V$ may be uniquely decomposed as $x_N + x_S$. This proves existence and uniqueness of the spline x_S. The original vector (function) x and x_S interpolate the same data y. For $\Lambda(x) = \Lambda(x_N + x_S) = \Lambda(x_S)$.

6.4. Minimum properties of splines.

Theorem: Let $x \in V$ and let x_S be its spline. Then x_S solves the following two extremum problems:

$$(I) \quad \min_{z \in S} \|\theta(x) - \theta(z)\|_W = \|\theta(x) - \theta(x_S)\|$$

$$(II) \quad \min_{z \in V,\, \Lambda(z) = \Lambda(x)} \|\theta(z)\|_W = \|\theta(x_S)\|$$

Remark: Problem (II) was the problem stated in section 6.1. It shows that splines have "the minimum norm property". Among all functions interpolating the same data as x does, $\theta(x_S)$ has the smallest norm in W. Problem (I) is complementary to problem (II). Confer chapter A.11 for a general discussion of complementary least squares problems. Problem (I) shows that x_S has the "best approximation property". Out of the space of splines, the spline x_S is closest to a given function x.

Proof: Given $x \in V$, decompose $x = x_N + x_S$. In case of problem (I) we let

z vary over the spline space S. Then

$$x - z = x_N + (x_S - z), \quad x_N \in N, \quad x_S - z \in S$$

Thus

$$\|\theta(x-z)\|^2 = \|\theta(x_N)\|^2 + \|\theta(x_S-z)\|^2$$

This is minimal for $z = x_S$.

In case of problem (II) we let z vary over the set $\{z \mid \Lambda(z) = \Lambda(x)\}$. Thus $z-x \in N$ or $z = x + u$, $u \in N$. Hence

$$z = x_N + u + x_S, \quad x_N + u \in N$$
$$\|\theta(z)\|^2 = \|\theta(x_N+u)\|^2 + \|\theta(x_S)\|^2$$

The minimum is obtained for $u = -x_N$. This proves (II).

6.5. Other examples.

The examples 1 and 1a treated earlier refer to the case of the free spline with nodes $\xi_1, \ldots, \xi_n$. It has been shown that the free spline interpolating $f(\xi)$ at $\xi_1, \ldots, \xi_n$ has certain minimal properties (I) and (II). Similar properties hold for the constrained spline and for the periodic spline. It is then preferable to put $\xi_1 = \alpha$ and $\xi_n = \beta$. Among all functions $s(\xi)$ interpolating $f(\xi)$ at $\xi_1, \ldots, \xi_n$, and having the same derivatives as $f(\xi)$ at the boundary nodes ξ_1

and ξ_n, the constrained spline minimizes $\|s''(\xi)\|$. A similar statement holds for the periodic spline.

Also a generalization to the bicubic splines is possible. It is interesting to note that in the two-dimensional case the operator θ is given by

$$\theta(f) = f_{\xi\xi\eta\eta}$$

Thus the squared norm in W is given by

$$\|f\|_W^2 = \int_{\xi_1}^{\xi_n} \int_{\eta_1}^{\eta_n} f_{\xi\xi\eta\eta}(\xi,\eta)^2 \, d\xi \, d\eta$$

6.6. Prediction as a special case of spline interpolation.

Let

$$V = R^{n+m}$$

A vector $x \in V$ is represented as

$$x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \qquad x_1 \in R^n, \quad x_2 \in R^m$$

Let Λ be implied by

$$\ell = (I, 0)\, x = x_1$$

Let the space W coincide with V and let the norm in $V = W$ be implied by

$$P = \begin{bmatrix} P_{11} & P_{12} \\ P_{21} & P_{22} \end{bmatrix}$$

We denote, as usual

$$Q = P^{-1} = \begin{bmatrix} Q_{11} & Q_{12} \\ Q_{21} & Q_{22} \end{bmatrix}$$

Let the operator θ be the identity:

$$\theta = I$$

Consider problem (II): Given $\ell \in R^n$, find $x \in V$ such that

$x_1 = \ell$ ... interpolation requirement

$\|x\|$ = minimum ... minimum norm requirement

This reduces to the problem

$$x^T P x = \text{minimum}$$

subject to

$$(I, 0)\, x = \ell$$

The problem is formally equivalent to a conditioned adjustment problem with zero observations, residuals x and discrepancies $-\ell$; i.e.

