NETHERLANDS GEODETIC COMMISSION

PUBLICATIONS ON GEODESY NEW SERIES

VOLUME 8 NUMBER 1

THE GEOMETRY OF GEODETIC INVERSE LINEAR MAPPING AND NON-LINEAR ADJUSTMENT

by

P. J. G. TEUNISSEN

1985

RIJKSCOMMISSIE VOOR GEODESIE, THIJSSEWEG 11, DELFT, THE NETHERLANDS

PRINTED BY W. D. MEINEMA B.V., DELFT, THE NETHERLANDS

ISBN 90 6132 233 2

SUMMARY

This publication discusses

1^{o} The problem of inverse linear mapping

and

2^{o} The problem of non-linear adjustment.

After the introduction, which contains a motivation of our emphasis on geometric thinking, we commence in chapter **II** with the theory of inverse linear mapping. Amongst other things we show that every inverse **B** of a given linear map **A** can be uniquely characterized through the choice of three linear subspaces, denoted by $\mathcal{S}$, $\mathcal{C}$ and $\mathcal{D}$.

Chapter **III** elaborates on the consequences of the inverse linear mapping problem for planar, ellipsoidal and three dimensional geodetic networks. For various situations we construct sets of base vectors for the nullspace $\mathcal{N}u(\mathbf{A})$ of the designmap. The chapter is concluded with a discussion on the problem of connecting geodetic networks. We discuss, under fairly general assumptions concerning the admitted degrees of freedom of the networks involved, three alternative methods of connection.

Chapter **IV** treats the problem of non-linear adjustment. After a general problem statement and a brief introduction into Riemannian geometry, we discuss the local convergence behaviour of Gauss' iteration method (GM). A differential geometric approach is used throughout.
For both one dimensional and higher dimensional curved manifolds we show that the local behaviour of GM is asymptotically linear. Important conclusions are further that the local convergence behaviour of GM, 1^{o}. is predominantly determined by the least-squares residual vector and the corresponding extrinsic curvature of the manifold, 2^{o}. is invariant against reparametrizations in case of asymptotic linear convergence, 3^{o}. is asymptotically quadratic in case either the least-squares residual vector or the normal field **B** vanishes, 4^{o}. is determined by the Christoffel symbols of the second kind in case of asymptotic quadratic convergence and 5^{o}. will practically not be affected by line search strategies if both the least-squares residual vector and extrinsic curvature are small enough.
Next we discuss some conditions which assure global convergence of GM.
Thereupon we show that for a particular class of manifolds, namely ruled surfaces, important simplifications of the non-linear least-squares adjustment problem can be obtained through dimensional reduction. Application of this idea made it possible to obtain an inversion-free solution of a non-linear variant of the classical two dimensional Helmert transformation. This non-linear variant has been called the Symmetric Helmert transformation. We also give an inversion-free solution of the two dimensional Symmetric Helmert transformation when a non-trivial rotational invariant covariance structure is pre-supposed. After this we generalize our results to three dimensions.
In the remaining sections of chapter **IV** we give some suggestions as to how to estimate the extrinsic curvatures in practice; we estimate the curvature of some simple 2-dimensional geodetic networks and we briefly discuss some of the consequences of non-linearity for the statistical treatment of an adjustment. Hereby it is also shown that the bias of the least-squares residual vector is determined by

the mean curvature of the manifold and that the bias of the least-squares parameter estimator is determined by the trace of the Christoffelsymbols of the second kind.
The chapter is concluded with a brief discussion of some problems which are still open for future research.

ACKNOWLEDGEMENTS

The author greatfully acknowledges the support received from the following organisations:

The Netherlands Geodetic Commission for granting travelling funds,

The Netherlands Organisation for the Advancement of Pure Research (Nederlandse Organisatie voor Zuiver-Wetenschappelijk Onderzoek Z.W.O.) for awarding a research-grant, and

The Geodetic Institute of the Stuttgart University (FRG) for the facilities offered during the author's stay in Stuttgart.

Finally,special thanks go to miss Janna Blotwijk for the excellent job she did in typing and preparing the final version of this publication.

THE GEOMETRY OF GEODETIC INVERSE LINEAR MAPPING

AND NON-LINEAR ADJUSTMENT

SUMMARY iii
ACKNOWLEDGEMENTS v

I INTRODUCTION 1

II GEOMETRY OF INVERSE LINEAR MAPPING

1. The Principles 10
2. Arbitrary Inverses Uniquely Characterized 13
3. Injective and Surjective Maps 18
4. Arbitrary Systems of Linear Equations and Arbitrary Inverses 22
5. Some Common Type of Inverses and their Relation to the Subspaces S, C and $\mathcal{D}$ 24
6. C - and S-Transformations 30

III GEODETIC INVERSE MAPPING

1. Introduction 35
2. Geodetic Networks and their Degrees of Freedom 36
   2.1. Planar networks 36
   2.2. Ellipsoidal networks 42
   2.3. Three dimensional networks 52
3. (Free)Networks and their Connection 65
   3.1. Types of networks considered 65
   3.2. Three alternatives 68

IV GEOMETRY OF NON-LINEAR ADJUSTMENT

1. General Problem Statement 84
2. A Brief Introduction into Riemannian Geometry 87
3. Orthogonal Projection onto a Parametrized Space Curve 91
   3.1. Gauss' iteration method 91
   3.2. The Frenet frame 92
   3.3. The "Kissing" circle 95
   3.4. One dimensional Gauss- and Weingarten equations 97
   3.5. Local convergence behaviour of Gauss' iteration method 98
   3.6. Examples 102

3.7. Conclusions 109
4. Orthogonal Projection onto a Parametrized Submanifold 110
4.1. Gauss' method 110
4.2. The Gauss' equation 112
4.3. The normalfield **B** 116
4.4. The local rate of convergence 118
4.5. Global convergence 125
5. Supplements and Examples 134
5.1. The two dimensional Helmert transformation 134
5.2. Orthogonal projection onto a ruled surface 139
5.3. The two dimensional Symmetric Helmert transformation 141
5.4. The two dimensional Symmetric Helmert transformation with a non-trivial rotational invariant covariance structure 145
5.5. The three dimensional Helmert transformation and its symmetrical generalization 148
5.6. The extrinsic curvatures estimated 156
5.7. Some two dimensional networks 163
6. Some Statistical Considerations 166
7. Epilogue 170

REFERENCES 173

I. INTRODUCTION

This publication has the intention to give a contribution to the theory of geodetic adjustment. The two main topics discussed are

1° The problem of inverse linear mapping

and

2° The problem of non-linear adjustment

In our discussion of these two problems there is a strong emphasis on geometric thinking as a means of visualizing and thereby improving our understanding of methods of adjustment. It is namely our belief that a geometric approach to adjustment renders a more general and simpler treatment of various aspects of adjustment theory possible. So is it possible to carry through quite rigorous trains of reasoning in geometrical terms without translating them into algebra. This gives a considerable economy both in thought and in communication of thought. Also does it enable us to recognize and understand more easily the basic notions and essential concepts involved. And most important, perhaps, is the fact that our geometrical imagery in two and three dimensions suggests results for more dimensions and offers us a powerful tool of inductive and creative reasoning. At the same time, when precise mathematical reasoning is required it will be carried out in terms of the theory of finite dimensional vector spaces. This theory may be regarded as a precise mathematical framework underlying the heuristic patterns of geometric thought.

In Geodesy it is very common to use geometric reasoning. In fact, geodesy benefited considerably from the development of the study of differential geometry which was begun very early in history. Practical tasks in cartography and geodesy caused and influenced the creation of the classical theory of surfaces (Gauss, 1827; Helmert, 1880). And differential geometry can now be said to constitute an essential part of the foundation of both mathematical and physical geodesy (Marussi, 1952; Hotine, 1969; Grafarend, 1973).
But it was not only in the development of geodetic models that geometry played such a pivotal rôle. Also in geodetic adjustment theory, adjustment was soon considered as a geometrical problem. Very early (Tienstra, 1947; 1948; 1956) already advocated the use of the Ricci-calculus in adjustment theory. It permits a consistent geometrization of the adjustment of correlated observations. His approach was later followed by (Baarda, 1967 a,b; 1969), (Kooimans, 1958) and many others.
More recently we witness a renewed interest in the geometrization of adjustment theory. See e.g. (Vanicek, 1979), (Eeg, 1982), (Meissl, 1982), (Blais, 1983) or (Blaha, 1984). The incentive to this renewed interest is probably due to the introduction into geodesy of the modern theory of Hilbert spaces with kernel functions (Krarup, 1969). As (Moritz, 1979) has put it rather plainly, this theory can be seen as an infinitely dimensional generalization of Tienstra's theory of correlated observations in its geometrical interpretation.

Probably the best motivation for taking a geometric standpoint in discussing adjustment problems in linear models is given by the following discussion which emphasizes the geometric interplay between

best linear unbiased estimation and least-squares estimation:

Let $\mathbf{y}$ be a random vector in the m-dimensional Euclidean space M with metric tensor $\langle .,. \rangle_M$. We assume that $\mathbf{y}$ has an expected value $\tilde{\mathbf{y}} \in M$, i.e.,

$$E\{\mathbf{y}\} = \tilde{\mathbf{y}} \in M, \tag{1.1}$$

where $E\{.\}$ is the mathematical expectation operator, and that $\mathbf{y}$ has a covariance map

$$\mathbf{Q}_y : M^* \to M, \quad \text{defined by } \mathbf{Q}_y^{-1} y_1 = \langle y_1, . \rangle_M \quad \forall y_1 \in M \quad . \tag{1.2}$$

The linear vector space M^* denotes the dual space of M and is defined as the set of all real-valued (homogeneous) linear functions defined on M. Thus each $y^* \in M^*$ is a linear function $y^* : M \to \mathbb{R}$. Instead of writing $y^*(y_1)$ we will use a more symmetric formulation, by considering $y^*(y_1)$ as a bilinear function in the two variables y^* and y_1. This bilinear function is denoted by $(.,.) : M^* \times M \to \mathbb{R}$ and is defined by $(y^*, y_1) = y^*(y_1) \ \forall \ y^* \in M^*, y_1 \in M$. The function $(.,.)$ is called the duality pairing of M^* and M into $\mathbb{R}$.

We define a **linear model** as

$$\tilde{\mathbf{y}} \in \bar{N} \subset M, \quad \mathbf{Q}_y, \tag{1.3}$$

where $\bar{N}$ is a linear manifold in M. A linear manifold can best be viewed as a translated subspace. We will assume that $\bar{N} = \{y_1\} + U$, where y_1 is a fixed vector of M and U is an n-dimensional proper subspace of M.

The problem of linear estimation can now be formulated as: given an observation y_s on the random vector $\mathbf{y}$, its covariance map $\mathbf{Q}_y$ and the linear manifold $\bar{N}$, estimate the position of $\tilde{\mathbf{y}}$ in $\bar{N} \subset M$.

If we restrict ourselves to Best Linear Unbiased Estimation (BLUE), then the problem of linear estimation can be formulated dually as: given an $y_s^* \in M^*$, find $\hat{a} \in \mathbb{R}$ and $\hat{y}^* \in M^*$ such that the inhomogeneous linear function $h(\mathbf{y}) = \hat{a} + (\hat{y}^*, \mathbf{y})$ is a BLUE's estimator of $(y_s^*, \tilde{\mathbf{y}})$. The function $h(\mathbf{y})$ is said to be a BLUE's estimator of $(y_s^*, \tilde{\mathbf{y}})$ if,

1^o $h(\mathbf{y})$ is a linear unbiased estimator of $(y_s^*, \tilde{\mathbf{y}})$, i.e.,
if $E\{h(\mathbf{y})\} = (y_s^*, \tilde{\mathbf{y}}), \quad \forall \ \tilde{\mathbf{y}} \in \bar{N}$,

and (1.4)

2^o $h(\mathbf{y})$ is best, i.e.,
Variance $\{h(\mathbf{y})\} \leq$ Variance $\{g(\mathbf{y})\}$ for all linear unbiased estimators $g(\mathbf{y}) = a + (y^*, \mathbf{y}), \quad a \in \mathbb{R}, \ y^* \in M^*$, of $(y_s^*, \tilde{\mathbf{y}})$.

From (1.4.1^o) follows that

$$\hat{a} + (\hat{y}^*, \tilde{\mathbf{y}}) = (y_s^*, \tilde{\mathbf{y}}), \quad \forall \ \tilde{\mathbf{y}} \in \bar{N} \ ,$$

or

$$\hat{a} = (y_s^* - \hat{y}^*, \tilde{y}), \quad \forall\ \tilde{y} \in \bar{N} ,$$

or

$$\hat{a} = (y_s^* - \hat{y}^*, y_1) \quad \text{for some } y_1 \in \bar{N} \quad \text{and} \quad (y_s^* - \hat{y}^*, U) = 0, \tag{1.5}$$

since $\bar{N} = \{y_1\} + U$.
The set of $y^* \in M^*$ for which $(y^*, U) = 0$, forms a subspace of M^*. It is called the annihilator of $U \subset M$ and is denoted by $U^o \subset M^*$, i.e. $(U^o, U) = 0$. This gives for (1.5),

$$\hat{a} = (y_s^* - \hat{y}^*, y_1) \quad \text{for some } y_1 \in \bar{N}, \text{ and } y_s^* - \hat{y}^* \in U^o. \tag{1.6}$$

From (1.4.2^o) follows with (1.6) that $\hat{y}^* \in \{y_s^*\} + U^o$ must satisfy

$$(\hat{y}^*, Q_y \hat{y}^*) \leq (y^*, Q_y y^*) , \ \forall\ y^* \in \{y_s^*\} + U^o . \tag{1.7}$$

If we now define the dual metric of M^* by pulling the metric of M back by Q_y, i.e.,

$$\langle y^*, \tilde{y}^* \rangle_{M^*} = \langle Q_y y^*, Q_y \tilde{y}^* \rangle_M \qquad \forall\ y^*, \tilde{y}^* \in M^*,$$

it follows that $\hat{y}^* \in \{y_s^*\} + U^o$ must satisfy

$$\langle \hat{y}^*, \hat{y}^* \rangle_{M^*} \leq \langle y^*, y^* \rangle_{M^*} \quad \forall\ y^* \in \{y_s^*\} + U^o. \tag{1.8}$$

Geometrically this problem can be seen as the problem of finding that point $\hat{y}^*$ in $\{y_s^*\} + U^o$ which has least distance to the origin of M^*. And it will be intuitively clear that $\hat{y}^*$ is found by orthogonally projecting y_s^* onto the orthogonal complement $(U^o)^\perp$ of U^o (see figure 1).

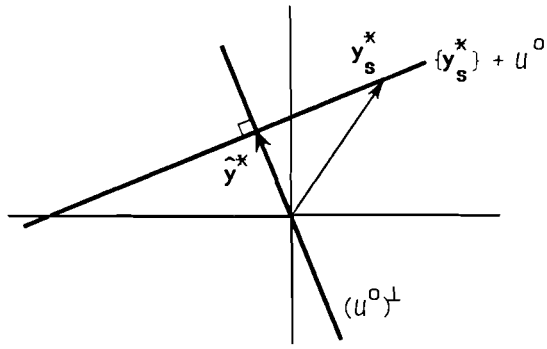


figure 1

Now, before we characterize the map which maps y_s^* into $\hat{y}^*$, let us first present some generalities on linear maps.

Let N and M be two linear vectorspaces of dimensions n and m respectively, and let $\mathbf{A}: N \to M$ be a linear map between them. Then we define the **image of** $U \subset N$ **under A** as

$$\mathbf{A}\,U = \{\, \mathbf{y} \in M \mid \mathbf{y} = \mathbf{A}\,\mathbf{x} \text{ for some } \mathbf{x} \in U \,\}\ . \tag{1.9}$$

The **inverse image of** $V \subset M$ **under A** is defined as

$$\mathbf{A}^{-1}(V) = \{\, \mathbf{x} \in N \mid \mathbf{A}\,\mathbf{x} \in V \,\}\ . \tag{1.10}$$

In the special case that $U = N$, the image of U under **A** is called the **range space** $R(\mathbf{A})$ of **A.** And the inverse image of $\{\mathbf{0}\} \in M$ under **A** is called the **nullspace** $Nu(\mathbf{A})$ of **A.** It is easily verified that if V and U are linear subspaces of M and N respectively, so are $\mathbf{A}U$ and $\mathbf{A}^{-1}(V)$.
A linear map $\mathbf{A}: N \to M$ is **injective** or one-to-one if for every $\mathbf{x}_1, \mathbf{x}_2 \in N$, $\mathbf{x}_1 \neq \mathbf{x}_2$ implies that $\mathbf{A}\,\mathbf{x}_1 \neq \mathbf{A}\,\mathbf{x}_2$. The map **A** is **surjective** or onto if $\mathbf{A}\,N = M$. And **A** is called **bijective** or a bijection if **A** is both injective and surjective.
With the linear map $\mathbf{A} : N \to M$ and the dual vector (or 1-form) $\mathbf{y}^* \in M^*$ it follows that the composition $\mathbf{y}^* \circ \mathbf{A}$ is a linear function which maps N into $\mathbb{R}$, i.e. $\mathbf{y}^* \circ \mathbf{A} \in N^*$. Since the map **A** assigns the 1-form $\mathbf{y}^* \circ \mathbf{A} \in N^*$ to $\mathbf{y}^* \in M^*$ we see that the map **A** induces another linear map, $\mathbf{A}^*$ say, which maps M^* into N^*. This map $\mathbf{A}^*$ is called the **dual map** to **A** and is defined as $\mathbf{A}^*\mathbf{y}^* = \mathbf{y}^* \circ \mathbf{A}$. With the duality pairing it is easily verified that

$$(\mathbf{A}^*\mathbf{y}^*, \mathbf{x}) = (\mathbf{y}^*, \mathbf{A}\,\mathbf{x})\ . \tag{1.11}$$

An important consequence of this bilinear identity is that for a non-empty inverse image of subspace $V \subset M$ under **A**, we have the duality relation

$$(\mathbf{A}^{-1}(V))^{\circ} = \mathbf{A}^*(V^{\circ})\ . \tag{1.12}$$

Note that here the four concepts of image, inverse image, annihilation and duality come together in one formula. For the special case that $V = \{\mathbf{0}\}$ the relation reduces to $Nu(\mathbf{A})^{\circ} = R(\mathbf{A}^*)$.

Maps that play an important role in linear estimation are the so-called projector maps. Assume that the subspaces U and V of N are complementary, i.e. $N = U \oplus V$, with "$\oplus$" denoting the direct sum. Then for $\mathbf{x} \in N$ we have the unique decomposition $\mathbf{x} = \mathbf{x}_1 + \mathbf{x}_2$ with $\mathbf{x}_1 \in U$, $\mathbf{x}_2 \in V$. We can now define a linear map $\mathbf{P}: N \to N$ through

$$\mathbf{P}\,\mathbf{x} = \mathbf{x}_1\ , \qquad \text{with } \mathbf{x} = \mathbf{x}_1 + \mathbf{x}_2,\ \mathbf{x}_1 \in U,\ \mathbf{x}_2 \in V \text{ and } N = U \oplus V\ . \tag{1.13}$$

This map is called the projector which projects onto U and along V. It is denoted by $\mathbf{P}_{U,V}$ (see figure 2).

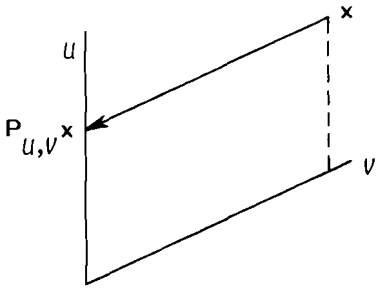


figure 2

If **P** projects onto U and along V then **I - P**, with **I** the identity map, projects onto V and along U. Thus

$$\mathbf{I} - \mathbf{P}_{U,V} = \mathbf{P}_{V,U} \quad . \tag{1.14}$$

For their images and inverse images we have

$$\left.\begin{aligned} \mathbf{P}_{U,V}\, N = U \;, \qquad & \mathbf{P}^{-1}_{U,V}(\mathbf{0}) = V \;, \qquad \mathbf{P}^{-1}_{U,V}(U) = N \\ (\mathbf{I} - \mathbf{P}_{U,V})\, N = V \;, \quad & (\mathbf{I}-\mathbf{P}_{U,V})^{-1}(\mathbf{0}) = U \;, \quad (\mathbf{I}-\mathbf{P}_{U,V})^{-1}(V) = N \end{aligned}\right\} \tag{1.15}$$

It is easily verified that the dual $\mathbf{P}^*$ of a projector **P** is again a projector operating on the dual space. For we have with (1.12) and (1.15):

$$(\mathbf{P}^{-1}_{U,V}(\mathbf{0}))^{o} = V^{o} = \mathbf{P}^{*}_{U,V} N^{*} \quad \text{and} \quad (\mathbf{P}^{-1}_{U,V}(U))^{o} = N^{o} = \{\mathbf{0}\} = \mathbf{P}^{*}_{U,V} U^{o}.$$

Thus,

$$\mathbf{P}^{*}_{U,V} = \mathbf{P}_{V^{o},U^{o}} \quad \text{and} \quad (\mathbf{I}-\mathbf{P}_{U,V})^{*} = \mathbf{P}^{*}_{V,U} = \mathbf{P}_{U^{o},V^{o}} \quad . \tag{1.16}$$

Finally we mention that one can check whether a linear map is a projector, by verifying whether the iterated operator coincides with the operator itself (Idempotence).

Now let us return to the point where we left our BLUE's problem. We noted that $\hat{\mathbf{y}}^*$ could be found by orthogonally projecting $\mathbf{y}^*_{\mathbf{s}}$ onto $(U^{o})^{\perp}$. Hence, the projector map needed is the one which projects onto $(U^{o})^{\perp}$ and along U^{o}, i.e.,

$$\hat{\mathbf{y}}^* = \mathbf{P}_{(U^{o})^{\perp},U^{o}}\mathbf{y}^*_{\mathbf{s}} \quad . \tag{1.17}$$

From (1.6) and (1.17) follows then that the linear function $h(\mathbf{y})$ is the unique BLUE's estimator of $(\mathbf{y}^*_{\mathbf{s}}, \tilde{\mathbf{y}})$ if

$$h(y) = \hat{a} + (\hat{y}^*, y) = ((I - P_{(U^o)^\perp, U^o}) y_s^*, y_1) + (P_{(U^o)^\perp, U^o} y_s^*, y),$$

or (1.18)

$$h(y) = (y_s^*, y_1) + (P_{(U^o)^\perp, U^o} y_s^*, y - y_1),$$

where y_1 is an arbitrary element of $\bar{N}$.
Application of the definition of the dual map gives

$$h(y) = (y_s^*, y_1) + (y_s^*, P^*_{(U^o)^\perp, U^o}(y - y_1)).$$

And since

$$P^*_{(U^o)^\perp, U^o} = P_{U, U^\perp},$$

we get

$$h(y) = (y_s^*, y_1 + P_{U, U^\perp}(y - y_1)),$$

in which we recognize the least-squares estimate

$$\hat{y} = y_1 + P_{U, U^\perp}(y_s - y_1), \quad y_1 \in \bar{N}, \tag{1.19}$$

which solves the dual problem

$$\langle y_s - \hat{y}, y_s - \hat{y} \rangle_M \leq \langle y_s - y, y_s - y \rangle_M \quad \forall\, y \in \bar{N}, \tag{1.20}$$

(see figure 3).

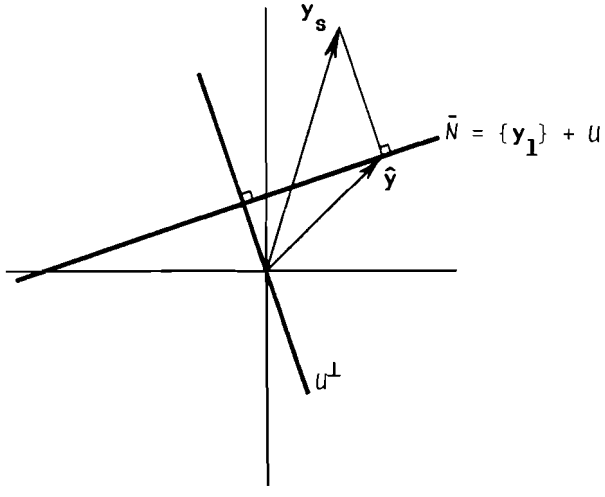


figure 3

Thus we have recovered the existing duality between BLUE's estimation and least-squares estimation. We minimize a sum of squares (1.20) and emerge with an optimum estimator, namely one which minimizes another sum of squares (1.8), the variance. From the geometrical viewpoint this arises simply from the duality between the so-called observation space M and estimator space M^*, established by the duality pairing $(\mathbf{y}^*, \mathbf{y})$.
The above given result is of course the well known Gauss-Markov theorem which probabilistically justifies least-squares estimation in case of linear models.

Observe that the above discussion shows another advantage of geometric reasoning, namely that the language of geometry embodies an element of invariance. That is, geometric reasoning avoids unnecessary reference to particular sets of coordinate axes. Concepts such as linear projections and linear manifolds for instance, may be visualized in a coordinate-free or invariant way. All results obtained by an invariant approach therefore necessarily apply to all possible representations of the linear manifold $\bar{N}$. That is, one could define $\bar{N}$ by a linear map $\mathbf{A}$ from the parameter space N into the observation space M (in Tienstra's terminology this would be "standard problem II") or implicitly by a set of linear constraints ("standard problem I"). Even a mixed representation is possible. Consequently, in general we have that if a coordinate representation is needed one can take the one which seems to be the most appropriate. That is, the use of a convenient basis rather than a basis fixed at the outset is a good illustration of the fact that coordinate-free does not mean freedom from coordinates so much as it means freedom to choose the appropriate coordinates for the task at hand.
With respect to our first topic, note that a direct consequence of the coordinate-free formulation is that the difficulties are evaded which might possibly occur when a non-injective linear map $\mathbf{A}$ is used to specify the linear model. This indicates that the actual problem of inverse linear mapping should not be considered to constitute an essential part of the problem of adjustment. That is, in the context of BLUE's estimation it is insignificant which pre-image of $\hat{\mathbf{y}}$ under $\mathbf{A}$ is taken. This viewpoint seems, however, still not generally agreed upon. The usually merely algebraic approach taken often makes one omit to distinguish between the actual adjustment problem and the actual inverse mapping problem. As a consequence, published studies in the geodetic literature dealing with the theory of inverse linear mapping surpass in our view often the essential concepts involved. We have therefore tried to present an alternative approach; one that is based on the idea that once the causes of the general inverse mapping problem are classified, also the problem of inverse linear mapping itself is solved. Our approach starts from the identification of the basic subspaces involved and next shows that the problem of inverse linear mapping can be reduced to a few essentials.

As to our second topic, that of non-linear adjustment, note that the Gauss-Markov theorem formulates a lot of "ifs" before it states why least-squares should be used: if the mean $\tilde{\mathbf{y}}$ lies in a linear manifold $\bar{N}$, if the covariance map is known to be $\mathbf{Q}_y$, if we are willing to confine ourselves to estimates that are unbiased in the mean and if we are willing to apply the quality criterium of minimum variance, then the best estimate is to be had by least-squares. These are a lot of "ifs" and it would be interesting to ask "and if not?". For all "ifs" this would become a complicated task indeed. But it will be clear that the first "if" which called for manifold $\bar{N}$ to be linear, already breaks down in case of non-linear models. Furthermore, in non-linear models a restriction to linear estimators does not seem reasonable anymore, because any estimator of $\tilde{\mathbf{y}}$ must be a mapping from M into

$\bar{N}$, which will be curved in general. Hence, strictly speaking the Gauss-Markov theorem does not apply anymore in the non-linear case. And consequently one might question whether the excessive use of the theorem in the geodetic literature for theoretical developments is justifiable in all cases.
Since almost all functional relations in our geodetic models are non-linear, one may be surprised to realize how little attention the complicated problem area of non-linear geodesic adjustment has received. One has used and is still predominantly using the ideas, concepts and results from the theory of linear estimation. Of course, one may argue that probably most non-linear models are only moderately non-linear and thus permit the use of a linear(ized) model. This is true. However, it does in no way release us from the obligation of really proving whether a linear(ized) model is sufficient as approximation. What we need therefore is knowledge of how non-linearity manifests itself at the various stages of adjustment. Here we agree with (Kubik, 1967), who points out that a general theoretical and practical investigation into the various aspects of non-linear adjustment is still lacking.
In the geodetic literature we only know of a few publications in which non-linear adjustment problems are discussed. In the papers by (Pope, 1972), (Stark and Mikhail, 1973), (Pope, 1974) and (Celmins, 1981; 1982) some pitfalls to be avoided when applying variable transformations or when updating and re-evaluating function values in an iteration procedure, are discussed. And in (Kubik, 1967) and (Kelley and Thompson, 1978) a brief review is given of some iteration methods. An investigation into the various effects of non-linearity was started in (Baarda, 1967 a,b), (Alberda, 1969), (Grafarend, 1970) and more recently in (Krarup, 1982a). (Alberda, 1969) discusses the effect of non-linearity on the misclosures of condition equations when a linear least-squares estimator is used and illustrates the things mentioned with a quadrilateral. A similar discussion can be found in (Baarda, 1967b), where also an expression is derived for the bias in the estimators. (Grafarend, 1970) discusses a case where the circular normal distribution should replace the ordinary normal distribution. And finally (Baarda, 1967a) and (Krarup, 1982a) exemplify the effect of non-linearity with the aid of a circular model.
Although we accentuate some different and new aspects of non-linear adjustment, our contribution to the problem of non-linear geodesic adjustment should be seen as a continuation of the work done by the above mentioned authors. We must admit though that unfortunately we do not have a cut and dried answer to all questions. We do hope, however, that our discussion of non-linear adjustment will make one more susceptible to the intrinsic difficulties of non-linear adjustment and that the problem will receive more attention than it has received hitherto.

The plan of this publication is the following:
In chapter II we consider the geometry of inverse linear mapping. We will show that every inverse **B** of a linear map **A** can be uniquely characterized through the choice of three subspaces S, C and D. Furthermore, each of these three subspaces has an interesting interpretation of its own. In order to facilitate reference the basic results are summarized in table 1.
In chapter III we start by showing the consequences of the inverse mapping problem for 2 and 3-dimensional geodetic networks. This part is easy-going since the planar case has to some extent already been treated elsewhere in the geodetic literature. The second part of this chapter presents a discussion on the in geodesy almost omnipresent problem of connecting geodetic networks.
Finally, chapter IV makes a start with the problem of non-linear adjustment. A differential geometric approach is used throughout. We discuss Gauss' method in some detail and show how the extrinsic

curvatures of submanifold $\bar{N}$ affects its local behaviour. And amongst other things, we also show how in some cases the geometry of the problem suggests important simplifications. Typical examples are our generalizations of the classical Helmert transformation.

II. GEOMETRY OF INVERSE LINEAR MAPPING

1. The principles

Many problems in physical science involve the estimation or computation of a number of unknown parameters which bear a linear (or linearized) relationship to a set of experimental data. The data may be contaminated by (systematic or random) errors, insufficient to determine the unknowns, redundant, or all of the above and consequently, questions as existence, uniqueness, stability, approximation and the physical description of the set of solutions are all of interest.

In econometrics for instance (see e.g. Neeleman, 1973) the problem of insufficient data is discussed under the heading of "multi-collinearity" and the consequent lack of determinability of the parameters from the observations is known there as the "identification problem". And in geophysics, where the physical interpretation of an anomalous gravitational field involves deduction of the mass distribution which produces the anomalous field, there is a fundamental non-uniqueness in potential field inversion, such that, for instance, even complete, perfect data on the earth's surface cannot distinguish between two buried spherical density anomalies having the same anomalous mass but different radii (see e.g. Backus and Gilbert, 1968).
Also in geodesy similar problems can be recognized. The fact that the data are generally only measured at discrete points, leaves one in physical geodesy for instance with the problem of determining a continuous unknown function from a finite set of data (see e.g. Rummel and Teunissen, 1982). Also the non-uniqueness in coordinate-system definitions makes itself felt when identifying, interpreting, qualifying and comparing results from geodetic network adjustments (see e.g. Baarda, 1973). The problem of connecting geodetic networks, which will be studied in chapter three, is a prime example in this respect.

All the above mentioned problems are very similar and even formally equivalent, if they are described in terms of a possible inconsistent and under-determined linear system

$$\mathbf{y} \doteq \mathbf{A}\mathbf{x} \quad , \tag{1.1}$$

where $\mathbf{A}$ is a linear map from the n-dimensional parameter space N into the m-dimensional observation space M.
The first question that arises is whether a solution to (1.1) exists at all, i.e. whether the given vector $\mathbf{y}$ is an element of the range space $R(\mathbf{A})$, $\mathbf{y} \in R(\mathbf{A})$. If this is the case we call the system **consistent.**
The system is certainly consistent if the rank of $\mathbf{A}$, which is defined as rank $\mathbf{A}$ = dim. $R(\mathbf{A})$ = r, equals the dimension of M. In this case namely the range space $R(\mathbf{A})$ equals M and therefore $\mathbf{y} \in M = R(\mathbf{A})$. In all other cases, r <dim. M, consistency is no longer guaranteed, since it would be a mere coincidence if the given vector $\mathbf{y} \in M$ lies in the smaller dimensioned subspace $R(\mathbf{A}) \subset M$. Consistency is thus guaranteed if $\mathbf{y} \in R(\mathbf{A}) = Nu(\mathbf{A}^*)^o$.
Assuming consistency, the next question one might ask is whether the solution of (1.1) is unique or

not, i.e. whether the vector **y** contains enough information to determine the vector **x**. If not, the system is said to be **under-determined.** The solution is only unique if the rank of **A** equals the dimension of its domain space N, i.e. if $r = \dim. N$. To see this, assume $\mathbf{x}_1$ and $\mathbf{x}_2 \neq \mathbf{x}_1$ to be two solutions to (1.1). Then $\mathbf{A}\mathbf{x}_1 = \mathbf{A}\mathbf{x}_2$ or $\mathbf{A}(\mathbf{x}_1-\mathbf{x}_2) = \mathbf{0}$ must hold. But this means that $r <$ dim. N.

From the above considerations follows that it is the relation of $r = \dim. \mathcal{R}(\mathbf{A})$ to $m = \dim. M$ and $n = \dim. N$, which decides on the general character of a linear system. In case $r = m = n$, we know that a unique inverse map **B** of the bijective map **A** exists, with the properties

$$\mathbf{B}\,\mathbf{A} = \mathbf{I} \quad \text{and} \quad \mathbf{A}\,\mathbf{B} = \mathbf{I} \ . \tag{1.2}$$

For non-bijective maps **A**, however, in general no map **B** can be found for which (1.2) holds. For such maps therefore a more relaxed type of inverse property is used. Guided by the idea that an inverse-like map **B** should solve any consistent system ,that is, map **B** should furnish for each $\mathbf{y} \in \mathcal{R}(\mathbf{A})$, some solution $\mathbf{x} = \mathbf{B}\mathbf{y}$ such that $\mathbf{y} = \mathbf{A}\mathbf{B}\mathbf{y}$, one obtains as defining property of **B**

$$\mathbf{A}\,\mathbf{B}\,\mathbf{A} = \mathbf{A} \ . \tag{1.3}$$

Maps $\mathbf{B}: M \rightarrow N$, which satisfy this relaxed type of inverse condition are now called **generalized** inverses of **A.**

In the geodetic literature there is an overwhelming list of papers which deal with the theory of generalized inverses (see e.g. Teunissen, 1984a and the references cited in it). It more or less started with the pioneering work of Bjerhammar (Bjerhammar, 1951) ,who defined a generalized inverse for rectangular matrices. And after the publication of Penrose (Penrose, 1955) the literature of generalized inverses has proliferated rapidly ever since.
Many of the published studies, however, follow a rather algebraic approach making use of anonymous inverses which merely produce a solution to the linear system under consideration. As a consequence of this anonymity the essential concepts involved in the problem of inverse linear mapping often stay concealed. Sometimes it even seems that algebraic manipulations and the stacking of theorems, lemma's, corollaries, and what have you, are preferred to a clear geometric interpretation of what really is involved in the problem of inverse linear mapping.
In this chapter we therefore approach the problem of inverse mapping from a different viewpoint. Our approach is based on the idea that once the causes of the inverse mapping problem are classified, also the problem of inverse mapping itself is solved. The following reminder may be helpful. We know that a map is uniquely determined once its basis values are given. But as the theorem of the next section shows, condition (1.3) does not fully specify all the basis values of the map **B.** Hence its non-uniqueness. This means, however, that analogously to the case where a basis of a subspace can be extended in many ways to a basis which generates the whole space, various maps satisfying (1.3) can be found by specifying their failing basis values.

To give a pictorial explanation of our procedure, observe that in the general case of rank $\mathbf{A} = r <$ min.(m,n), the nullspace $\mathcal{N}u(\mathbf{A}) \subset N$ and range space $\mathcal{R}(\mathbf{A}) \subset M$ both are proper subspaces. That is,

they do not coincide with respectively N and M (see figure 4).

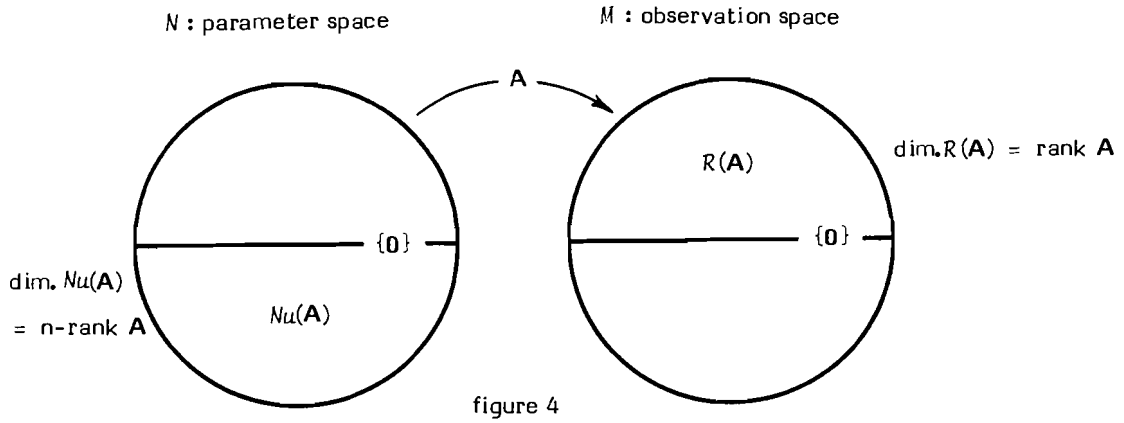


figure 4

Now, just like there are many ways in which a basis of a subspace can be extended to a basis which generates the whole space, there are many ways to extend the subspaces $Nu(\mathbf{A}) \subset N$ and $R(\mathbf{A}) \subset M$ to fill N and M respectively (see figure 5).