$$(0 + x)^T P (0 + x) = \text{minimum}$$

subject to

$$(I \;\; 0)(0 + x) = \ell$$

Thus the solution is

$$x = P^{-1} \begin{bmatrix} I \\ 0 \end{bmatrix} k, \quad \text{i.e.} \quad x = \begin{bmatrix} Q_{11} \\ Q_{21} \end{bmatrix} k$$

where the correlates k follow from the normal equations

$$(I \;\; 0)\, P^{-1} \begin{bmatrix} I \\ 0 \end{bmatrix} k - \ell = 0$$

i.e.

$$Q_{11}\, k = \ell, \quad k = Q_{11}^{-1}\, \ell$$

Thus

$$x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} Q_{11} \\ Q_{21} \end{bmatrix} Q_{11}^{-1} \ell = \begin{bmatrix} \ell \\ Q_{21}Q_{11}^{-1}\ell \end{bmatrix}$$

The formula

$$x_2 = Q_{21}Q_{11}^{-1}\ell = Q_{21}Q_{11}^{-1}x_1$$

is the familiar prediction formula. (Usually written $s_2 = \Sigma_{21}\Sigma_{11}^{-1}s_1$).

Replacing the solution x by s, which stands for "spline", we know that the set of all splines is obtained by letting ℓ vary all over R^n.

Any spline $s \in S$ is represented as

$$s = \begin{bmatrix} I \\ Q_{21}Q_{11}^{-1} \end{bmatrix} \ell \qquad \text{for some } \ell \in R^n.$$

We know that interpolating splines are also the solution of problem (I):

Given $x \in V$, find s such that

$$\|x - s\| = \text{minimum}$$

subject to

$$s \in S$$

Let us verify this directly. The problem is restated as follows

$$(x - s)^T P (x - s) = \text{minimum}$$

subject to

$$s = \begin{bmatrix} I \\ Q_{21}Q_{11}^{-1} \end{bmatrix} \ell$$

This is formally identical to an adjustment problem by parameters with observations x, adjusted observations s and parameters ℓ. The solution is obtained via the normal equations

$$(I \quad Q_{11}^{-1}Q_{12}) \begin{bmatrix} P_{11} & P_{12} \\ P_{21} & P_{22} \end{bmatrix} \cdot \begin{bmatrix} I \\ Q_{21}Q_{11}^{-1} \end{bmatrix} \ell = (I \quad Q_{11}^{-1}Q_{12}) \begin{bmatrix} P_{11} & P_{12} \\ P_{21} & P_{22} \end{bmatrix} \cdot \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

Rewriting this as

$$Q_{11}^{-1}(Q_{11} \quad Q_{12})Q^{-1} \begin{bmatrix} Q_{11} \\ Q_{12} \end{bmatrix} Q_{11}^{-1}\ell = Q_{11}^{-1}(Q_{11} \quad Q_{12})Q^{-1}x$$

or

$$Q_{11}^{-1}Q_{11}Q_{11}^{-1}\ell = Q_{11}^{-1}x_1$$

i.e.

$$\ell = x_1$$

which was to be shown!

6.7. Noise-free collocation with trend parameters.

Consider the problem

Given $\ell \in R^n$, find $x \in R^m$, $y \in R^p$ such that

$$\ell = A x + B y \quad \text{and} \quad y^T P y = \text{minimum}$$

The vector ℓ is called observation vector. (Its coordinates are linear functionals). The vector $x \in R^m$ comprises the trend parameters. The vector $B y$ is sometimes called "the signal".

We put

$$V = R^{m+p}$$

such that for $z \in V$ we have

$$z = \begin{bmatrix} x \\ y \end{bmatrix}$$

The operator Λ is given by

$$\ell = \Lambda(z) = (A \quad B) \begin{bmatrix} x \\ y \end{bmatrix}$$

The space R^p of signals is taken as the space W. Its inner product is given by the matrix P.

The operator θ is taken as

$$w = \theta(z) = (0, \ I)\ z \ = \ y$$

Our above stated extremum problem is precisely problem (II) of section 6.4. and section 6.1.