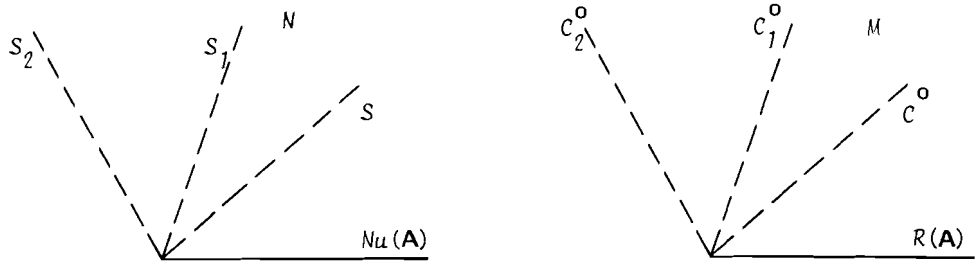


figure 5

Let us choose two arbitrary subspaces, say $S \subset N$ and $C^o \subset M$, such that the direct sums $S \oplus Nu(\mathbf{A})$ and $R(\mathbf{A}) \oplus C^o$ coincide with N and M (see figure 6).

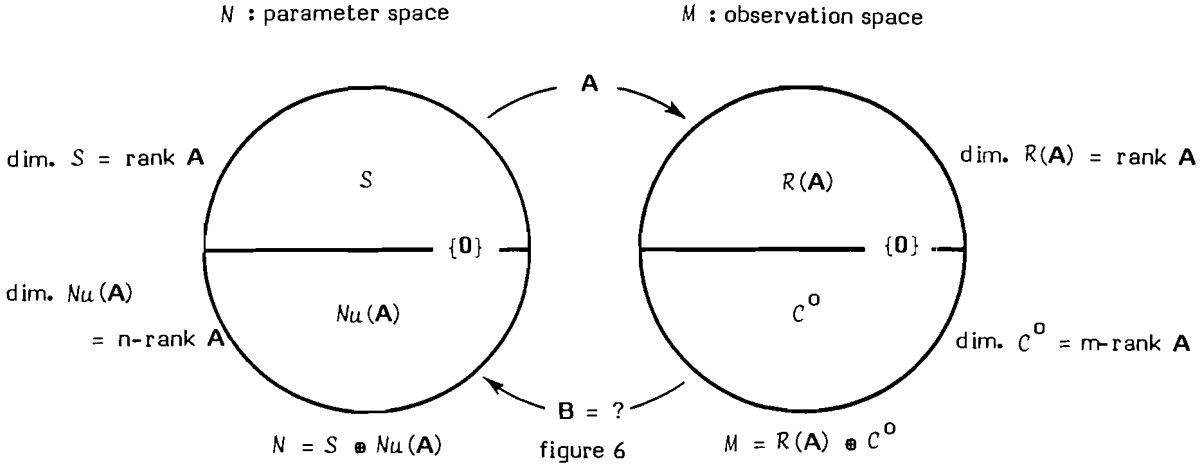


figure 6

The complementarity of S and $\mathcal{N}u(\mathbf{A})$ then implies that the subspace S has a dimension which equals that of $\mathcal{R}(\mathbf{A})$, i.e. dim. S = dim. $\mathcal{R}(\mathbf{A})$. But this means that map $\mathbf{A}$, when restricted to S, $\mathbf{A}|_S$, is bijective. There exist therefore linear maps $\mathbf{B}: M \to N$ which, when restricted to $\mathcal{R}(\mathbf{A})$, become the inverse of $\mathbf{A}|_S$ (see figure 7):

$$\mathbf{B}|_{\mathcal{R}(\mathbf{A})}\,\mathbf{A}|_S = \mathbf{I} \quad \text{and} \quad \mathbf{A}|_S\,\mathbf{B}|_{\mathcal{R}(\mathbf{A})} = \mathbf{I} \ . \tag{1.4}$$

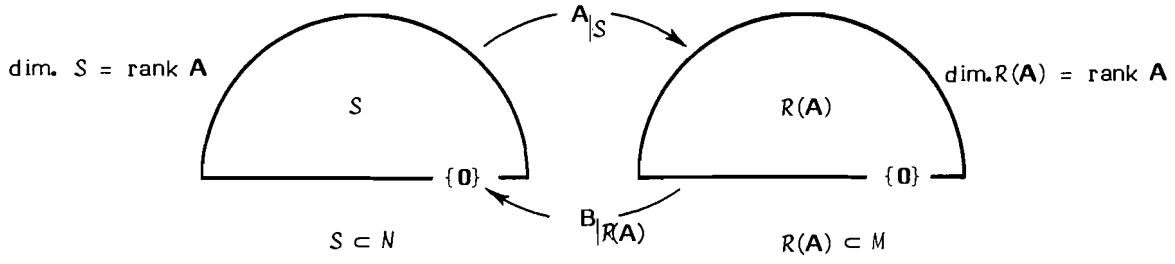


figure 7

The inverse-like properties (1.4) are thus the ones which replace (1.2) in the general case of rank $\mathbf{A} = r < \min.(m,n)$. The second equation of (1.4) can be rephrased as $\mathbf{ABA} = \mathbf{A}$, and therefore constitutes the classical definition of a generalized inverse of $\mathbf{A}$. The first equation of (1.4) states that

$$\mathbf{B}\,\mathbf{A}\,\mathbf{x} = \mathbf{x} \ , \qquad \forall\ \mathbf{x} \in S \ . \tag{1.5}$$

In the next section we will prove what is already intuitively clear, namely that equation (1.5) is equivalent to the classical definition (1.3), and therefore (1.5) can just as well be used as a definition of a generalized inverse. In fact, (1.5) has the advantage over (1.3) that it clearly shows why generalized inverses are not unique. The image of S under $\mathbf{A}$ is namely only a proper subspace of M. To find a particular map $\mathbf{B}$ which satisfies (1.5), we therefore need to specify its failing basis values.

2. Arbitrary inverses uniquely characterized

In this section we will follow our lead that a map is only uniquely determined once its basis values are completely specified.

As said, the usual way to define generalized inverses $\mathbf{B}$ of $\mathbf{A}$ is by requiring

$$\mathbf{A}\,\mathbf{B}\,\mathbf{A} = \mathbf{A}. \tag{2.1}$$

This expression, however, is not a very illuminating one, since it does not tell us what generalized inverses of $\mathbf{A}$ look like or how they can be computed. We will therefore rewrite expression (2.1) in such a form that it becomes relatively easy to understand the mapping characteristics of $\mathbf{B}$. This is done by the following theorem:

Theorem

1^o $\quad \mathbf{A\,B\,A = A} \iff$ For some unique $S \subset N$,
complementary to $Nu(\mathbf{A})$,
$\mathbf{B\,Ax = x}$, $\forall\ \mathbf{x} \in S$, holds.

2^o $\quad \mathbf{A\,B\,A = A} \iff \mathbf{A\,By = y}$, $\forall\ \mathbf{y} \in R(\mathbf{A})$.

Proof of 1^o

($\rightarrow$) From premultiplying **ABA = A** with **B** follows **BABA = BA.** The map **BA** is thus idempotent and therefore a projector from N into N.
From **ABA = A** also follows that $Nu(\mathbf{BA}) = Nu(\mathbf{A})$.
To see this, consider $\mathbf{x} \in Nu(\mathbf{BA})$. Then **BAx = 0** or **ABAx = Ax = 0,** which means that $\mathbf{x} \in Nu(\mathbf{A})$. Thus $Nu(\mathbf{BA}) \subset Nu(\mathbf{A})$. Conversely, if $\mathbf{x} \in Nu(\mathbf{A})$, then **Ax = 0** or **BAx = 0,** which means $\mathbf{x} \in Nu(\mathbf{BA})$. Thus we also have $Nu(\mathbf{A}) \subset Nu(\mathbf{BA})$. Hence $Nu(\mathbf{BA}) = Nu(\mathbf{A})$.
Now let us denote the subspace $R(\mathbf{BA})$ by S, i.e. $R(\mathbf{BA}) = S$. The projector property of **BA** then implies that $\mathbf{BAx = x}$, $\forall\ \mathbf{x} \in S$. And it also implies that $N = R(\mathbf{BA}) \oplus Nu(\mathbf{BA})$. With $R(\mathbf{BA}) = S$ and $Nu(\mathbf{BA}) = Nu(\mathbf{A})$ we therefore have that $N = S \oplus Nu(\mathbf{A})$. Hence the complementarity of S and $Nu(\mathbf{A})$.

($\leftarrow$) From $N = S \oplus Nu(\mathbf{A})$ follows the complementarity of S and $Nu(\mathbf{A})$. We can therefore construct the projector $P_{S,\,Nu(\mathbf{A})} = \mathbf{I} - P_{Nu(\mathbf{A}),S}$. With this projector we can now replace

$$\mathbf{B\,Ax = x}, \quad \forall \mathbf{x} \in S ,$$

by

$$\mathbf{B\,A}\,P_{S,\,Nu(\mathbf{A})}\mathbf{x} = P_{S,\,Nu(\mathbf{A})}\mathbf{x}, \quad \forall \mathbf{x} \in N .$$

And since $\mathbf{A}\,P_{S,\,Nu(\mathbf{A})} = \mathbf{A}(\mathbf{I} - P_{Nu(\mathbf{A}),S}) = \mathbf{A}$, we get

$$\mathbf{B\,A}\,P_{S,\,Nu(\mathbf{A})}\mathbf{x} = \mathbf{B\,Ax} = P_{S,\,Nu(\mathbf{A})}\mathbf{x}, \quad \forall \mathbf{x} \in N ,$$

or finally, after premultiplication with **A** ,

$$\mathbf{A\,B\,Ax = Ax}, \quad \forall \mathbf{x} \in N .$$

Proof of 2^o

We omit the proof since it is straightforward.

The above theorem thus makes precise what already was made intuitively clear in section one. There are now two important points which are put forward by the theorem. First of all, it states that every linear map $\mathbf{B}$: $M \to N$ which satisfies

$$\mathbf{B}\,\mathbf{A}x = x, \qquad \forall x \in S \ , \tag{2.2}$$

with $N = S \oplus Nu(\mathbf{A})$, is a generalized inverse of $\mathbf{A}$. And since

$$\begin{aligned} R(\mathbf{A}) = \mathbf{A}N &= \{y \in M \mid y = \mathbf{A}x \text{ for some } x \in N\} \\ &= \{y \in M \mid y = \mathbf{A}x \text{ for some } x = x_1 + x_2, x_i \in S, x_2 \in Nu(\mathbf{A})\} \\ &= \{y \in M \mid y = \mathbf{A}x \text{ for some } x \in S\} \\ &= \mathbf{A}S \ , \end{aligned}$$

this implies that a generalized inverse $\mathbf{B}$ of $\mathbf{A}$ maps the subspace $R(\mathbf{A}) \subset M$ onto a subspace $S \subset N$ complementary to $Nu(\mathbf{A})$. Map $\mathbf{B}$ therefore determines a one-to-one relation between $R(\mathbf{A})$ and S, and is injective when restricted to the subspace $R(\mathbf{A})$.

A second point that should be noted about the theorem is that it gives a way of constructing arbitrary generalized inverses of $\mathbf{A}$. To see this, consider expression (2.2). Since $R(\mathbf{A}) = \mathbf{A}N = \mathbf{A}S$, expression (2.2) only specifies how $\mathbf{B}$ maps a **subspace,** namely $R(\mathbf{A})$, of M. Condition (2.2) is therefore not sufficient for determining map $\mathbf{B}$ uniquely. Thus in order to be able to compute a particular generalized inverse of $\mathbf{A}$ one also needs to specify how $\mathbf{B}$ maps a basis of a subspace complementary to $R(\mathbf{A})$. Let us denote such a subspace by $C^o \subset M$, i.e. $M = R(\mathbf{A}) \oplus C^o$ Then if $\mathbf{e}_i$, i=1,...,m, and $\mathbf{e}_\alpha$, α=1,...,n, are bases of M and N, and $C_p^{\perp i}\mathbf{e}_i$, p=1,...,(m-r), *) forms a basis of C^o, a particular generalized inverse $\mathbf{B}$ of $\mathbf{A}$ is uniquely characterized by specifying in addition to (2.2) how it maps C^o, say:

$$\mathbf{B}\, C_p^{\perp i}\, \mathbf{e}_i = D_p^{\alpha}\, \mathbf{e}_\alpha \ , \qquad i=1,\ldots,m;\ \alpha=1,\ldots,n;\ p=1,\ldots,(m-r) \tag{2.3}$$

(Einstein's summation convention).

Thus if $\mathcal{D}$ denotes the subspace spanned by $D_p^{\alpha}\, \mathbf{e}_\alpha$, we have,

$$\mathbf{B}\, C^o = \mathcal{D} \subset N \ , \text{ with } M = R(\mathbf{A}) \oplus C^o \ . \tag{2.4}$$

Although the choice for $\mathcal{D} \subset N$ is completely free, we will show that one can impose an extra condition, namely $\mathcal{D} \subset Nu(\mathbf{A})$, without affecting generality. Note that point 2^o of the theorem says that $\mathbf{AB}$ is a projector, projecting onto the rangespace $R(\mathbf{A})$ and along a space, say $\bar{C}^o$, complementary to $R(\mathbf{A})$. With (2.4) we therefore get that

---

*) The kernel letter "$C^{\perp}$" expresses the fact that $C_p^{\perp i}\, \delta_{ij}\, C_q^j = 0$, $i,j = 1,\ldots,m$; $p=1,\ldots,(m-r)$; $q = 1,\ldots,r$, or in matrix notation that $(C^{\perp})^t\, C = O$ $\quad$ ($p \times m$, $m \times (m-p)$, $p \times (m-p)$).

$$\mathbf{P}_{R(\mathbf{A}),\mathcal{C}^o}\ \mathcal{C}^o = \mathbf{A}\,\mathcal{D} \quad .$$

But this means that if **B** is characterized by mapping $\mathcal{C}^o$ onto $\mathcal{D}$, there exists another subspace of M complementary to $R(\mathbf{A})$ which is mapped by **B** to a subspace of $Nu(\mathbf{A})$. We can therefore just as well start characterizing a particular generalized inverse **B** of **A** by (2.2) and (2.4), but now with the additional condition that $\mathcal{D} \subset Nu(\mathbf{A})$.

Summarizing, we have for the images of the two complementary subspaces $R(\mathbf{A}) = \mathbf{A}S$ and $\mathcal{C}^o$ under **B**:

$$\mathbf{B}\,\mathbf{A}\,S = S \qquad \text{and} \qquad \mathbf{B}\,\mathcal{C}^o = \mathcal{D}\ ,$$

with

$$N = S \oplus Nu(\mathbf{A}), \quad M = R(\mathbf{A}) \oplus \mathcal{C}^o \qquad (2.5)$$

and

$$\mathcal{D} \subset Nu(\mathbf{A})$$

A few things are depicted in figure 8.

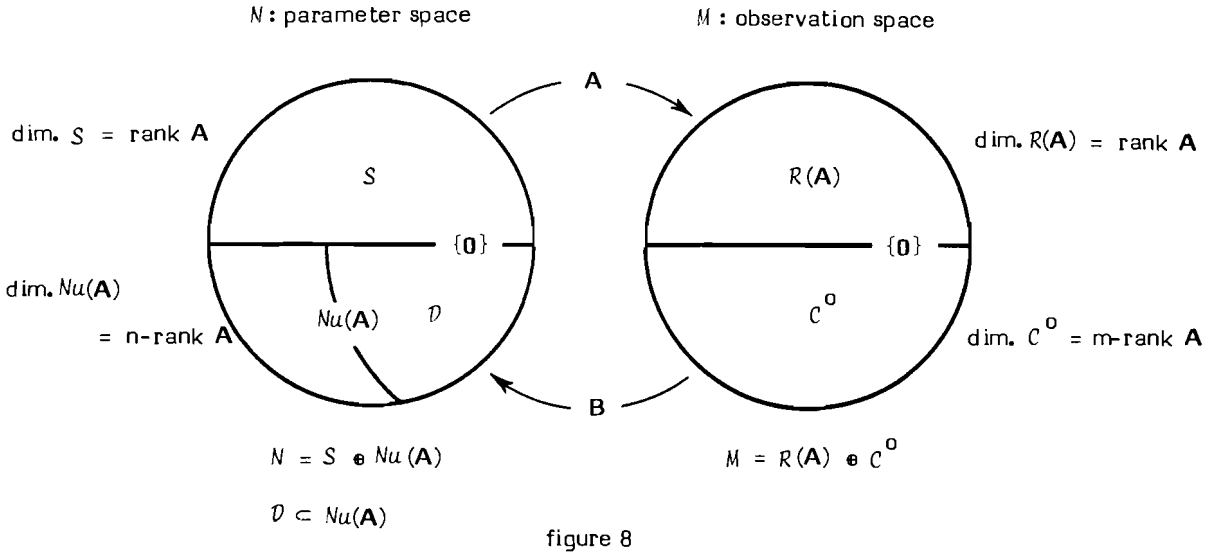


figure 8

Our objective of finding a unique representation of an arbitrary generalized inverse **B** of **A** can now be reached in a very simple way indeed. The only thing we have to do is to combine (2.2) and (2.3). If we take the coordinate expressions of **B** and **A** to be

$$\mathbf{B}\,\mathbf{e}_i = B_i^{\ \alpha}\,\mathbf{e}_\alpha \qquad \text{and} \qquad \mathbf{A}\,\mathbf{e}_\alpha = A_\alpha^{\ i}\,\mathbf{e}_i\ ,$$

where $\mathbf{e}_i$, $i=1,\ldots,m$, and $\mathbf{e}_\alpha$, $\alpha=1,\ldots,n$ are bases of M and N, and if we take as bases of S, $\mathcal{C}^o$ and $\mathcal{D}$,

$$S_q^{\ \alpha}\,\mathbf{e}_\alpha\ ,\ C_p^{\perp i}\,\mathbf{e}_i \quad \text{and} \quad D_p^{\ \alpha}\,\mathbf{e}_\alpha\ ,\ p=1,\ldots,(m-r);\ q=1,\ldots,r,$$

then (2.2) and (2.3) can be expressed as

$$\mathbf{B}\,\mathbf{A}\,S_q^{\alpha}\mathbf{e}_{\alpha} = S_q^{\alpha}\mathbf{B}\,A_{\alpha}^{i}\mathbf{e}_i = S_q^{\alpha}\,A_{\alpha}^{i}\,B_i^{\beta}\mathbf{e}_{\beta} = S_q^{\beta}\mathbf{e}_{\beta}$$

and

$$\mathbf{B}\,C_p^{\perp i}\mathbf{e}_i = C_p^{\perp i}\,B_i^{\beta}\mathbf{e}_{\beta} = D_p^{\beta}\mathbf{e}_{\beta},$$

or as

$$B_i^{\beta}\,(A_{\alpha}^{i}\,S_q^{\alpha} \;\vdots\; C_p^{\perp i})\,\mathbf{e}_{\beta} = (S_q^{\beta} \;\vdots\; D_p^{\beta})\,\mathbf{e}_{\beta},$$

which gives in matrix notation

$$\underset{n\times m}{B}\,(\underset{m\times n}{A}\,\underset{n\times r}{S} \;\vdots\; \underset{m\times(m-r)}{C^{\perp}}) = (\underset{n\times r}{S} \;\vdots\; \underset{n\times(m-r)}{D}). \tag{2.6}$$

Now, since the subspaces $\mathcal{R}(\mathbf{A}) = \mathbf{A}S$ and $\mathcal{C}^o$ are complementary, the mxm matrix $(AS \;\vdots\; C^{\perp})$ has full rank and is thus invertible. The unique representation of a particular generalized inverse B of **A** therefore becomes

$$\boxed{\underset{n\times m}{B} = (\underset{n\times r}{S} \;\vdots\; \underset{n\times(m-r)}{D})(\underset{m\times r}{AS} \;\vdots\; \underset{m\times(m-r)}{C^{\perp}})^{-1}} \tag{2.7}$$

A more symmetric representation is obtained if we substitute the easily verified matrix identity

$$(AS \;\vdots\; C^{\perp})^{-1} = \begin{pmatrix} (C^tAS)^{-1}C^t \\ ((U^{\perp})^tC^{\perp})^{-1}(U^{\perp})^t \end{pmatrix},$$

with $\mathcal{U}^o = \mathcal{R}(\mathbf{A})^o = \mathcal{N}u(\mathbf{A}^*)$, into (2.7) (recall that $C^{\perp}$ and $U^{\perp}$ are matrix representations of respectively the subspaces $\mathcal{C}^o$ and $\mathcal{U}^o$):

$$\boxed{\underset{n\times m}{B} = S(\underset{n\times m}{C^tAS})^{-1}C^t + D\,((\underset{n\times m}{U^{\perp}})^tC^{\perp})^{-1}(U^{\perp})^t} \tag{2.8}$$

With (2.7) or (2.8) we thus have found **one** expression which covers **all** the generalized inverses of **A**. Furthermore we have the important result that each particular generalized inverse of **A** ,defined through (2.2) and (2.3), is uniquely characterized by the choices made for the subspaces $\mathcal{S}$, complementary to $\mathcal{N}u(\mathbf{A})$, $\mathcal{C}^o$ complementary to $\mathcal{R}(\mathbf{A})$ and $\mathcal{D}$, a subspace of $\mathcal{N}u(\mathbf{A})$.

In the next two sections we will give the interpretation associated with the three subspaces S, $\mathcal{C}^0$ and $\mathcal{D}$. Also the relation with the problem of solving an arbitrary system of linear equations will become clear then.

3. Injective and surjective maps

From the theorem of the previous section we learnt that the inverse-like properties

$$\mathbf{B}_{|R(\mathbf{A})}\,\mathbf{A}_{|S} = \mathbf{I} \quad \text{and} \quad \mathbf{A}_{|S}\,\mathbf{B}_{|R(\mathbf{A})} = \mathbf{I} \tag{3.1}$$

hold for any arbitrary generalized inverse **B** of **A**. That is, the maps **BA** and **AB** behave like identity maps on respectively the subspaces $S \subset N$ and $R(\mathbf{A}) \subset M$. Thus in the special case that rank **A** = r = n, the generalized inverses of **A** become left-inverses, since then **BA** = **I**. And similarly they become right-inverses if rank **A** = r = m, because then **AB** = **I** holds.

In order to give an interpretation of the subspace $S \subset N$, let us now first concentrate on the special case that rank **A** = r = m.

If rank **A** = r = m then $R(\mathbf{A}) = M$, which implies that the subspaces complementary to $R(\mathbf{A})$ reduce to $\mathcal{C}^0 = \{\mathbf{0}\}$. With (2.5) we then also have that $\mathcal{D} = \{\mathbf{0}\}$ (see figure 9). The general expression of right-inverses therefore readily follows from (2.8) as

$$\mathbf{B}_{n\times m} = \mathbf{S}_{n\times m}\,(\mathbf{AS})^{-1}_{m\times m}, \text{ with } N = S \oplus Nu(\mathbf{A}) \tag{3.2}$$

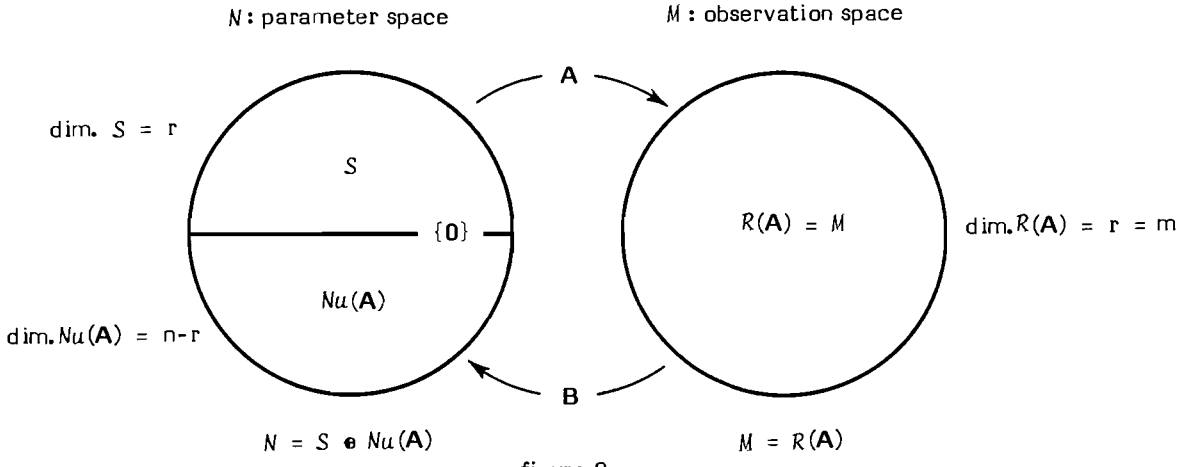


figure 9

Thus the only subspaces which play a role in the inverses of surjective maps are the subspaces S complementary to $Nu(\mathbf{A})$.

In order to find out how (3.2) is related to the problem of solving a system of linear equations

$$\underset{m\times 1}{y} = \underset{m\times n}{A}\ \underset{n\times 1}{x}, \tag{3.3}$$

for which matrix $\underset{m\times n}{A}$ has full row rank m, first observe that the system is consistent for all $\underset{m\times 1}{y} \in R^m$. With a particular generalized inverse (right-inverse), say $\underset{n\times m}{B}$, of $\underset{m\times n}{A}$, and $V^o = Nu(\mathbf{A})$, the solution set of (3.3), which actually represents a linear manifold in N, can therefore be written as

$$\{\underset{n\times 1}{x}\} = \{\underset{n\times 1}{x} \mid \underset{n\times 1}{x} = \underset{n\times 1}{By} + \underset{n\times(n-r)}{V^{\perp}}\ \underset{(n-r)\times 1}{\alpha}\} .$$

By choosing α, say $\alpha := \alpha_1$, we get thus as a particular solution $x_1 \in \{x\}$:

$$\underset{n\times 1}{x_1} = \underset{n\times 1}{By} + \underset{n\times 1}{V^{\perp}\alpha_1}, \tag{3.4}$$

where α_1 so to say contributes the extra information, which is lacking in y, to determine x_1. Since $R(\mathbf{B}) = S$, it follows from (3.4) that

$$\underset{(n-r)\times n}{(S^{\perp})^t}\ \underset{n\times 1}{x_1} = \underset{(n-r)\times(n-r)}{[(S^{\perp})^t\ V^{\perp}]}\ \underset{(n-r)\times 1}{\alpha_1} \overset{call}{=} \underset{(n-r)\times 1}{c_1} . \tag{3.5}$$

But this means that, since α_1 or c_1 contributes the extra information which is lacking in y to determine x_1, equation (3.5) and (3.3) together suffice to determine x_1 uniquely. Or in other words, the solution of the uniquely solvable system

$$\underset{(m+n-r)\times 1}{\begin{bmatrix} y \\ c_1 \end{bmatrix}} = \underset{(m+n-r)\times n}{\begin{bmatrix} A \\ (S^{\perp})^t \end{bmatrix}}\ \underset{n\times 1}{x} \tag{3.6}$$

is precisely x_1:

$$\underset{n\times 1}{x_1} = \underset{n\times(m+n-r)}{\begin{bmatrix} A \\ (S^{\perp})^t \end{bmatrix}^{-1}}\ \underset{(m+n-r)}{\begin{bmatrix} y \\ c_1 \end{bmatrix}} = \underset{n\times m}{[S(AS)^{-1}}\ \vdots\ \underset{n\times(n-r)}{V^{\perp}[(S^{\perp})^t V^{\perp}]^{-1}}]\underset{(m+n-r)\times 1}{\begin{bmatrix} y \\ c_1 \end{bmatrix}} , \tag{3.7}$$

with $V^o = Nu(\mathbf{A})$.

Thus we have recovered the rule, that in order to find a particular solution to (3.3), say x_1, we merely

need to extend the system of linear equations from (3.3) to (3.6) by introducing the additional equations $c_1 = (S^{\perp})^t x$, so that the extended matrix

$$\underset{(m+n-r)\times n}{\begin{bmatrix} A \\ (S^{\perp})^t \end{bmatrix}}$$

becomes square and regular. Futhermore the corresponding right-inverse of **A** is obtainable from the inverse of this extended matrix.

Let us now consider the case rank **A** = r = n. Then all generalized inverses of **A** become left-inverses. Because of the injectivity of **A** we have that its nullspace reduces to $Nu(\mathbf{A}) = \{\mathbf{0}\}$. But this implies that $S = N$ and $\mathcal{D} = \{\mathbf{0}\}$, since $\mathcal{D} \subset Nu(\mathbf{A})$. (see figure 10).

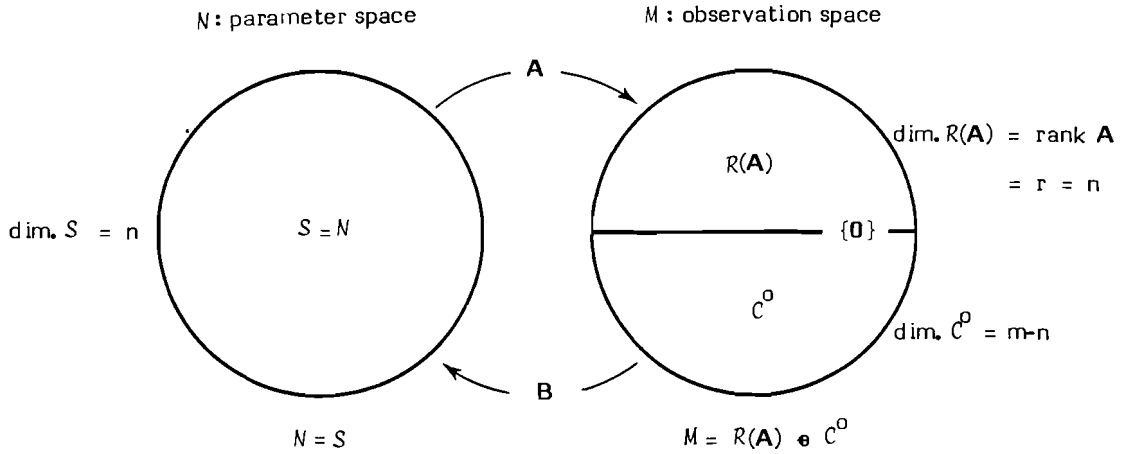


figure 10

For the dual map $\mathbf{A}^*: M^* \to N^*$ we therefore have a situation which is comparable to the one sketched in figure 9 (see figure 11). Now, taking advantage of our result (3.2), we find the general matrix-representation of an arbitrary generalized inverse $\mathbf{B}^*$ of $\mathbf{A}^*$ to be

$$\underset{m\times n}{B^t} = \underset{m\times n}{C}\,(\underset{n\times n}{A^t C})^{-1}.$$

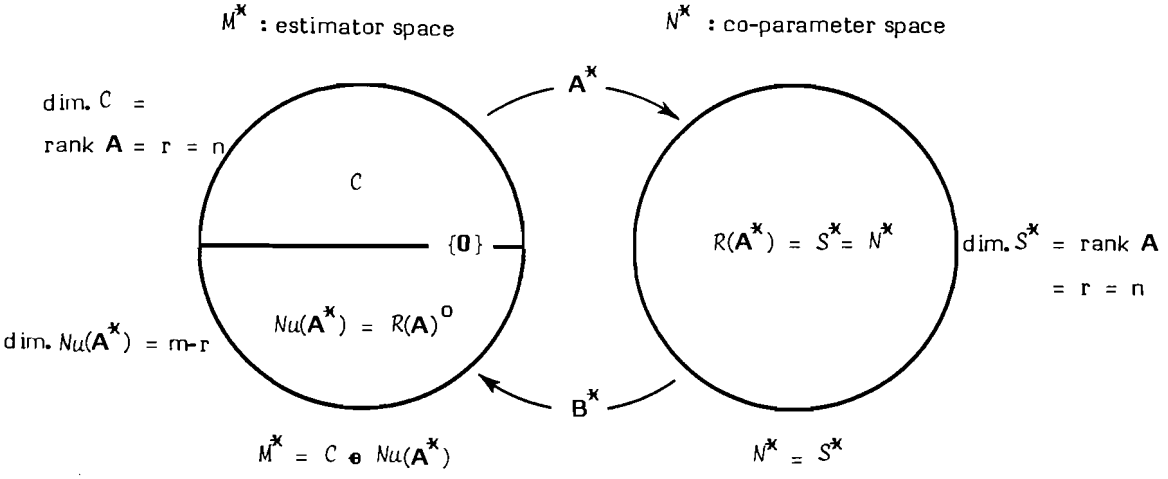


figure 11

The general expression of left-inverses therefore readily follows as

$$\underset{n\times m}{B} = \underset{n\times n}{(C^t A)^{-1}}\,\underset{n\times m}{C^t}\,, \quad \text{with} \quad M = R(A) \oplus C^o \tag{3.8}$$

Thus dual to our result (3.2), we find that the only subspaces which play a rôle in the inverses of injective maps, are the subspaces C^o complementary to $R(A)$.
With the established duality relations it now also becomes easy to see how (3.8) is related to the problem of solving a generally inconsistent but otherwise uniquely determined system of linear equations

$$\underset{m\times 1}{y} \doteq \underset{m\times n}{A}\ \underset{n\times 1}{x}\,, \quad \text{with rank } A = r = n. \tag{3.9}$$

The dual of (3.6) modified to our present situation gives namely

$$\underset{m\times 1}{y} = \underset{m\times n}{(A}\ \vdots\ \underset{m\times(m-r)}{C^{\perp})} \begin{bmatrix} x \\ \lambda \end{bmatrix}_{(m+n-r)\times 1}. \tag{3.10}$$

And dual to (3.7), the unique solution of (3.10) is given by:

$$\begin{bmatrix} x \\ \lambda \end{bmatrix}_{(n+m-r)\times 1} = \underset{(n+m-r)\times m}{(A\ \vdots\ C^{\perp})^{-1}}\ \underset{m\times 1}{y} = \begin{bmatrix} (C^t A)^{-1} C^t \\ \left[(U^{\perp})^t C^{\perp}\right]^{-1} (U^{\perp})^t \end{bmatrix}_{(n+m-r)\times m} \underset{m\times 1}{y}\,, \tag{3.11}$$

with $U^o = Nu(A^*)$.

We therefore have recovered the dual rule that in order to find a particular solution to (3.9), we need to extend the system of linear equations from (3.9) to (3.10) by introducing additional unknowns such that the extended matrix

$$\underset{m\times n}{(A}\ \vdots\ \underset{m\times(m-r)}{C^{\perp}}) \tag{3.12}$$

becomes square and regular. Furthermore the corresponding left-inverse of A is obtainable from the inverse of this extended matrix.

4. Arbitrary systems of linear equations and arbitrary inverses

In the previous section we showed that a particular solution of an underdetermined but otherwise consistent system of linear equations could be obtained by extending the matrix $\underset{m\times n}{A}$ rowwise. And especially the principal rôle played by the subspace $S \subset N$ complementary to $Nu(\underset{m\times n}{A})$ in removing the underdeterminability was demonstrated. Similarly we saw how consistency of an inconsistent, but otherwise uniquely determined system of linear equations was restored by extending the matrix $\underset{m\times n}{A}$ columnwise. And here the subspace $C^o \subset M$ complementary to $R(A)$ played the decisive rôle. We also observed a complete duality between these results; for the dual of an injective map is surjective and vice versa.

These results are, however, still not general enough. In particular we note that the subspace $\mathcal{D} \subset Nu(A)$ was annihilated as a consequence of the assumed injectivity and surjectivity. The reason for this will become clear if we consider the interpretation associated with the subspace $\mathcal{D}$. Since $S \cap \mathcal{D} = \{0\}$ it follows from expression (2.8) that $R(B) = S \oplus \mathcal{D}$. With dim.$S$ = dim $R(A)$ = rank **A** we therefore have that rank **B** $\geq$ rank **A**, with equality if and only if $\mathcal{D} = \{0\}$. But this shows why the subspace $\mathcal{D}$ gets annihilated in case of injective and surjective maps. The left (right) inverses have namely the same rank as the injective (surjective) maps. From the above it also becomes clear that the rank of **B** is completely determined by the choice made for $\mathcal{D}$. In particular **B** will have minimum rank if $\mathcal{D}$ is chosen to be $\mathcal{D} = \{0\}$, and maximum rank, rank **B** = min.(m,n), if one can choose $\mathcal{D}$ such that dim. $\mathcal{D}$ = min.(m,n)-r. Now to see how the subspace $\mathcal{D} \subset Nu(A)$ gets incorporated in the general case, we consider a system of linear equations

$$\underset{m\times 1}{y} \doteq \underset{m\times n}{A}\ \underset{n\times 1}{x}\ , \text{ with rank } \mathbf{A} = r < \min.(m,n), \tag{4.1}$$

i.e. a system which is possibly inconsistent and underdetermined at the same time. From the rank-deficiency of **A** in (4.1) follows that the unknowns **x** cannot be determined uniquely, even if $\mathbf{y} \in R(\mathbf{A})$. Thus the information contained in **y** is not sufficient to determine **x** uniquely. Following the same approach as before, we can at once remove this underdeterminability by extending (4.1) to

$$\underset{(m+n-r)\times 1}{\begin{bmatrix} y \\ c \end{bmatrix}} \doteq \underset{(m+n-r)\times n}{\begin{bmatrix} A \\ (S^{\perp})^t \end{bmatrix}}\ \underset{n\times 1}{x}\ , \text{ with } N = S \oplus Nu(A)\ . \tag{4.2}$$

But although the extended matrix of (4.2) has full column rank, the system can still be inconsistent. To remove possible inconsistency we therefore have to extend the matrix of (4.2) columnwise so that the resulting matrix becomes square and regular. Now since $M = R(A) \oplus C^o$, the following extension is a feasible one:

$$\underset{(m+n-r)\times 1}{\begin{bmatrix} y \\ c \end{bmatrix}} = \underset{(m+n-r)\times(m+n-r)}{\begin{bmatrix} A & C^{\perp} \\ (S^{\perp})^t & 0 \end{bmatrix}} \underset{(m+n-r)\times 1}{\begin{bmatrix} x \\ \lambda \end{bmatrix}}\ , \text{ with } M = R(A) \oplus C^o\ .$$

But the most general extension would be

$$\underset{(m+n-r)\times 1}{\begin{bmatrix} y \\ c \end{bmatrix}} = \underset{(m+n-r)\times(m+n-r)}{\begin{bmatrix} A & C^{\perp} \\ (S^{\perp})^{t} & X \end{bmatrix}} \underset{(m+n-r)\times 1}{\begin{bmatrix} x \\ \lambda \end{bmatrix}}, \tag{4.3}$$

with $N = S \oplus Nu(\mathbf{A})$, $M = R(\mathbf{A}) \oplus C^{o}$ and $\underset{(n-r)\times(m-r)}{X}$ being arbitrary. The unique solution of (4.3) is then given by:

$$\begin{bmatrix} x \\ \lambda \end{bmatrix} = \begin{bmatrix} A & C^{\perp} \\ (S^{\perp})^{t} & X \end{bmatrix}^{-1} \begin{bmatrix} y \\ c \end{bmatrix} =$$

$$= \left[\begin{array}{c:c} S(C^{t}AS)^{-1}C^{t} - V^{\perp}\left[(S^{\perp})^{t}V^{\perp}\right]^{-1}X\left[(U^{\perp})^{t}C^{\perp}\right]^{-1}(U^{\perp})^{t} & V^{\perp}\left[(S^{\perp})^{t}V^{\perp}\right]^{-1} \\ \hline \left[(U^{\perp})^{t}C^{\perp}\right]^{-1}(U^{\perp})^{t} & 0 \end{array}\right] \begin{bmatrix} y \\ c \end{bmatrix}, \tag{4.4}$$

with $N = S \oplus Nu(\mathbf{A})$, $M = R(\mathbf{A}) \oplus C^{o}$, $V^{o} = Nu(\mathbf{A})$ and $U^{o} = Nu(\mathbf{A}^{*})$.
In this expression we recognize, if we put $-V^{\perp}\left[(S^{\perp})^{t}V^{\perp}\right]^{-1}X = D$ or $X = -(S^{\perp})^{t}D$, our general matrix representation (2.8) of an arbitrary generalized inverse **B** of **A**. Thus as a generalization of (3.7) and (3.11) we have:

$$\begin{bmatrix} x \\ \lambda \end{bmatrix} = \begin{bmatrix} A & C^{\perp} \\ (S^{\perp})^{t} & -(S^{\perp})^{t}D \end{bmatrix}^{-1} \begin{bmatrix} y \\ c \end{bmatrix} = \left[\begin{array}{c:c} S(C^{t}AS)^{-1}C^{t} + D\left[(U^{\perp})^{t}C^{\perp}\right]^{-1}(U^{\perp})^{t} & V^{\perp}\left[(S^{\perp})^{t}V^{\perp}\right]^{-1} \\ \hline \left[(U^{\perp})^{t}C^{\perp}\right]^{-1}(U^{\perp})^{t} & 0 \end{array}\right] \begin{bmatrix} y \\ c \end{bmatrix},$$

with $V^{o} = Nu(\mathbf{A})$ and $U^{o} = Nu(\mathbf{A}^{*})$.