Given $\ell \in R^n$, find $z \in V$ such that

$$\|\theta(z)\|_W = \text{minimum}$$

subject to

$$\Lambda(z) = \ell$$

The special problem of this section is also formally equivalent to the socalled "general adjustment problem". Let us solve the problem in the familiar way by means of Lagrange multipliers. Lagrange's function is

$$\Phi(x,y) = y^T P\, y \ - \ 2\, \lambda^T (A\, x \ + \ B\, y \ - \ \ell)$$

Thus

$$\frac{1}{2}\frac{\partial \Phi}{\partial y} = P\, y \ - \ B^T \lambda \ = \ 0$$

$$\frac{1}{2}\frac{\partial \Phi}{\partial x} = A^T \lambda \qquad = \ 0$$

The first of these equations gives

$$y \ = \ P^{-1} B^T \lambda$$

Inserting into $A\,x + B\,y = \ell$ we get the system

$$\begin{aligned} B\,P^{-1}B^T\lambda + A\,x &= \ell \\ A^T\lambda &= 0 \end{aligned}$$

Solving for x by partial reduction we find the system and its solution

$$A^T(B\,P^{-1}B^T)^{-1}A\,x = A^T(B\,P^{-1}B^T)^{-1}\ell$$

$$\boxed{x = (A^T(B\,P^{-1}B^T)^{-1}A)^{-1}A^T(B\,P^{-1}B^T)^{-1}\ell}$$

From the first equation one finds

$$\lambda = (B\,P^{-1}B^T)^{-1}(\ell - Ax)$$

From $y = P^{-1}B^T\lambda$ one finds

$$\boxed{y = P^{-1}B^T(B\,P^{-1}B^T)^{-1}(\ell - Ax)}$$

These are the formulas for noise free collocation with trend parameters x. Note that the usual notation is

$$s = B\,y$$

$$\Sigma_{ss} = B\,P^{-1}B^T \qquad \Sigma_{ys} = P^{-1}B^T$$

We prefer to introduce the notation

$$s = \begin{bmatrix} u \\ v \end{bmatrix}$$

for the solution of our problem. Given $\ell \in R^n$, we find u, v from the equations

$$\begin{aligned} B\,P^{-1}B^T\lambda + A\,u &= \ell \\ A^T\lambda &= 0 \\ v &= P^{-1}B^T\lambda \end{aligned}$$

Letting ℓ run through R^n, we obtain the set S of splines s. Let us check, whether also our abstract approach arrives at this set.

The null-space of Λ is given by the solutions of

$$A\,x + B\,y = 0$$

The space $U^\perp$ consists of all vectors w which are orthogonal to all y's which fulfill $A\,x + B\,y = 0$ together with a suitable x. Thus

$$A\,x + B\,y = 0$$

must imply

$$w^T P\,y = 0$$

It follows that the row vector

$$(0, \quad w^T P)$$

must be a linear combination of the rows of $(A, \quad B)$:

$$(0, \quad w^T P) \quad = \quad \lambda^T (A, \quad B)$$

or

$$A^T \lambda \quad = \quad 0$$

$$w \quad = \quad P^{-1} B^T \lambda$$

The pre-images of such w's are the splines. Thus the set S of splines is given by

$$s = \begin{bmatrix} u \\ P^{-1} B^T \lambda \end{bmatrix} \qquad \text{with u arbitrary and} \quad A^T \lambda = 0$$

Let us solve the problem (I) of section 6.4:

Given

$$z \quad = \quad \begin{bmatrix} x \\ y \end{bmatrix} \quad \varepsilon \quad V$$

find $s \varepsilon S$ such that

$$\| z - s \|_w \quad = \quad \text{minimum}$$

Using the above representation for s, as well as the definition of the inner product in W, we get

$$(y - P^{-1}B^T\lambda)^T P (y - P^{-1}B^T\lambda) = \text{minimum}$$

subject to

$$A^T\lambda = 0$$

This problem is formally equivalent to an adjustment problem by parameters which fulfill additional conditions. (Observations ... y, parameters ... λ, adjusted observations $P^{-1}B^T\lambda$, conditions $A^T\lambda = 0$).

Introducing the Lagrangean

$$\Phi(\lambda,\mu) = (y - P^{-1}B^T\lambda)^T P (y - P^{-1}B^T\lambda) + 2\,\mu^T A^T\lambda$$

we find

$$B\,P^{-1}B^T\lambda + A\,\mu = B\,y$$
$$A^T\lambda = 0$$

If we identify $\ell = B\,y$ and $\mu = x$, we arrive at the earlier equations. This concludes the successful verification of the equivalence of problem (I) and problem (II).

7. Approximation with splines.

7.1. Introduction.

Frequently we want that our data are not exactly interpolated but only approximated. Our spline functions shall not precisely match the data; there will be residuals or discrepancies. Data may be distributed irregularly. In the 1-dimensional case we can design spline functions matching irregular data exactly. However, even in the 1-dimensional case this is not always desirable. The data may be noisy, and we want that our approximating functions filter out some of the noise. In the two- or higher dimensional case it is virtually impossible to interpolate irregular data by bi- or n-cubic splines.

7.2. Approximation in one dimension.

Let (ξ_k,η_k), k=1,...,N denote the data. This means that at the locations ξ_k, function values $\eta_k = f(\xi_k)$ are prescribed. Let x_i, i=0,...,n denote the nodes of the spline s(x), i.e. the location of the discontinuities of its third derivative. The x_i generally do not coincide with the ξ_k. Occasional coincidence is, however, not excluded. Also the number of nodes n is typically less than N, the number of data. Let y_i denote the ordinates of the spline s(x) at the locations x_i. The y_i are now unknown. The approximating spline is parameterized in terms of y_i and y'_i, i=0,...,n. Thus we have 2n+2 unknowns. The equations of section 4.3, enforcing continuity of the second derivatives at x_i, i=1,...,n-1, represent a set of n-1 constraints among x_i, y_i. Additional boundary constraints at x_0 and x_n may or may not be prescribed.

The discrepancy $v_i = -[\eta_i - s(\xi_i)]$ may be linearly expressed in terms of y_i, y'_i. This can be done by means of the formula in section 4.2 giving the coefficients $a^{(i,i+1)}$ as linear functions of y_i, y'_i. Then one has:

$$v_k = -[\eta_k - \sum_{l=0}^{3} a_l^{(i,i+1)} \xi_k^l]$$

The size of the discrepancies can be measured by their squared norm

$$\|v\|^2 = \sum_{k=1}^{N} v_k^2$$

This expression can be minimized subject to the formulated constraints. We obtain a mixed adjustment model. It is the model of variation of parameters with additional constraints. It is a feasible way to obtain an approximating spline. A better way is outlined in the next section.

7.3. Basis splines with local support.

This remarkable type of spline is already described in Schoenberg (1946). We assume equidistant data with

$$x_i = i, \quad i=0,\ldots,n$$

We start with the function B(x) defined as follows:

$$B(x) = \begin{cases} 0 & \ldots \quad x \leq -2 \\ \frac{1}{6}(x+2)^3 & \ldots \quad -2 \leq x \leq -1 \\ \frac{1}{6}(x+2)^3 - \frac{4}{6}(x+1)^3 & \ldots \quad -1 \leq x \leq 0 \\ \frac{1}{6}(-x+2)^3 - \frac{4}{6}(-x+1)^3 & \ldots \quad 0 \leq x \leq 1 \\ \frac{1}{6}(-x+2)^3 & \ldots \quad 1 \leq x \leq 2 \\ 0 & \ldots \quad 2 \leq x \end{cases}$$

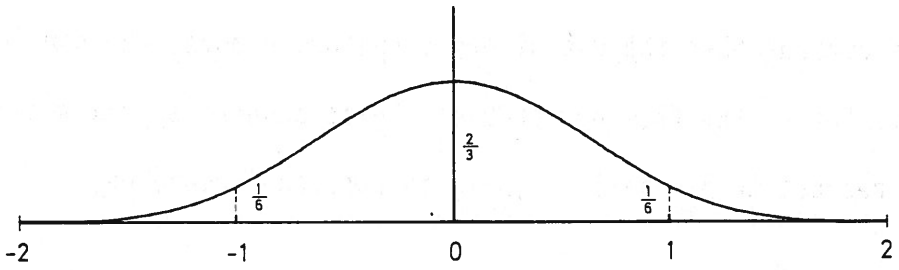


Fig. 7.1 Basis spline B(x)

The function B(x) is shown in figure 7.1. It is a spline. Its second derivative is continuous. The decisive property of B(x) is its local support. It vanishes outside a finite interval, namely the interval $-2 \leq x \leq 2$.