(4.5)

This result then completes the circle. In section one namely, we started by describing the geometric principles behind inverse linear mapping. In section two these principles were made precise by the stated theorem. This theorem enabled us to find a unique representation concerning all generalized inverses **B** of a linear map **A**. In section three we then specialized to injective and surjective maps, showing the relation between the corresponding inverses and the solutions of the corresponding systems of linear equations. And finally this section generalized these results to arbitrary systems of linear equations whereby our general expression of generalized inverses was again obtained.

5. Some common type of inverses and their relation to the subspaces S, C and $\mathcal{D}$

With our interpretation of the three subspaces S, C and $\mathcal{D}$, and an expression like (2.8) it now becomes very simple indeed to derive most of the standard results which one can find in the many textbooks available. See e.g. (Rao and Mitra, 1971). As a means of exemplification we show what rôle is played by the three subspaces S, C and $\mathcal{D}$ in the more common type of inverses used:

-- least-squares inverses --

Let M be Euclidean with metric tensor $\langle .,. \rangle_M$ and let $\mathbf{Q}_y : M^* \to M$ be the covariance map defined by $\mathbf{Q}_y^{-1} \mathbf{y} = \langle \mathbf{y}, . \rangle_M$.
We know from chapter one that for

$$\hat{\mathbf{x}} = \mathbf{B}\,\mathbf{y}$$

to be a least-squares solution of $\min_{\mathbf{x}} . \langle \mathbf{y}-\mathbf{A}\,\mathbf{x}, \mathbf{y}-\mathbf{A}\,\mathbf{x} \rangle_M$,

$$\mathbf{A}\,\mathbf{B} = \mathbf{P}_{U,U^\perp}, \quad \text{with} \quad U = R(\mathbf{A}), \tag{5.1}$$

must hold. From (2.8) follows, however, that in general

$$\mathbf{A}\,\mathbf{B} = \mathbf{P}_{U,C^o}, \quad \text{with} \quad U = R(\mathbf{A}). \tag{5.2}$$

Namely, expression (2.8) shows that

$$\underset{m\times m}{\mathbf{A}\,\mathbf{B}} = \underset{m\times m}{\mathbf{A}\,S(C^t A\,S)^{-1}C^t}. \tag{5.3}$$

And since

$$\underset{m\times m}{\mathbf{A}\,S(C^t A\,S)^{-1}C^t}.\ \underset{m\times(m-r)}{C^\perp} = \underset{m\times(m-r)}{0},$$

and

$$\underset{m\times m}{\mathbf{A}\,S(C^t A\,S)^{-1}C^t}.\underset{m\times r}{\mathbf{A}\,S} = \underset{m\times r}{\mathbf{A}\,S},$$

it follows that (5.3) is the matrix representation of the projector $\mathbf{P}_{U,C^o}$. From comparing (5.1) and (5.2) we thus conclude that least-squares inverses are obtained by choosing

$$\boxed{C^o = U^\perp = R(\mathbf{A})^\perp}, \tag{5.4}$$

while S and $\mathcal{V}$ may still be chosen arbitrarily. In matrices condition (5.4) reads

$$\underset{mx(m-r)}{C^{\perp}} = \underset{mxm}{Q_y} \; \underset{mx(m-r)}{U^{\perp}} \quad . \tag{5.5}$$

-- minimum norm inverses --

Let N be Euclidean with metric tensor $\langle .,. \rangle_N$ and let $\mathbf{Q}_x : N^* \to N$ be the covariance map defined by $\mathbf{Q}_x^{-1}\mathbf{x} = \langle \mathbf{x}, . \rangle_N$.
For

$$\hat{\mathbf{x}} = \mathbf{B}\,\mathbf{y}$$

to be the minimum norm solution of $\min_{\mathbf{x}}. \langle \mathbf{x},\mathbf{x} \rangle_N$ subject to $\mathbf{y} = \mathbf{A}\,\mathbf{x}$, $\hat{\mathbf{x}}$ must be the orthogonal projection of the origin onto the linear manifold specified by $\mathbf{y} = \mathbf{A}\,\mathbf{x}$. Hence,

$$\mathbf{B}\,\mathbf{A} = \underset{(V^o)^{\perp}, V^o}{\mathbf{P}}, \quad \text{with } V^o = Nu(\mathbf{A}), \tag{5.6}$$

must hold. With the same reasoning as above we then find that the minimum norm inverses are obtained by choosing

$$\boxed{S = V^{o\perp} = Nu\,(\mathbf{A})^{\perp}} \quad , \tag{5.7}$$

while C^o and $\mathcal{V}$ may still be chosen arbitrarily. In matrices condition (5.7) reads

$$\underset{nxr}{S} = \underset{nxn}{Q_x^{-1}} \; \underset{nxr}{V} \quad . \tag{5.8}$$

Note that since (5.7) implies that $S^o = R\,(\mathbf{A}^*)^{\perp}$, (5.4) and (5.7) are dually related.

-- maximum- and minimum rank inverses --

In the previous section we already indicated that by varying the choices for $\mathcal{V} \subset Nu\,(\mathbf{A})$ one could manipulate the rank of the corresponding generalized inverse. Inverses with maximum rank min.(m,n) were obtained if one could choose $\mathcal{V}$ such that $\dim.\,\mathcal{V} = \min.(m,n)-r$, and minimum rank inverses were characterized by the choice $\mathcal{V} = \{\mathbf{0}\}$.
As we will see in the next section the minimum rank inverses are by far the most important for statistical applications.
There is an interesting transformation property involved in the class of minimum rank inverses, which enables one to transform from an arbitrary inverse to a prespecified minimum rank inverse. To see this, recall that a minimum rank inverse, $\mathbf{B}_1$ say, of $\mathbf{A}$, which is uniquely characterized by the choices S_1, C_1^o and $\mathcal{V}_1 = \{\mathbf{0}\}$, satisfies the conditions

$$
\left.
\begin{aligned}
& B_1 A\, x = x, \quad \forall\, x \in S_1; \quad B_1 C_1^o = \{0\}\ , \\
& \text{with} \\
& N = S_1 \oplus V^o = B_1 R(A) \oplus V^o, \quad M = U \oplus C_1^o = A\, S_1 \oplus Nu(B_1) \\
& \text{and} \\
& U = R(A), \quad V^o = Nu(A)\ .
\end{aligned}
\right\} \tag{5.9}
$$

And it can be represented as

$$
B_1 = S_1 (C_1^t A\, S_1)^{-1} C_1^t . \tag{5.10}
$$

But the linear map **A** itself also satisfies similar conditions. For an arbitrary generalized inverse, **B** say, of **A** we have namely

$$
\left.
\begin{aligned}
& A\, B\, y = y, \quad \forall\, y \in U; \quad A\, V^o = \{0\}\ , \\
& \text{with} \\
& M = U \oplus C^o = A\, R(B) \oplus C^o, \quad N = S \oplus V^o = B\, U \oplus Nu(A) \\
& \text{and} \\
& U = R(A), \quad V^o = Nu(A)\ .
\end{aligned}
\right\} \tag{5.11}
$$

Upon comparing (5.11) with (5.9) we therefore conclude that the linear map **A** is representable in a way similar to that of B_1 in (5.10), i.e.

$$
\underset{m\times n}{A} = \underset{m\times n}{U}(\underset{m\times n}{V^t B\, U})^{-1} V^t\ , \tag{5.12}
$$

with $U = R(A)$, $V = Nu(A)^o$ and where **B** may be any arbitrary inverse of **A**. Now, substitution of (5.12) into (5.10) gives

$$
\begin{aligned}
& \underset{n\times m}{B_1} = S_1 (C_1^t \underset{n\times m}{U(V^t B\, U)^{-1} V^t} S_1)^{-1} C_1^t \\
\text{or} \quad & \underset{n\times m}{B_1} = (S_1 \underset{n\times n}{(V^t S_1)^{-1} V^t}).\ \underset{n\times m}{B}\ .\ (U \underset{m\times m}{(C_1^t U)^{-1}} C_1^t).
\end{aligned}
$$

In this last expression we recognize the matrix representations of the projectors $P_{S_1, Nu(A)}$ and $P_{R(A), C_1^o}$. Thus we have found the transformation rule

$$
\boxed{B_1 = P_{S_1, Nu(A)} . B . P_{R(A), C_1^o}} \ , \tag{5.13}
$$

which shows how to obtain a prespecified minimum rank inverse from any arbitrary generalized inverse of **A**. Because of the reciprocal character of minimum rank inverses - **A** is namely again an inverse of its minimum rank inverses - they are often called **reflexive** inverses.

— minimum norm least-squares inverses —

The minimum norm least-squares solution

$$\hat{\mathbf{x}} = \mathbf{B}\ \mathbf{y} \tag{5.14}$$

of an inconsistent and underdetermined system of linear equations

$$\mathbf{y} \doteq \mathbf{A}\ \mathbf{x}, \quad \text{with rank } \mathbf{A} = r < \min.(m,n), \tag{5.15}$$

is defined as the solution for which $\hat{\mathbf{x}}$ is the minimum norm solution of

$$\hat{\mathbf{y}} = \mathbf{A}\ \mathbf{x}\ , \tag{5.16}$$

and $\hat{\mathbf{y}}$ is the least-squares solution of (5.15).
Since the minimum norm solution of (5.16) is given by

$$\hat{\mathbf{x}} = \bar{\mathbf{B}}\ \hat{\mathbf{y}}\ , \tag{5.17}$$

where the inverse $\bar{\mathbf{B}}$ of $\mathbf{A}$ is characterized by (5.7), and the least-squares solution of (5.15) is given by

$$\hat{\mathbf{y}} = \mathbf{P}_{U,U^{\perp}}\ \mathbf{y}, \quad \text{with} \quad U = R(\mathbf{A}), \tag{5.18}$$

it follows from the combination of (5.17) and (5.18) together with the transformation rule (5.13), that the minimum norm least-squares inverse of $\mathbf{A}$ is uniquely characterized by

$$\boxed{S = V^{o\perp} = N\!u(\mathbf{A})^{\perp}, \quad C^{o} = U^{\perp} = R(\mathbf{A})^{\perp} \text{ and } \mathcal{D} = \{\mathbf{0}\}} \tag{5.19}$$

Note that since no freedom is left in choosing the three subspaces, the minimum norm least-squares inverse must be unique.
In the special case that both N and M are endowed with the ordinary canonical metric, the minimum norm least-squares inverse is commonly known as the **pseudo-inverse.**

— constrained inverses —

So far we have been careful in stating the complementarity conditions for $S \subset N$ and $C^{o} \subset M$. In the method of prolonging a matrix $\underset{m\times n}{A}$ this was reached by adding the minimum number of equations needed to the system $\underset{m\times 1}{y} \doteq \underset{m\times n}{A}\ \underset{n\times 1}{x}$ so that determinability of $\mathbf{x}$ was restored, and the minimum number of unknowns so that the prolonged matrix became square and regular, i.e. so that consistency was restored.

Sometimes, however, one can come across the situation where the system of linear equations $\underset{m\times 1}{y} \doteq \underset{m\times n}{A}\ \underset{n\times 1}{x}$ is appended with the restrictions $\underset{q\times n}{(T^{\perp})^{t}}\ \underset{n\times 1}{x} = \underset{q\times 1}{c}$, $q > n-r$. That is, with the restrictions that $\mathbf{x}$ should lie in a linear manifold parallel to a subspace T which is a proper subspace of an S, complementary to $Nu(\mathbf{A})$. In this case T thus fails to be complementary to $Nu(\mathbf{A})$. Although this situation differs from the ones we considered so far, it can be handled just as easy. By the method of prolongation we get namely

$$\underset{(m+q)\times 1}{\begin{bmatrix} y \\ c \end{bmatrix}} = \underset{(m+q)\ \times\ (m+q)}{\begin{bmatrix} A & C^{\perp} \\ (T^{\perp})^{t} & 0 \end{bmatrix}} \underset{(m+q)\times 1}{\begin{bmatrix} x \\ \lambda \end{bmatrix}}, \tag{5.20}$$

with $T \subset S$, $N = S \oplus Nu(\mathbf{A})$, $M = \mathbf{A}T \oplus C^{o}$.
The solution of (5.20) then follows as

$$\begin{bmatrix} x \\ \lambda \end{bmatrix} = \begin{bmatrix} T(C^{t}A\,T)^{-1}C^{t} & : & [I-T(C^{t}A\,T)^{-1}C^{t}A]T^{\perp}[(T^{\perp})^{t}T^{\perp}]^{-1} \\ \cdots & \cdots & \cdots \\ [((A\,T)^{\perp})^{t}C^{\perp}]^{-1}((A\,T)^{\perp})^{t} & : & -[((AT)^{\perp})^{t}C^{\perp}]^{-1}((A\,T)^{\perp})^{t}A\,T^{\perp}[(T^{\perp})^{t}T^{\perp}]^{-1} \end{bmatrix} \begin{bmatrix} y \\ c \end{bmatrix}, \tag{5.21}$$

where matrix $\underset{n\times m}{T(C^{t}A\,T)^{-1}C^{t}}$ is known as a **constrained** inverse of A (see e.g. Rao and Mitra, 1971). Other types of constrained inverses can be obtained in a similar way.

To conclude this section we have summarized, in order to facilitate reference, the basic results in table 1.

A , B $m \times n$ $n \times m$ rank $A = r_A$ rank $B = r_B$	INVERSES	INVERSELIKE PROPERTIES	SOLUTIONS OF LINEAR SYSTEMS OF EQUATIONS	
$m = n = r_A = r_B$	**inverse** $B = A^{-1}$	$AB = I_m$, $BA = I_n$	$y = Ax$	$x = A^{-1}y = By$
$m = r_A = r_B$	**right-inverse** $B = S(AS)^{-1}$ $N = S \oplus Nu(A)$	$AB = I_m$, $BA = P_{S, Nu(A)}$	$\begin{pmatrix} y \\ c \end{pmatrix} = \begin{pmatrix} A \\ \cdots \\ (S^{\perp})^t \end{pmatrix} x$	$x = \begin{pmatrix} A \\ \cdots \\ (S^{\perp})^t \end{pmatrix}^{-1} \begin{pmatrix} y \\ c \end{pmatrix} = \left(B \;\vdots\; V^{\perp}[(S^{\perp})^t V^{\perp}]^{-1} \right) \begin{pmatrix} y \\ c \end{pmatrix}$
$n = r_A = r_B$	**left-inverse** $B = (C^t A)^{-1} C^t$ $M = R(A) \oplus C^o$	$AB = P_{R(A), C^o}$, $BA = I_n$	$y = \begin{pmatrix} A & \vdots & C^{\perp} \end{pmatrix} \begin{pmatrix} x \\ \lambda \end{pmatrix}$	$\begin{pmatrix} x \\ \lambda \end{pmatrix} = \begin{pmatrix} A & \vdots & C^{\perp} \end{pmatrix}^{-1} y = \begin{pmatrix} B \\ \cdots \\ [(U^{\perp})^t C^{\perp}]^{-1} (U^{\perp})^t \end{pmatrix} y$
$r_A \leq \min.(m,n)$ $r_A \leq r_B \leq \min.(m,n)$	**generalized inverse** $B = S(C^t A S)^{-1} C^t + D[(U^{\perp})^t C^{\perp}]^{-1} (U^{\perp})^t$ $N = S \oplus Nu(A)$, $M = R(A) \oplus C^o$, $D \subset Nu(A)$	$AB = P_{R(A), C^o}$, $BA = P_{S, Nu(A)}$	$\begin{pmatrix} y \\ c \end{pmatrix} = \begin{pmatrix} A & \vdots & C^{\perp} \\ \cdots & \vdots & \cdots \\ (S^{\perp})^t & \vdots & -(S^{\perp})^t D \end{pmatrix} \begin{pmatrix} x \\ \lambda \end{pmatrix}$	$\begin{pmatrix} x \\ \lambda \end{pmatrix} = \begin{pmatrix} A & \vdots & C^{\perp} \\ \cdots & \vdots & \cdots \\ (S^{\perp})^t & \vdots & -(S^{\perp})^t D \end{pmatrix}^{-1} \begin{pmatrix} y \\ c \end{pmatrix} = \begin{pmatrix} B & \vdots & V^{\perp}[(S^{\perp})^t V^{\perp}]^{-1} \\ \cdots & \vdots & \cdots \\ [(U^{\perp})^t C^{\perp}]^{-1} (U^{\perp})^t & \vdots & O \end{pmatrix} \begin{pmatrix} y \\ c \end{pmatrix}$
$r_A \leq \min.(m,n)$ $\dim. T = n - q$ $r_B = n - q < r_A$	**constrained inverse** $B = T(C^t A T)^{-1} C^t$ $N = S \oplus Nu(A)$, $M = AT \oplus C^o$ $T \subset S$	$AB = P_{AT, C^o}$, $BA = P_{T, (A^* C)^o}$	$\begin{pmatrix} y \\ c \end{pmatrix} = \begin{pmatrix} A & \vdots & C^{\perp} \\ \cdots & \vdots & \cdots \\ (T^{\perp})^t & \vdots & O \end{pmatrix} \begin{pmatrix} x \\ \lambda \end{pmatrix}$	$\begin{pmatrix} x \\ \lambda \end{pmatrix} = \begin{pmatrix} A & \vdots & C^{\perp} \\ \cdots & \vdots & \cdots \\ (T^{\perp})^t & \vdots & O \end{pmatrix}^{-1} \begin{pmatrix} y \\ c \end{pmatrix} = \begin{pmatrix} B & \vdots & [I - T(C^t AT)^{-1} C^t A] T^{\perp}[(T^{\perp})^t T^{\perp}]^{-1} \\ \cdots & \vdots & \cdots \\ [((AT)^{\perp})^t C^{\perp}]^{-1} ((AT)^{\perp})^t & \vdots & -[((AT)^{\perp})^t C^{\perp}]^{-1} ((AT)^{\perp})^t AT^{\perp}[(T^{\perp})^t T^{\perp}]^{-1} \end{pmatrix} \begin{pmatrix} y \\ c \end{pmatrix}$

least-squares inverse : $C^o := U^{\perp}$ ($C^{\perp}_{m \times (m-r)} := Q_{m \times (m-r)} U^{\perp}$)

minimum norm inverse : $S := V^{o\perp}$ ($S := Q_{n \times r} V$)

reflexive inverse : $D := \{0\}$

minimum norm least-squares : $C^o := U^{\perp}$, $S := V^{o\perp}$, $D := \{0\}$

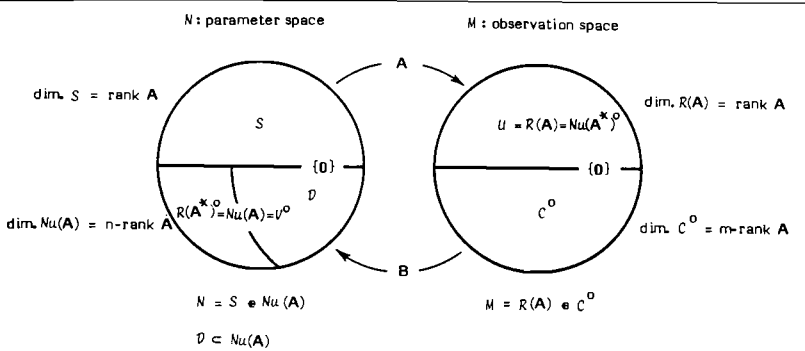


TABLE I

6. C- and S-transformations

Now that we have found a geometric characterization of the inverse linear mapping problem, let us return to the linear estimation problem which was considered in chapter I.
Consider the linear model

$$\tilde{y} \in \bar{N} \subset M, \quad \mathbf{Q}_y. \tag{6.1}$$

As we know (see (I.1.6)) the necessary and sufficiency conditions for the linear function $h(y) = \hat{a} + (\hat{y}^*, y)$ to be a linear unbiased estimator (LUE) of $(y_s^*, \tilde{y})$, are:

$$\begin{aligned} &\hat{a} = (y_s^* - \hat{y}^*, y_1) \text{ for some } y_1 \in \bar{N} \\ &\text{and} \\ &\hat{y}^* \in \{y_s^*\} + U^o. \end{aligned} \tag{6.2}$$

That is, $\hat{y}^*$ needs to be a point on the linear manifold $\{y_s^*\} + U^o$ in M^* (see figure 12).

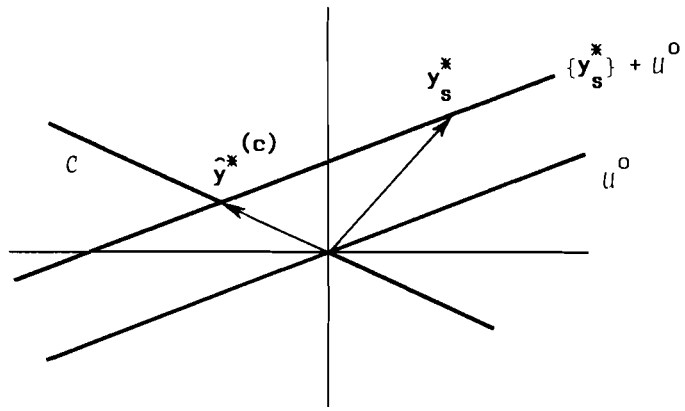


figure 12

It will be clear that every point $\hat{y}^*$ on this linear manifold can be obtained by choosing an appropriate subspace $C \subset M^*$ complementary to U^o and then projecting the 1-form y_s^* along the linear manifold $\{y_s^*\} + U^o$ onto C. Hence,

$$\hat{y}^{*(c)} = \mathbf{P}_{C, U^o} y_s^*,$$

or if $U = \mathbf{A}N$,

$$\hat{y}^{*(c)} = \mathbf{P}_{C, Nu(\mathbf{A}^*)} y_s^*. \tag{6.3}$$

With (6.2) then follows that the class of linear unbiased estimable functions of $(\mathbf{y}_s^*, \tilde{\mathbf{y}})$ is given by:

$$\begin{aligned} h(\mathbf{y}) &= (\mathbf{y}_s^* - \mathbf{P}_{C,\,Nu(\mathbf{A}^*)}\,\mathbf{y}_s^*, \mathbf{y}_1) + (\mathbf{P}_{C,\,Nu(\mathbf{A}^*)}\,\mathbf{y}_s^*, \mathbf{y}) = \\ &= (\mathbf{y}_s^*, \mathbf{y}_1) + (\mathbf{P}_{C,\,Nu(\mathbf{A}^*)}\,\mathbf{y}_s^*, \mathbf{y}-\mathbf{y}_1)\ , \end{aligned} \tag{6.4}$$

where $\bar{N} = \{\mathbf{y}_1\} + \mathbf{A}N$ and $C \subset M^*$ is arbitrary provided that $M^* = C \oplus Nu(\mathbf{A}^*)$. Every such linear function is thus uniquely characterized by the choice made for C. And by varying the choices for C one varies the type of unbiased estimator. Since the projector $\mathbf{P}_{C,\,Nu(\mathbf{A}^*)}$ always projects along the nullspace of $\mathbf{A}^*$ (see figure 13), we have that

$$\mathbf{P}_{C_i,\,Nu(\mathbf{A}^*)} \cdot \mathbf{P}_{C_j,\,Nu(\mathbf{A}^*)} = \mathbf{P}_{C_i,\,Nu(\mathbf{A}^*)} \cdot \tag{6.5}$$

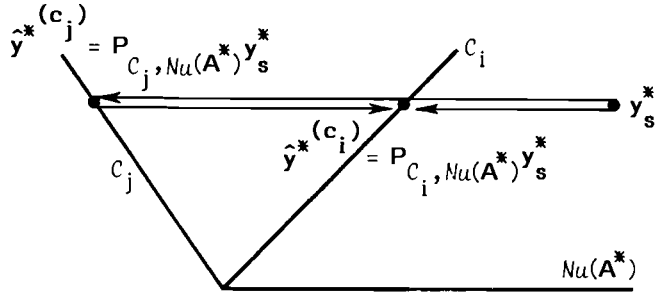


figure 13

The transformation between the corresponding 1-forms is therefore given by

$$\boxed{\hat{\mathbf{y}}^{*(c_i)} = \mathbf{P}_{C_i,\,Nu(\mathbf{A}^*)}\,\hat{\mathbf{y}}^{*(c_j)}}\ , \tag{6.6}$$

and in accordance with the current terminology one could call such transformations, C-**transformations.**

A typical example in which a particular choice for C is made can be found in the **method of averages** due to T. Mayer (Whittaker and Robinson, 1944, p. 258). In this method, which is sometimes used for polynomial approximations (see e.g. Morduchow and Levin, 1959), C is chosen such that the equations of a linear system $\underset{m\times 1}{\mathbf{y}} = \underset{m\times n}{\mathbf{A}}\ \underset{n\times 1}{\mathbf{x}}$ are separated into n groups and after that groupwise summed. Although more of such examples can be given, the most common applied estimator is of course the BLUE's estimator which is, as we know, characterized by the choice $C = Nu(\mathbf{A}^*)^{\perp}$. It is interesting to note though, that since every (oblique) projector can be interpreted as an orthogonal projector with respect to an appropriate metric tensor, every unbiased estimator can be interpreted as a BLUE's estimator with respect to an appropriate covariance map, a fact which was already

pointed out by (Baarda, 1967b, p. 34). To see this, assume that

$$U(C^tU)^{-1}C^t \quad , \text{ with } \quad U = R(A) \text{ and } \quad M = R(A) \oplus C^o ,$$

is a matrix representation of the oblique projector $P_{R(A),C^o}$.
With the symmetric and positive-definite matrix

$$Q = C^\perp[(U^\perp)^tC^\perp]^{-1}(U^\perp)^t(U^\perp)[(C^\perp)^tU^\perp]^{-1}(C^\perp)^t + \\ + U[C^tU]^{-1}C^tC[U^tC]^{-1}U^t ,$$

or

$$Q^{-1} = C(C^tC)^{-1}C^t + U^\perp[(U^\perp)^tU^\perp]^{-1}(U^\perp)^t$$

follows then that

$$U(C^tU)^{-1}C^t = U[U^tQ^{-1}U]^{-1}U^tQ^{-1}.$$

Thus the problem of comparing different unbiased estimators can in principle be restricted to the problem of analyzing the effect of assumptions on the metric tensor. See e.g. (Krarup, 1972).

Now let us assume that we have picked one particular 1-form, $\hat{y}^{*(c)}$ say. It follows then from (6.4) that the corresponding unbiased estimate of $\tilde{y} \in \bar{N} \subset M$ is given by:

$$\hat{y}^{(c)} = y_1 + P^*_{C,\,Nu(A^*)}(y_s - y_1),$$

or

$$\hat{y}^{(c)} = y_1 + P_{R(A),C^o}(y_s - y_1), \text{ with } y_1 \in \bar{N} . \tag{6.7}$$

And since the problem of removing inconsistency is in the above context of linear estimation essentially the problem of finding the estimate $\hat{y}^{(c)}$, one could say that one has concluded the actual adjustment problem once $\hat{y}^{(c)}$ is computed. In practice, however, one often requires a parameter representation of $\hat{y}^{(c)} \in \bar{N}$. And here is thus where the actual inverse mapping problem enters. That is, in order to find a parameter representation of $\hat{y}^{(c)}$ one needs a particular pre- or inverse image of $\hat{y}^{(c)} - y_1 \in AN$ under **A**. By means of a generalized inverse, **B** say, of **A** such an inverse image is obtained as

$$\hat{x} = B(\hat{y}^{(c)} - y_1).$$

From the transformation rule (5.13) and (6.7) follows then that every inverse image of $\hat{y}^{(c)} - y_1$ under **A** can be written as

$$\hat{x}^{(s,c)} = P_{S,\,Nu(A)} \cdot \bar{B} \cdot P_{R(A),C^o}(y_s - y_1) , \tag{6.8}$$

with $N = S \oplus Nu(\mathbf{A})$ and $M^* = C \oplus Nu(\mathbf{A}^*)$, and where $\bar{\mathbf{B}}$ is allowed to be any arbitrary inverse of **A**. The estimate $\hat{\mathbf{x}}^{(s,c)}$ is thus uniquely characterized by the choices made for C and S. To understand what $\hat{\mathbf{x}}^{(s,c)}$ actually estimates, consider the following equivalencies:

$$(\mathbf{y}_s^*, \mathbf{y}_1) + (\hat{\mathbf{y}}^{*(c)}, \mathbf{y}-\mathbf{y}_1) \quad \text{is a LUE of}$$

$$(\mathbf{y}_s^*, \tilde{\mathbf{y}}),\ \tilde{\mathbf{y}}, \mathbf{y}_1 \in \bar{N} = \{\mathbf{y}_1\} + \mathbf{A}N\ ,\quad \forall\ \mathbf{y}_s^* \in M^*\ ; \quad \hat{=}$$

$$(\mathbf{y}_s^*, \mathbf{y}_1) + (\mathbf{P}_{C,\, Nu(\mathbf{A}^*)}\, \mathbf{y}_s^*, \mathbf{y}-\mathbf{y}_1) \quad \text{is a LUE of}$$

$$(\mathbf{y}_s^*, \mathbf{y}_1) + (\mathbf{y}_s^*, \tilde{\mathbf{y}}-\mathbf{y}_1),\ \tilde{\mathbf{y}}, \mathbf{y}_1 \in \bar{N} = \{\mathbf{y}_1\} + \mathbf{A}N\ ,\ \forall\ \mathbf{y}_s^* \in M^*\ ; \quad \hat{=}$$

$$(\mathbf{y}_s^*, \mathbf{y}_1) + (\mathbf{P}_{C,\, Nu(\mathbf{A}^*)}\, \mathbf{y}_s^*, \mathbf{y}-\mathbf{y}_1) \quad \text{is a LUE of}$$

$$(\mathbf{y}_s^*, \mathbf{y}_1) + (\mathbf{A}^*\mathbf{y}_s^*, \mathbf{x}),\ \tilde{\mathbf{y}}-\mathbf{y}_1 = \mathbf{A}\,\mathbf{x},\quad \forall\ \mathbf{y}_s^* \in M^*\ ; \quad \hat{=}$$

$$(\mathbf{P}_{C,\, Nu(\mathbf{A}^*)} \cdot \bar{\mathbf{B}}^*\mathbf{x}_s^*, \mathbf{y}-\mathbf{y}_1) \quad \text{is a LUE of}$$

$$(\mathbf{x}_s^*, \mathbf{x}),\quad \forall\ \mathbf{x}_s^* \in R(\mathbf{A}^*),\ \text{arbitrary inverses } \bar{\mathbf{B}}^* \text{ of } \mathbf{A}^*\ ; \quad \hat{=}$$

$$(\mathbf{P}_{C,\, Nu(\mathbf{A}^*)} \cdot \bar{\mathbf{B}}^* \cdot \mathbf{P}_{R(\mathbf{A}^*),\, S^o}\, \mathbf{x}_s^*, \mathbf{y}-\mathbf{y}_1) \quad \text{is a LUE of}$$

$$(\mathbf{P}_{R(\mathbf{A}^*),\, S^o}\, \mathbf{x}_s^*, \mathbf{x}),\quad \forall\ \mathbf{x}_s^* \in N^*\ ; \quad \hat{=}$$

$$(\mathbf{x}_s^*, \mathbf{P}_{S,\, Nu(\mathbf{A})} \cdot \bar{\mathbf{B}} \cdot \mathbf{P}_{R(\mathbf{A}),\, C^o}(\mathbf{y}-\mathbf{y}_1)) \quad \text{is a LUE of}$$

$$(\mathbf{x}_s^*, \mathbf{P}_{S,\, Nu(\mathbf{A})}\, \mathbf{x}),\quad \forall\ \mathbf{x}_s^* \in N^*\ .$$

In other words $\hat{\mathbf{x}}^{(s,c)}$ is an unbiased estimate of $\mathbf{x}^{(s)} = \mathbf{P}_{S,\, Nu(\mathbf{A})}\,\mathbf{x}$, but **not** of **x** itself. This subtle difference as to what $\hat{\mathbf{x}}^{(s,c)}$ actually estimates has sometimes been a source of confusion. See e.g. (Jackson, 1982).

Since the projector $\mathbf{P}_{S,\, Nu(\mathbf{A})}$ always projects along the nullspace of **A** (see figure 14) we have that

$$\mathbf{P}_{S_i,\, Nu(\mathbf{A})} \cdot \mathbf{P}_{S_j,\, Nu(\mathbf{A})} = \mathbf{P}_{S_i,\, Nu(\mathbf{A})}\ . \tag{6.9}$$

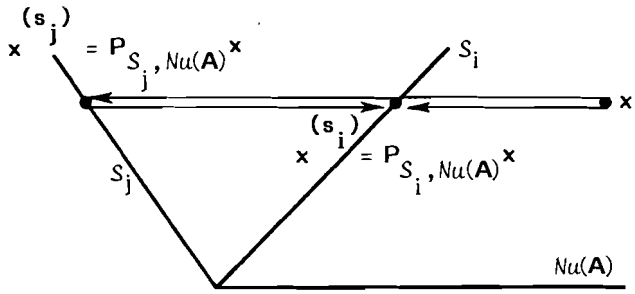


figure 14

The transformation between the various inverse images of $\hat{y}^{(c)} - y_1$ under **A** is therefore given by

$$\hat{x}^{(s_i,c)} = P_{S_i,\, Nu(A)}\, \hat{x}^{(s_j,c)} \,. \qquad (6.10)$$

Such transformations are now known as ***S*-transformations.** They were first introduced by Baarda in the context of free networks and used to obtain an invariant precision description of geodetic networks (see e.g. Baarda, 1973; Molenaar, 1981; Van Mierlo, 1979; or Teunissen, 1984a). Baarda has used the term "S-transformation", since the projector $P_{S,\, Nu(A)}$ is in case of geodetic networks derivable from the differential Similarity transformation. In the above general context, however, it would perhaps be more appropriate to call transformation (6.10) a Singularity transformation. This as opposed to the Consistency transformation (6.6).

Note the great resemblance between (6.6) and (6.10). From this comparison also follows the duality result that the *C*-transformations of **A** are the *S*-transformations of $\mathbf{A}^*$, or, the projector $P_{C,\, Nu(A^*)}$ is the *S*-transformation of $\mathbf{A}^*$ and the projector $P_{S,\, Nu(A)}$ is the *C*-transformation of $\mathbf{A}^*$.

In this section we have seen how the inverse linear mapping theory applies to the problem of linear estimation. We have seen that the actual problem of adjustment and the actual problem of inverse mapping, although dually related, are essentially two problems of a different kind. Were we only interested in adjustment, i.e. in removing inconsistency, then we would only be concerned with the subspace $C \subset M^*$. But if one, in addition to removing inconsistency, is also interested in finding a particular pre- or inverse image of $\hat{y}^{(c)} \in R(A)$ under **A**, then the choice of $S \subset N$ comes to the fore. We would like to stress here the importance of the definite ordering: first adjustment and then inverse mapping, since it shows that in an estimation context no great value should be attached to the subspace $\mathcal{D}$. In fact the only inverses of **A** which map y_s into the pre-image $\hat{x}^{(s,c)}$ of $\hat{y}^{(c)}$, are the minimum rank inverses ($\hat{=}\ \mathcal{D} = \{0\}$). And in particular one should be aware that one can not get an arbitrary pre-image $\hat{x}^{(s)} \in N$ of the least-squares estimate $\hat{y} = P_{R(A),\, R(A)^\perp} y_s$, by mapping $y_s \in M$ with an arbitrary least-squares inverse of **A** into N.

III. GEODETIC INVERSE MAPPING

1. Introduction

In the preceding chapter we have seen how to characterize an arbitrary inverse of a linear map $\mathbf{A}: N \rightarrow M$ uniquely. In particular we saw how by choosing C^o complementary to $R(\mathbf{A})$ one could make an inconsistent system of linear equations consistent, and how S complementary to $Nu(\mathbf{A})$ gave a way of restoring determinability. We also noted that although inconsistency and underdeterminability generally occur simultaneously if rank $\mathbf{A} = r < \min.(m,n)$, the actual problem of adjustment, i.e. the problem of removing inconsistency, and the actual problem of inverse mapping are essentially two problems of a different kind. They can therefore be dealt with separately.