We now add two auxiliary nodes of $x_{-1} = -1$ and $x_{n+1} = n+1$. We associate with any node x_i, $i=-1,0,\ldots,n,n+1$, a basis spline $B_i(x)$ by shifting the function B(x):

$$B_i(x) = B(x-x_i) = B(x-i), \qquad i=-1,\ldots,n+1$$

We consider a linear combination of the basis splines

$$s(x) = \sum_{i=-1}^{n+1} \beta_i B_i(x)$$

Obviously s(x) is a spline. It is parameterized in terms of its coordinates β_i with respect to the basis built up of the B_i's. (Splines with nodes x_i, $i=0,\ldots,n$, form an n+3-dimensional vector space. The B_i, $i=-1,\ldots,n+1$, are a basis. This basis is different from that one specified in section 4.8).

It becomes obvious that the use of basis splines removes the continuity constraints for s''(x) from our problem. Least squares approximation may now be done with respect to β_i, $i=-1,\ldots,n+1$, in straightforward way.

We could have used the basis splines defined in section 4.8 for our approximation procedure. Theoretically this is sound; practically, it is not. The basis splines $a_i(x)$, $i=0,\ldots,n$, $\alpha_0(x)$, $\alpha_n(x)$ introduced in section 4.8 do not have local support.

Consequently, the normal equations of the adjustment problem are not sparse. The splines $B_i(x)$ have local support. Hence the normal equations are sparse and can be solved much faster. (The computational effort is $O(n)$ versus $O(n^3)$ in the case of a full system). The basis splines of section 4.8 are good for exact interpolation. For approximation, the $B_i(x)$ are much better.

7.3. Two dimensions.

We assume a grid of nm square meshes of unit side length. The basic two-dimensional spline is

$$B(x,y) = B(x)\ B(y)$$

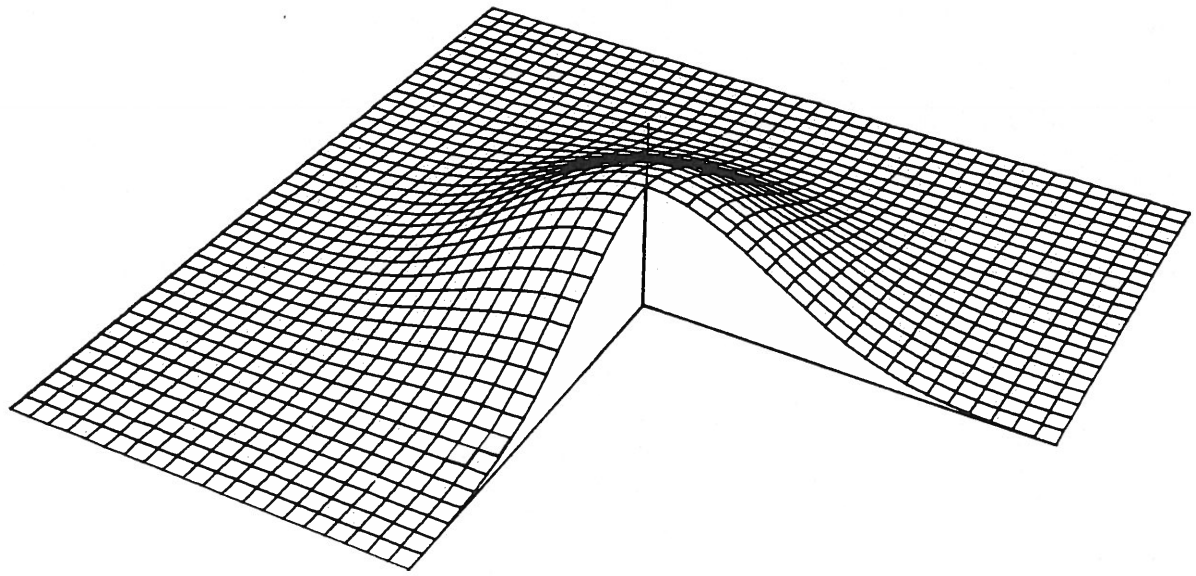


Fig. 7.2

Fig 7.2 shows this function. A complete basis is again obtained by shifting

$$B_{ij}(x,y) = B(x-i,y-j) = B(x-i)\,B(y-j)$$

$$i=-1,\ldots,m+1; \quad j=-1,\ldots,n+1$$

Using these basis functions, a spline is represented as

$$s(x,y) = \sum_{i=-1}^{m+1} \sum_{j=-1}^{n+1} \beta_{ij} B_{ij}(x,y)$$

The approximation problem with irregular data can be solved routinely. The normal equations will be sparse. The computational effort to solve this system is $O\{(nm)^{1.5}\}$ as opposed to $O\{(nm)^{3}\}$ if basis functions not having a local support are used.