In this chapter we will concentrate on the actual inverse mapping problem of geodetic networks. As an exemplification of the theory of S-transformations we discuss the non-uniqueness in coordinate system definitions and construct sets of base vectors for $Nu(\mathbf{A})$. We also discuss the related problem of connecting geodetic networks.
Section two is devoted to the inverse mapping problem and section three to the problem of connecting networks. In section two we discuss successively the planar, ellipsoidal- and three dimensional case. Although we recognize that the inverse mapping problem of two dimensional planar geodetic networks has already been discussed at length in the geodetic literature (see e.g. Teunissen, 1984a, and the references listed therein), we have reiterated some of the theory since it indicates very well the principles involved. Generalization to the ellipsoidal- and three dimensional case becomes then rather straightforward.
For practical ellipsoidal networks an interesting feature turns out to be the numerical ill-conditioning of the inverse map. One will find namely that even after the admitted degree of freedom of the ellipsoidal model is taken care of, the estimated geodetic coordinates of practical ellipsoidal networks still lack precision. As a consequence the estimation problem of the ellipsoidal model turns out to be not too different from that of the planar model.
In our discussion of three dimensional networks we make a distinction between local surveys and networks covering a large area. For local surveys (e.g. for the purpose of construction works), it is likely that one is only interested in describing the point configuration of the network. Therefore, for such networks S -transformations that only transform coordinates (and their co-variances) will do. As an example we have given an analytic expression of the three dimensional S-transformation advocated by (Baarda, 1979). For large networks however, it will not be sufficient to consider only the coordinate transforming S-transformations. In these cases one is almost surely also interested in a description of the fundamental directions like local verticals and the average terrestrial pole. That is, besides the network's point configuration also the configuration of the fundamental directions becomes of interest then. Hence, we also need S-transformations that transform both coordinates and orientation parameters.

Having given the various representations of $Nu(\mathbf{A})$ which are needed to derive the appropriate S-

transformations, we turn our attention in section three to the problem of connecting geodetic networks. Without exaggeration one can consider this problem of comparing and connecting overlapping pointfields to be almost omnipresent in geodesy. In cartography for instance, the problem occurs when digitized map material needs to be transformed to a well established known coordinate system such as a national system. And in photogrammetry when photogrammetric blocks need to be connected with terrestrial coordinate systems or in case of stripwise block adjustment when the various strips need to be connected (Molenaar, 1981b). Also in surveying practice where densification networks need to be tied to existing (often higher order) networks the connection problem appears repeatedly (Brouwer et.al., 1982). And on a more global scale when connecting satellite networks to national networks (Adam et.al., 1982). Even in case of gravity surveys one sometimes needs to connect networks, e.g. relative gravity networks to existing well established absolute gravity systems. And finally similar problems are encountered in deformation analysis (Van Mierlo, 1978). There networks measured at two or possibly more epochs need to be compared in order to affirm projected geophysical hypotheses.

In all the above cases the same principles for connecting networks can be applied although of course the elaboration can differ from application to application, depending e.g. on the information available and the purposes one needs to serve. That is, although different solution strategies exist, all methods rely on the self-evident principle that the only information suited for comparing networks, is the information which is **common** to both networks.

In our presentation we will discuss three methods for connecting geodetic networks. Although all three alternatives are considered to some extent in the geodetic literature, the treatment below accentuates some aspects which are not discussed elsewhere.

2. Geodetic networks and their degrees of freedom

2.1. Planar networks

Let us commence, in order to fix our minds, with the simple example of a two dimensional planar triangulation network in which only angles are measured (see e.g. figure 15).

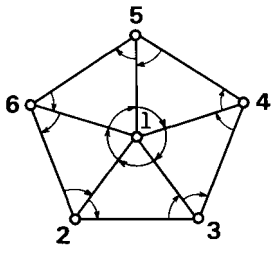


figure 15

After adjusting the network (using e.g. a first standard problem formulation) we obtain a consistent set of adjusted angles which determines the shape of the network. In order to describe this shape we have many possiblities at hand. Each set of mutually independent adjusted angles for instance, will do. In practice, however, one usually wants the result of an adjustment to be presented by means of coordinates, since they are more manageable than individual angles. The advantage of working with coordinates is namely that, once they are introduced, they all have **one and the**

same reference in common. The benefit being that with coordinates the relative position of any two points in a network is easily obtained without need to bother about the way in which these two network points are connected by the measured elements. Consequently, coordinates are very tractable for drawing maps or making profiles of the whole or parts of the network.

With this motivation in mind we are thus looking for a way to present our results of adjustment by means of (cartesian) coordinates.

However, in order to compute coordinates we first need to fix some reference, i.e. in the case of a planar triangulation network we need to fix the position, orientation and scale of the network. One way to accomplish this is of course by fixing two points of the network,i.e. by assigning arbitrary and non-stochastic coordinates to two points of the network. For instance, we can start by fixing the points P_1 and P_2 and then compute, with the aid of the adjusted angles, the coordinates of the points P_3, P_4, P_5 and P_6. Or, we can fix the points P_3 and P_1, and then compute the points P_4, P_5, P_6 and P_2. Let us for the moment leave in the middle which two points we fix. Let's just call them P_r and P_s. We then can write (see figure 16)

$$\begin{aligned} x_i &= x_r + l_{rs} \sin A_{rs} + l_{si} \sin (A_{rs} + \pi + \alpha_{rsi}) \\ y_i &= y_r + l_{rs} \cos A_{rs} + l_{si} \cos (A_{rs} + \pi + \alpha_{rsi}) \end{aligned} \qquad (2.1)$$

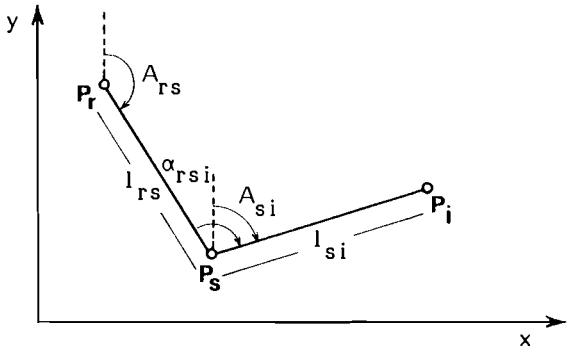


figure 16

Linearization of (2.1) gives (the upperindices "o" indicate the approximate values):

$$\begin{aligned} \Delta x_i &= \Delta x_r + x^o_{rs} \Delta \ln l_{rs} + y^o_{rs} \Delta A_{rs} + x^o_{si} \Delta \ln l_{si} + y^o_{si} \Delta A_{rs} + y^o_{si} \Delta \alpha_{rsi} \\ \Delta y_i &= \Delta y_r + y^o_{rs} \Delta \ln l_{rs} - x^o_{rs} \Delta A_{rs} + y^o_{si} \Delta \ln l_{si} - x^o_{si} \Delta A_{rs} - x^o_{si} \Delta \alpha_{rsi} \end{aligned} \qquad (2.2)$$

which we can write as

$$\begin{pmatrix} \Delta x_i \\ \Delta y_i \end{pmatrix} = \begin{pmatrix} y^o_{si} & x^o_{si} \\ -x^o_{si} & y^o_{si} \end{pmatrix} \begin{pmatrix} \Delta \alpha_{rsi} \\ \Delta \ln \frac{l_{si}}{l_{sr}} \end{pmatrix} + \begin{pmatrix} 1 & 0 & y^o_{ri} & x^o_{ri} \\ 0 & 1 & -x^o_{ri} & y^o_{ri} \end{pmatrix} \begin{pmatrix} \Delta x_r \\ \Delta y_r \\ \Delta A_{rs} \\ \Delta \ln l_{rs} \end{pmatrix} \qquad (2.3)$$

Since all the angular type of information is collected in the first term on the right-hand side of (2.3) we see that, in order to be able to introduce coordinates, we need to assign à priori values to the second term. One way is of course to take points P_r and P_s as reference- or base points by assigning to them the non-stochastic approximate coordinates x_r^o, y_r^o and x_s^o, y_s^o, i.e. by assuming that $\Delta x_r = \Delta y_r = \Delta A_{rs} = \Delta \ln l_{rs} = 0$ or

$$\Delta x_r = \Delta y_r = \Delta x_s = \Delta y_s = 0 \ . \tag{2.4}$$

The coordinates of any other point P_i of the network are then computed as

$$\begin{pmatrix} \Delta x \\ \Delta y \end{pmatrix}_i^{(r,s)} = \begin{pmatrix} y_{si}^o & x_{si}^o \\ -x_{si}^o & y_{si}^o \end{pmatrix} \begin{pmatrix} \Delta\alpha_{rsi} \\ \Delta \ln \frac{l_{si}}{l_{sr}} \end{pmatrix} \ , \tag{2.5}$$

where the upperindices (r,s) indicate that these coordinates are computed with respect to the basepoints P_r and P_s.

Although the choice of fixing the two points P_r and P_s in (2.3) is an obvious one, there are also other ways of introducing coordinates. One could for instance take two other points of the network as base points, or fix linear combinations of the coordinate increments of network points. Essential is, irrespective the choice made, that the positional-, orientational- and scale degrees of freedom of the network are taken care of. This is best seen by observing that (2.3) combined with (2.5) essentially constitutes the two dimensional differential similarity transformation:

$$\begin{pmatrix} \Delta x \\ \Delta y \end{pmatrix}_i = \begin{pmatrix} \Delta x \\ \Delta y \end{pmatrix}_i^{(r,s)} + \begin{pmatrix} 1 & 0 & y_i^o & x_i^o \\ 0 & 1 & -x_i^o & y_i^o \end{pmatrix} \begin{pmatrix} \Delta t_x \\ \Delta t_y \\ \Delta\phi \\ \Delta\kappa \end{pmatrix} \ , \tag{2.6}$$

which follows from linearizing

$$\begin{pmatrix} x \\ y \end{pmatrix}_i = \kappa \begin{pmatrix} \cos\phi & \sin\phi \\ -\sin\phi & \cos\phi \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix}_i^{(r,s)} + \begin{pmatrix} t_x \\ t_y \end{pmatrix} \tag{2.7}$$

under the assumptions that $\kappa^o = 1$, $\phi^o = 0$ and $t_x^o = t_y^o = 0$.

Since there are many different ways of introducing coordinates, it is important that one recognizes that in general

$$\begin{pmatrix} \Delta x \\ \Delta y \end{pmatrix}_i^{(r,s)} \neq \begin{pmatrix} \Delta x \\ \Delta y \end{pmatrix}_i^{(p,q)} \ .$$

Hence, if one wants to compare two sets of coordinates, where the two sets are computed from two different and independent observational campaigns - for instance for the purpose of a deformation analysis - it is essential that these coordinates are all defined with respect to the same reference. Now in order to get all coordinates in the same reference system one needs to be able to transform from one system to another.

For the above defined (r,s)-system this transformation is easily obtained.

From substituting

$$\begin{pmatrix} \Delta A_{rs} \\ \Delta \ln l_{rs} \end{pmatrix} = \frac{1}{(l^o_{rs})^2} \begin{pmatrix} y^o_{rs} & x^o_{rs} \\ -x^o_{rs} & y^o_{rs} \end{pmatrix} \begin{pmatrix} \Delta x_{rs} \\ \Delta y_{rs} \end{pmatrix} \quad , \tag{2.8}$$

into (2.3) follows namely with (2.5) the transformation rule

$$(\Delta x_i^{(r,s)} \quad \Delta y_i^{(r,s)})^t = (\Delta x_i \quad \Delta y_i)^t -$$

$$\begin{pmatrix} 1 - \frac{x^o_{ri}x^o_{rs} + y^o_{ri}y^o_{rs}}{(l^o_{rs})^2} & - \frac{x^o_{ri}y^o_{rs} - x^o_{rs}y^o_{ri}}{(l^o_{rs})^2} & \frac{x^o_{ri}x^o_{rs} + y^o_{ri}y^o_{rs}}{(l^o_{rs})^2} & \frac{x^o_{ri}y^o_{rs} - y^o_{ri}x^o_{rs}}{(l^o_{rs})^2} \\ \frac{x^o_{ri}y^o_{rs} - x^o_{rs}y^o_{ri}}{(l^o_{rs})^2} & 1 - \frac{x^o_{ri}x^o_{rs} + y^o_{ri}y^o_{rs}}{(l^o_{rs})^2} & \frac{x^o_{ri}y^o_{rs} - y^o_{ri}x^o_{rs}}{(l^o_{rs})^2} & \frac{x^o_{ri}x^o_{rs} + y^o_{ri}y^o_{rs}}{(l^o_{rs})^2} \end{pmatrix} \begin{pmatrix} \Delta x_r \\ \Delta y_r \\ \Delta x_s \\ \Delta y_s \end{pmatrix} , \tag{2.9}$$

which shows how to transform from an arbitrary coordinate system to the prespecified (r,s)-system.

To find the general procedure for deriving such transformations, note that the definition of the (r,s)-system and the derivation of (2.9) followed from the decomposition formula (2.3). With (2.5) and (2.8) this decomposition formula reads in matrix notation as

$$x = x^{(r,s)} + V^{\perp}\left[(S^{\perp}_{(r,s)})^t V^{\perp}\right]^{-1}(S^{\perp}_{(r,s)})^t x, \tag{2.10}$$

where

$$x = (\Delta x_r, \Delta y_r, \Delta x_s, \Delta y_s \ldots \Delta x_i, \Delta y_i \ldots)^t,$$

$$x^{(r,s)} = (\Delta x_r^{(r,s)}, \Delta y_r^{(r,s)}, \Delta x_s^{(r,s)}, \Delta y_s^{(r,s)} \ldots \Delta x_i^{(r,s)}, \Delta y_i^{(r,s)} \ldots)^t$$

and

$$V^{\perp}_{nx4} = \begin{pmatrix} \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ 1 & 0 & y_i & x_i \\ 0 & 1 & -x_i & y_i \\ \cdot & \cdot & \cdot & \cdot \end{pmatrix}, \quad S^{\perp}_{(r,s)\,nx4} = \begin{pmatrix} 1 & & & \\ & 1 & & \\ & & 1 & \\ 0 & 0 & 0 & 1 \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \end{pmatrix}.$$

Decomposition (2.10) is however, just one of the many possible decompositions of x. An alternative decomposition follows if we premultiply (2.10) by

$$S_i(V^tS_i)^{-1}V^t + V^{\perp}[(S_i^{\perp})^tV^{\perp}]^{-1}(S_i^{\perp})^t = I,$$

where $R(S_i)$ is arbitrary but complementary to $R(V^{\perp})$. We then get

$$x = S_i(V^tS_i)^{-1}V^tx^{(r,s)} + V^{\perp}[(S_i^{\perp})^tV^{\perp}]^{-1}(S_i^{\perp})^t[x^{(r,s)} + V^{\perp}[(S^{\perp}_{(r,s)})^tV^{\perp}]^{-1}(S^{\perp}_{(r,s)})^tx]$$

or

$$x = S_i(V^tS_i)^{-1}V^tx^{(r,s)} + V^{\perp}[(S_i^{\perp})^tV^{\perp}]^{-1}(S_i^{\perp})^tx\,. \tag{2.11}$$

And this expression decomposes x just like (2.10) into a first part, which contains all the angular type of information and a second part for which additional a priori information is needed. Now, just like decomposition (2.10) suggested to choose the restrictions (2.4), (2.11) suggests that we take

$$(S_i^{\perp})^tx = 0\,. \tag{2.12}$$

The coordinates of the network points are then computed as

$$x^{(s_i)} = S_i(V^tS_i)^{-1}V^tx^{(r,s)}, \tag{2.13}$$

where the upperindex (s_i) refers to the choice (2.12). And analogously to (2.10) we find from substituting (2.13) into (2.11) that

$$x = x^{(s_i)} + V^{\perp}[(S_i^{\perp})^tV^{\perp}]^{-1}(S_i^{\perp})^tx.$$

Hence the transformation to the (s_i)-system is given by

$$\boxed{x^{(s_i)} = [I - V^{\perp}[(S_i^{\perp})^tV^{\perp}]^{-1}(S_i^{\perp})^t]x} \tag{2.14}$$

This is the general expression one can use for deriving transformations like (2.9). We thus see that in order to derive such a transformation we only need to know $R(V^{\perp})$ and to choose an $(S_i^{\perp})^t_{4\times n}$ such that $R(S_i)$ is complementary to $R(V^{\perp})$.

So far we discussed planar networks of the angular type. But formula (2.14) is of course valid for other types of networks too. The only difference is that we need to modify $R(V^{\perp})$ accordingly. For a network in which azimuths and distances are measured for instance, we find from

$$\begin{pmatrix} \Delta x_i \\ \Delta y_i \end{pmatrix} = \begin{pmatrix} y^o_{ri} & x^o_{ri} \\ -x^o_{ri} & y^o_{ri} \end{pmatrix} \begin{pmatrix} \Delta A_{ri} \\ \Delta \ln l_{ri} \end{pmatrix} + \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \Delta x_r \\ \Delta y_r \end{pmatrix} , \tag{2.15}$$

that

$$R(V^{\perp}) = R\left(\begin{pmatrix} \cdot & \cdot \\ 1 & 0 \\ 0 & 1 \\ \vdots & \vdots \end{pmatrix} \right), \tag{2.16}$$

i.e. the appropriate (differential) similarity transformation is in this case the one in which scale and rotation is excluded.

To link up with the theory of the previous chapter note that it follows from

$$0 = \{ I - V^{\perp} ((S_i^{\perp})^t V^{\perp})^{-1} (S_i^{\perp})^t \} V^{\perp}$$

that in case of, for instance, an angular type of network all linear(ized) functions of the angular observables are invariant to the differential similarity transformation (2.6). Thus if the adjustment of the planar triangulation network of e.g. figure 15 is formulated as

$$\tilde{\mathbf{y}} = \mathbf{A}\,\mathbf{x} \tag{2.17}$$

then

$$Nu(\mathbf{A}) = R\left(\begin{pmatrix} \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & y^o_i & x^o_i \\ 0 & 1 & -x^o_i & y^o_i \\ \vdots & \vdots & \vdots & \vdots \end{pmatrix} \right) . \tag{2.18}$$

Hence we recognize transformation (2.14) as an example of an *S*-transformation, i.e.

$$\mathbf{x}^{(s_i)} = \mathbf{P}_{S_i,\, Nu(\mathbf{A})}\,\mathbf{x} . \tag{2.19}$$

Following (Baarda, 1973) we will therefore call the coordinate systems corresponding with choices like (2.12), *S*-systems.

At this point of our discussion it is perhaps fitting to make the following

remark concerning the choice of $S = R(S)$ complementary to $R(V^{\perp})$

Some authors, when dealing with free network adjustments, prefer to take the coordinate system definition corresponding to the choice

$$S^{\perp} := V^{\perp} \quad . \tag{2.20}$$

This is of course a legitimate choice, since it is just one of the many possible. However, we cannot endorse their claim that one always should choose (2.20) because it gives the "best" coordinate system definition possible.

They motivate their claim by pointing out that the covariance map of the pre-image of the BLUE's estimate $\hat{\mathbf{y}}$ of $\tilde{\mathbf{y}} = \mathbf{Ax}$ corresponding with the choice (2.20), has minimum trace, i.e. that

$$\text{trace}\{(I - V^{\perp}[(V^{\perp})^t V^{\perp}]^{-1}(V^{\perp})^t) Q_{\hat{x}^{(s)}} (I - V^{\perp}[(V^{\perp})^t V^{\perp}]^{-1}(V^{\perp})^t)\} \leq \text{trace } Q_{\hat{x}^{(s)}}$$

for all pre-images $\hat{\mathbf{x}}^{(s)}$ of $\hat{\mathbf{y}}$ under **A**.

This in itself is true of course. In case of free networks however, it is unfortunately without any meaning. All the essential information available is namely contained in $\hat{\mathbf{y}}$ whereby $\hat{\mathbf{x}}^{(s)}$ is nothing but a convenient way of representing this information. A theoretical basis for prefering (2.20) does therefore not exist in free network adjustments. At the most one can decide to choose (2.20) on the basis of computational convenience which might in some cases be due to the symmetry of $I - V^{\perp}[(V^{\perp})^t V^{\perp}]^{-1}(V^{\perp})^t$.