MITTEILUNGEN DER GEODÄTISCHEN INSTITUTE
DER TECHNISCHEN UNIVERSITÄT GRAZ

(früher: Institutsmitteilungen des I. Geodätischen Institutes
der Technischen Hochschule in Graz)

Bisher erschienene Folgen:

Nr., Autor/Herausgeber, Titel, Seiten, Preis (ö.S.)

1 HUBENY K. (1959): Formeln und Tafeln zur Berechnung der 2. Hauptaufgabe auf dem Bessel'schen Ellipsoid für Strecken bis 300 km im Bereich zwischen den geographischen Breiten 45° und 57°30'. 33 Seiten, öS 45.

2 HUBENY K. (1959): Formeln und Tafeln zur Berechnung der geodätischen Hauptaufgaben über Normalschnitte für beliebige Ellipsoide und beliebige Entfernungen. 29 Seiten, öS 45.

3 MORITZ H. (1959): Untersuchungen über eine direkte Lösung der 2. Hauptaufgabe auf dem Rotationsellipsoid für beliebige Entfernungen. 16 Seiten, öS 35.

4 HUBENY K. (1960): Die Lösung der geodätischen Hauptaufgaben nach Bessel-Jordan. Erweiterte und neue Formeln sowie Tafeln für die Ellipsoide von Bessel und Hayford im Bereich zwischen den geographischen Breiten 45° und 58°. 65 Seiten, öS 50.

5 HUBENY K. (1960): Formeln und Tafeln zur Berechnung der geodätischen Hauptaufgaben über beliebige Entfernungen (Internationales Ellipsoid). 44 Seiten, öS 50.

6 HUBENY K., K. RINNER (1966): Vorlesungen am II. Fortbildungskurs für Praktiker an der Technischen Hochschule in Graz vom 5. bis 7. Oktober 1964. 156 Seiten, öS 50.

7 RINNER K. (1967): Geodätische Programme im Rechenzentrum Graz (Stand 9. Oktober 1967). 191 Seiten, öS 50.

8 RINNER K. (1968): Vorlesungen am III. Fortbildungskurs für Praktiker an der Technischen Hochschule in Graz vom 9. bis 12. Oktober 1967. 282 Seiten, öS 50.

9 RINNER K., G. BRANDSTÄTTER (1971): Forschungsberichte über Erdgezeiten und Satellitengeodäsie. 121 Seiten, öS 50.

10 FELDBACHER F., K. HUBENY, K. RINNER (1971): Beiträge zur ellipsoidischen Geometrie und zu Mikrowellen- und Lasermessungen für große Entfernungen. 78 Seiten, öS 50.

11 RINNER K. (1972): Proceedings of the International Symposium "Satellite and Terrestrial Triangulation"; 2 volumes: 1) Sessions of the West European Sub-Commission of the International Commission for Artificial Satellites, I.A.G.; 2) Sessions of the Special Study Group 1.26 of the I.A.G. 612 Seiten, öS 300.

12 BITTMANN O., G. KRAJICEK, P. MEISSL (1973): Microcomputer Compucorp 320 G und 322 G, die Benützung und Anwendungsbeispiele für die Vermessungstechnik. 53 Seiten, öS 30.

13 RINNER K. (1973): Berichte über Forschungsarbeiten. 57 Seiten. öS 50.

14 FRIEDL J., G. KRAJICEK, P. MEISSL (1974): Taschenrechner Hewlett-Packard HP-45, die Benützung und Anwendungsbeispiele für die Vermessungstechnik. 60 Seiten, öS 50.

15 BARTELME N., P. MEISSL (1974): Strength Analysis of Distance Networks. 57 Seiten, öS 50.

16 CHESI G., K. RINNER (1974): Tabellen zur meteorologischen Reduktion von Entfernungsmessungen mit dem Geodimeter 8. 100 Seiten, öS 100.

17 BITTMANN O., P. MEISSL (1974): Empfohlene Algorithmen zur Programmierung geodätischer Rechenaufgaben. I. Einfache Koordinatenrechnungen in der Ebene, 35 Seiten, öS 50.

18 MEISSL P., K. RINNER (1975): Vorträge am IV. Fortbildungskurs für Praktiker des Vermessungswesens an der Technischen Universität in Graz vom 25. bis 27. November 1974. 290 Seiten, öS 130.

19 LACHAPELLE G. (1975): Determination of the Geoid using Heteregeneous Data. 121 Seiten, öS 150.

20 MEISSL P., H. MORITZ, K. RINNER (1975): Contributions of the Graz Group to the XVI. General Assembly of IUGG/IAG in Grenoble. 308 Seiten, öS 150.

21 BENZ F., K. RINNER (1976): Verfahren zur Verminderung des Einflusses der Bodenreflexion bei der Entfernungsmessung mit Mikrowellen. 97 Seiten, öS 100.

22 RINNER K. (1976): Bericht über Laser- und Mikrowellenmessungen im Testnetz Steiermark. 109 Seiten, öS 100.

23 RINNER K. (1976): Bericht zur Meeresgeodäsie und Satellitengeodäsie. 111 Seiten, öS 100.

24 MEISSL P. (1976): Empfohlene Algorithmen zur Programmierung geodätischer Rechenaufgaben. II. Punktverwaltung mittels Massenspeicher. 69 Seiten, öS 50.

25 MEISSL P., K. STUBENVOLL (1977): Ein Computer-Programmsystem zur Verdichtung trigonometrischer Netze. 129 Seiten, öS 100.

26 KRYNSKI J., H. NOE, K.P. SCHWARZ, H. SONKEL (1977): Numerical Studies and Programs for Interpolation and Collocation. 67 Seiten, öS 50.

27 HUBENY K., A. REITHOFER (1977): Isotherme Koordinatensysteme und konforme Abbildungen des Rotationsellipsoides mit Tafeln und Programmen zur konformen Abbildung für die Ellipsoide von Bessel, Hayford, Krasowsky und für das Referenzellipsoid 1967. 222 Seiten, öS 200.

28 SONKEL H. (1977): Die Darstellung geodätischer Integralformeln durch bikubische Spline-Funktionen. 161 Seiten, öS 150.

29 LEBERL F. (1977): Proceedings of the International Symposium on Image Processing - Interactions with Photogrammetry and Remote Sensing, 3-5 October, 1977, Graz. 250 Seiten, öS 160.

30 ALLMER F. (1977): Dr. Ing. h.c. Eduard Ritter von Orel, dem Erfinder des Stereo-Autographen zum 100. Geburtstag. 41 Seiten, öS 50.

31 KRYNSKY J. (1978): Possibilities of Low-low Satellite Tracking for Local Geoid Improvement. 67 Seiten, öS 70.

32 GERONTOPOULOS P. (1978): Molodensky's Problem in the Plane. 160 Seiten, öS 200.

33 LEBERL F. (1980): Beiträge zur Radargrammetrie und digitalen Bildverarbeitung. 230 Seiten, öS 200.

34 HUBENY K. (1980): Die Klothoide (Formeln, Tafeln, Beispiele). 122 Seiten, öS 120.

35 RINNER K. et al. (1980): Festschrift zur Emeritierung von Prof. Dr. K. Hubeny. 200 Seiten, öS 200.

36 NOE H. (1980): Numerical Investigations on the Problem of Molodensky. 80 Seiten, öS 90.

37 BARTELME N., B. HOFMANN-WELLENHOF, P. MEISSL (1980): Empfohlene Algorithmen zur Programmierung geodätischer Rechenaufgaben. III. Zugriff auf Meßdatendatei. 200 Seiten, öS 180.

38 LICHTENEGGER H., K. RINNER (1982): Verzeichnis der Habilitationen, Dissertationen, Diplomarbeiten sowie von Seminar- und Proseminararbeiten 1960-1981. 81 Seiten, öS 100.

39 CHEN C.-Y. (1982): Geodetic Datum and Doppler Positioning (Dissertation). 255 Seiten, öS S 200.

40 MORITZ H. et al. (1982): Geodaesia Universalis. Festschrift Karl Rinner zum 70. Geburtstag. 382 Seiten, öS 250.

41 MORITZ H., H. SONKEL (1982): Geodesy and Global Geodynamics. 689 Seiten, öS 400.

42 RINNER K., H. LICHTENEGGER (1982): Proceedings of the International Symposium "Education in Geodesy", Sept. 27-29, 1982, Graz, Austria.

43 MEISSL P. (1982): Least Squares Adjustment. A Modern Approach. 440 Seiten, öS 450.

44 ALLMER F. (1983): Biographie Peter Meissl (in Vorbereitung -in preparation).