One could also rephrase the above as follows: Since every (oblique) projector can be interpreted as an orthogonal projector with respect to an appropriate chosen metric, the difference between the with choice (2.20) corresponding S-system and another arbitrary S-system can be interpreted as the difference in choosing a parameter-space norm, with (2.20) corresponding to the canonical parameter-space norm. And since there is no reason to prefer one particular norm above another, we do **not** have, as in physical geodesy, a norm choice problem in free network adjustments.

2.2 Ellipsoidal networks

So far we discussed the inverse linear mapping problem of planar geodetic networks. But let us now assume that we have to compute a geodetic network, the points of which are forced to lie **on** a given ellipsoid of revolution, defined by

$$\frac{x^2}{a^2} + \frac{y^2}{a^2} + \frac{z^2}{b^2} = 1, \tag{2.21}$$

where a and b are respectively the ellipsoid's major and minor axes.

In view of the foregoing discussion the three main questions we are interested in are then: (i) how does the theory of S-transformations apply to the ellipsoidal model, (ii) how does it compare to the

results we already obtained for the planar case and (iii) what are the consequences for practical network computations.

On an intuitive basis it is not too difficult to answer these three questions provisionally. From the rotational symmetry of the ellipsoid of revolution follows namely that the ellipsoidal model will at most admit one degree of freedom. And since this degree of freedom is of the longitudinal type it follows that the ellipsoidal counterpart of transformation (2.6) will read as

$$\begin{pmatrix} \vdots \\ \Delta\lambda_i \\ \vdots \end{pmatrix}^{(1)} = \begin{pmatrix} \vdots \\ \Delta\lambda_i \\ \vdots \end{pmatrix}^{(2)} + \begin{pmatrix} \vdots \\ 1 \\ \vdots \end{pmatrix} \Delta\epsilon_z , \tag{2.22}$$

where $\Delta\lambda_i$ is the geodetic longitude increment of point P_i and $\Delta\epsilon_z$ the differential rotation angle. Hence, transformation (2.22) can be used to derive the appropriate *S*-transformations for the ellipsoidal model.

As to the second question, if one wants to understand in what way and to what extent the ellipsoidal model differs from the planar model, we need a way of comparing both models. One can achieve this by considering the planar model as a special degenerate case of the ellipsoidal model. Assume therefore that we are given a geodetic triangle (i.e. a triangle bounded by geodesics) on the ellipsoid of revolution (2.21). By letting $e^2=(a^2-b^2)/a^2$, the first numerical eccentricity squared, approach zero we get for the limit $e^2 \to 0$ that the ellipsoid of revolution becomes a sphere with radius R:=a=b. Consequently, the given ellipsoidal triangle will become a spherical triangle for which then spherical geometry applies. Now, if we further proceed by letting the spherical curvature approach zero then for the limit $R \to \infty$ the sphere becomes identifiable with its own tangent planes. Hence, for increasing values of R the spherical triangle will ultimately reduce to an ordinary planar triangle.

Summarizing one could therefore say that the difference between ellipsoidal geometry and planar Euclidean geometry is primarily made up by the two factors e^2 and R. And one can thus expect that if both the ellipsoidal eccentricity factor e^2 and the spherical curvature 1/R are small enough, no significant differences will be recognizable between ellipsoidal geometry and planar Euclidean geometry.

But what about the admitted degrees of freedom? We note namely a drastic change in the maximal number of admitted degrees of freedom when the two limits $e^2 \to 0$ and $R \to \infty$ are taken: the ellipsoidal model only admits the longitudinal degree of freedom, whereas the planar model admits a maximum of four degrees of freedom. Still, despite this difference in admitted degrees of freedom it seems reasonable to expect that the actual estimation problem of the ellipsoidal model will not be too different from that of the planar model if e^2 and 1/R both are small enough. Consequently, it can be questioned whether in this case transformation (2.22) suffices to characterize the degrees of freedom admitted by the ellipsoidal model. Theoretically it does of course. But for practical applications it becomes questionable whether the rotational degree of freedom as described by (2.22) is the only degree of freedom the ellipsoidal model admits if both e^2 and 1/R are small.

This then brings us to the third question concerning the consequences for practical network computations. Namely, the smaller e^2 and 1/R get the worse the conditioning of the ellipsoidal networks' design matrix A can expected to be. That is, although theoretically the maximum defect of

A equals one, it can be expected that for small enough values of e^2 and 1/R more than one of the columns of the design matrix A will show near linear dependencies. As a consequence one can therefore expect that the ill-conditioning of A will affect the estimation of the explanatory variables x in the linear model $\tilde{y} = A\ x$. Intuitively one can understand this by realizing that the almost collinear variables do not provide information that is very different from that already inherent in others. It becomes difficult therefore to infer the separate influence of such explanatory variables on the response $\tilde{y}$. Consequently, the potential harm due to the ill-conditioning of the design matrix arises from the fact that a near collinear relation can readily result in a situation in which some of the observed systematic influences of the explanatory variables x on the response is swamped by the model's random error term. And it will be clear that under these circumstances, estimation can be hindered.

To find out whether for practical ellipsoidal networks the estimation of geodetic coordinates is indeed hindered by the expected ill-conditioning of A, one can follow different but related routes. One way is to investigate numerically to what extent the shape of an ellipsoidal network as measured by its angles, can considered to be invariant to a change of its position, orientation and scale. Another way is to compute the non-zero singular values of A or the non-zero eigenvalues of the normal matrix A^tA. Eigenvalues small relative to the largest eigenvalue of the normal matrix will then reflect the poor conditioning of A. And finally one could try to show analytically that the estimated geodetic coordinates lack precision if only the longitudinal degree of freedom is taken care of.

The first approach, which is based on the idea that for planar geodetic networks of the angular type the invariance to position, orientation and scale changes is complete, has been followed by (Nibbelke, 1984). And he found that for practical ellipsoidal triangulation networks one can indeed consider the network's position, orientation and scale as non-estimable. That is, one is, just as in the planar case, forced to fix four linear independent functions of the geodetic coordinate increments. The theoretical deformations of the network's shape, which possibly follow from these restrictions, are then negligible. The same conclusion was also reached by (Kube and Schnädelbach, 1975), who used the second approach. The reported eigenvalue computations which were performed for the European network show that in case of, for instance, an ellipsoidal triangulation network, four eigenvalues of the normal matrix will be so small that a sensible estimation of the network's position, orientation and scale is not attainable. This conclusion is also in agreement with the result found by (Krarup, 1982a), who indicated that the position of a trilateration network on an ellipsoid of revolution is practically non-estimable.

As an example and also to support the above mentioned findings we will now show analytically that the estimation of geodetic coordinates indeed lacks precision if only the longitudinal degree of freedom is taken care of. For this purpose assume that we have a full rank linear model

$$\underset{m\times 1}{\tilde{y}} = \underset{m\times n}{A}\ \underset{n\times 1}{x},$$

in which x_2 of $x = (x_1^t\ x_2^t)^t$ has been identified as the parameter which is degraded by the ill-conditioning of A.

From the partitioning

$$\underset{m\times 1}{\tilde{y}} = \underset{m\times(n-1)}{(A_1\quad A_2)} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}_{n\times 1}, \tag{2.23}$$

follows then that the column vector A_2 depends almost linearly on the columns of A_1. Using the reparametrization

$$\bar{x}_1 := x_1 + (A_1^t A_1)^{-1} A_1^t A_2 x_2; \qquad x_2 := x_2,$$

we can write (2.23) as

$$\tilde{y} = A_1(x_1 + (A_1^t A_1)^{-1} A_1^t A_2 x_2) + (I - A_1(A_1^t A_1)^{-1} A_1^t) A_2 x_2$$

or as

$$\tilde{y} = A_1 \bar{x}_1 + \bar{A}_2 x_2, \tag{2.24}$$

with

$$\bar{A}_2 = (I - A_1(A_1^t A_1)^{-1} A_1^t) A_2. \tag{2.25}$$

From the fact that A_2 depends almost linearly on the columns of A_1 now follows that one can reasonably expect $\bar{A}_2$ to be a rather short column vector. Geometrically this is seen as follows. Since $I - A_1(A_1^t A_1)^{-1} A_1^t$ is an orthogonal projector, we have that (see figure 17)

$$\bar{A}_2^t \bar{A}_2 = A_2^t (I - A_1(A_1^t A_1)^{-1} A_1^t) A_2 = A_2^t A_2 \sin^2 \theta, \tag{2.26}$$

where θ denotes the angle between A_2 and its orthogonal projection on the subspace spanned by the columns of A_1.

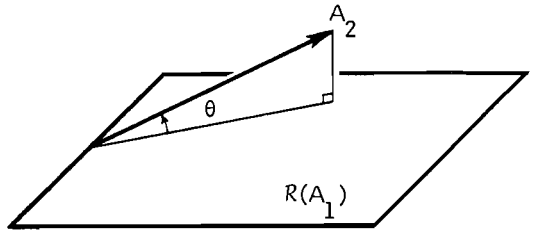


figure 17

From the near linear dependency of A_1 and A_2 thus follows that the angle θ will be small. Hence, the length of $\bar{A}_2$ can be expected to be small if the length of A_2 is not too large.
Now if we assume the covariance map of y to be $Q_y = \sigma^2 I$, it follows from (2.24) and the orthogonality of A_1 and $\bar{A}_2$ that the variance $\sigma^2_{x_2}$ of x_2 is given by

$$\sigma^2_{x_2} = \frac{\sigma^2}{\bar{A}_2^t \bar{A}_2} = \frac{\sigma^2}{A_2^t A_2 \sin^2 \theta} = \frac{\sigma^2}{A_2^t (I - A_1(A_1^t A_1)^{-1} A_1^t) A_2}. \tag{2.27}$$

Hence, the estimation of x_2 lacks precision if the length of $\bar{A}_2$ is too small. Thus in order to find out to what extent the diagnosed ill-conditioning of A affects the estimation of x_2 we need to have a reasonable estimate of (2.27).
Since we know that the possible lack of precision of the estimated parameter x_2 is a consequence of

the near linear dependency between A_1 and A_2, it follows that there must exist a vector, z say, for which

$$A\ z = \nabla \quad , \tag{2.28}$$

is small enough. From writing (2.28) as

$$(A_1\ A_2) \begin{pmatrix} z \\ 1 \end{pmatrix} = \nabla \quad ,$$

we get

$$A_2 = \nabla - A_1 z_1 \quad .$$

Hence, expression (2.27) can also be written as

$$\sigma^2_{x_2} = \frac{\sigma^2}{\nabla^t (I - A_1 (A_1^t A_1)^{-1} A_1^t) \nabla} \quad . \tag{2.29}$$

With $\nabla^t (I - A_1 (A_1^t A_1)^{-1} A_1^t) \nabla \leq \nabla^t \nabla$, we then get the lower bound

$$\sigma^2_{x_2} \geq \frac{\sigma^2}{\nabla^t \nabla} \quad . \tag{2.30}$$

Thus if we are able to find a vector z such that the length of $A\ z = \nabla$ is small enough, we can use the lower bound of (2.30) to prove that the estimation of x_2 indeed lacks precision.

Now, to apply the above to our case of ellipsoidal networks, recall that we made it plausible that the difference between ellipsoidal, spherical and planar Euclidean geometry can considered to be insignificant if both the factors e^2 and 1/R are small enough. One can therefore expect that for small enough values of e^2 and 1/R, the eigenvectors of spherical- and planar networks' design matrices belonging to zero eigenvalues are the proper candidates for the z-vector of (2.28). For this purpose we thus first need to find the spherical analogon of (2.3) (or (2.15)).

We will start from the three dimensional differential similarity transformation

$$\begin{pmatrix} \Delta x \\ \Delta y \\ \Delta z \end{pmatrix}_i^{(1)} = \begin{pmatrix} \Delta x \\ \Delta y \\ \Delta z \end{pmatrix}_i^{(2)} + \begin{pmatrix} 1 & 0 & 0 & 0 & -z_i^o & y_i^o & x_i^o \\ 0 & 1 & 0 & z_i^o & 0 & -x_i^o & y_i^o \\ 0 & 0 & 1 & -y_i^o & x_i^o & 0 & z_i^o \end{pmatrix} \begin{pmatrix} \Delta t \\ \Delta t^x \\ \Delta t^y \\ \Delta \varepsilon^z \\ \Delta \varepsilon^x \\ \Delta \varepsilon^y \\ \Delta \kappa^z \end{pmatrix} \quad . \tag{2.31}$$

With

$$\begin{pmatrix} x \\ y \\ z \end{pmatrix}_i = \begin{pmatrix} (N_i + h_i) \cos\phi_i \cos\lambda_i \\ (N_i + h_i) \cos\phi_i \sin\lambda_i \\ (N_i (1 - e^2) + h_i) \sin\phi_i \end{pmatrix} \quad ,$$

and

$$\begin{pmatrix} \Delta x \\ \Delta y \\ \Delta z \end{pmatrix}_i = \begin{pmatrix} \cos\phi_i^o \cos\lambda_i^o & -\sin\phi_i^o \cos\lambda_i^o & -\sin\lambda_i^o \\ \cos\phi_i^o \sin\lambda_i^o & -\sin\phi_i^o \sin\lambda_i^o & \cos\lambda_i^o \\ \sin\phi_i^o & \cos\phi_i^o & 0 \end{pmatrix} \begin{pmatrix} \Delta h \\ (M_i^o + h_i^o)\Delta\phi \\ N_i^o \cos\phi_i^o \Delta\lambda \end{pmatrix}_i ,$$

where ϕ_i, λ_i and h_i are respectively the geodetic latitude, longitude and geometric height above the ellipsoid of point P_i, and N_i, M_i are the east-west and north-south radii of curvature, one can rewrite (2.31) in geodetic coordinates as

$$\begin{pmatrix} (N_i^o + h_i^o)\cos\phi_i^o \Delta\lambda \\ (M_i^o + h_i^o)\Delta\phi \\ \Delta h \end{pmatrix}_i^{(1)} = \begin{pmatrix} (N_i^o + h_i^o)\cos\phi_i^o \Delta\lambda \\ (M_i^o + h_i^o)\Delta\phi \\ \Delta h \end{pmatrix}_i^{(2)} +$$

$$\begin{bmatrix} -\sin\lambda_i^o & \cos\lambda_i^o & 0 & (N_i^o(1-e^2)+h_i^o)\sin\phi_i^o\cos\lambda_i^o \\ -\sin\phi_i^o\cos\lambda_i^o & -\sin\phi_i^o\sin\lambda_i^o & \cos\phi_i^o & -(N_i^o(1-e^2\sin^2\phi_i^o)+h_i^o)\sin\lambda_i^o \\ \cos\phi_i^o\cos\lambda_i^o & \cos\phi_i^o\sin\lambda_i^o & \sin\phi_i^o & -e^2 N_i^o \cos\phi_i^o \sin\phi_i^o \sin\lambda_i^o \end{bmatrix}$$

$$\begin{bmatrix} (N_i^o(1-e^2)+h_i^o)\sin\phi_i^o\sin\lambda_i^o & -(N_i^o+h_i^o)\cos\phi_i^o & 0 \\ (N_i^o(1-e^2\sin^2\phi_i^o)+h_i^o)\cos\lambda_i^o & 0 & -e^2 N_i^o \cos\phi_i^o \sin\phi_i^o \\ e^2 N_i^o \cos\phi_i^o \sin\phi_i^o \cos\lambda_i^o & 0 & (N_i^o(1-e^2\sin^2\phi_i^o)+h_i^o) \end{bmatrix} \begin{pmatrix} \Delta t_x \\ \Delta t_y \\ \Delta t_z \\ \Delta\varepsilon_x \\ \Delta\varepsilon_y \\ \Delta\varepsilon_z \\ \Delta\kappa \end{pmatrix}_i . \tag{2.32}$$

Since the network points are forced to lie on the ellipsoid of revolution, we must have that

$$h_i^o = 0 \text{ and } \Delta h_i = 0 \qquad \forall i = 1,\ldots \tag{2.33}$$

Hence, it follows from (2.32) that

$$\begin{aligned} 0 = {} & \cos\phi_i^o\cos\lambda_i^o \Delta t_x + \cos\phi_i^o \sin\lambda_i^o \Delta t_y + \sin\phi_i^o \Delta t_z - e^2 N_i^o \cos\phi_i^o \sin\phi_i^o \sin\lambda_i^o \Delta\varepsilon_x + \\ & + e^2 N_i^o \cos\phi_i^o \sin\phi_i^o \cos\lambda_i^o \Delta\varepsilon_y + N_i^o(1-e^2\sin^2\phi_i^o)\Delta\kappa , \\ & \qquad \forall\ i = 1,\ldots \end{aligned} \tag{2.34}$$

But this means that for a regular network (i.e. a network which excludes cases like λ_i = constant, $\forall i = 1,\ldots$) situated on an ellipsoid of revolution we have that

$$\Delta t_x = \Delta t_y = \Delta t_z = \Delta\varepsilon_x = \Delta\varepsilon_y = \Delta\kappa = 0, \tag{2.35}$$

which confirms our earlier statement that the ellipsoidal model only admits the longitudinal degree of freedom.

In an analogous way we can find the type of degrees of freedom admitted by the spherical model. In spherical coordinates R_i, ϕ_i and λ_i, transformation (2.32) will namely read as

$$\begin{pmatrix} R_i^o \cos\phi_i^o \Delta\lambda_i \\ R_i^o \Delta\phi_i \\ \Delta R_i \end{pmatrix}^{(1)} = \begin{pmatrix} R_i^o \cos\phi_i^o \Delta\lambda_i \\ R_i^o \Delta\phi_i \\ \Delta R_i \end{pmatrix}^{(2)} +$$

$$+ \begin{pmatrix} -\sin\lambda_i^o & \cos\lambda_i^o & 0 & R_i^o \sin\phi_i^o \cos\lambda_i^o & R_i^o \sin\phi_i^o \sin\lambda_i^o & -R_i^o \cos\phi_i^o & 0 \\ -\sin\phi_i^o \cos\lambda_i^o & -\sin\phi_i^o \sin\lambda_i^o & \cos\phi_i^o & -R_i^o \sin\lambda_i^o & R_i^o \cos\lambda_i^o & 0 & 0 \\ \cos\phi_i^o \cos\lambda_i^o & \cos\phi_i^o \sin\lambda_i^o & \sin\phi_i^o & 0 & 0 & 0 & R_i^o \end{pmatrix} \begin{pmatrix} \Delta t_x \\ \Delta t_y \\ \Delta t_z \\ \Delta\varepsilon_x \\ \Delta\varepsilon_y \\ \Delta\varepsilon_z \\ \Delta\kappa \end{pmatrix} \tag{2.36}$$

And by setting

$$R_i^o = R \text{ and } \Delta R_i = 0 \qquad \forall i = 1,\ldots, \tag{2.37}$$

we get that

$$0 = \cos\phi_i^o \cos\lambda_i^o \Delta t_x + \cos\phi_i^o \sin\lambda_i^o \Delta t_y + \sin\phi_i^o \Delta t_z + R\,\Delta\kappa, \qquad \forall i = 1,\ldots, \tag{2.38}$$

from which follows with (2.36) that the spherical counterpart of (2.6) is given by

$$\begin{pmatrix} R\cos\phi_i^o \Delta\lambda_i \\ R\Delta\phi_i \end{pmatrix}^{(1)} = \begin{pmatrix} R\cos\phi_i^o \Delta\lambda_i \\ R\Delta\phi_i \end{pmatrix}^{(2)} + \begin{pmatrix} R\sin\phi_i^o \sin\lambda_i^o & R\sin\phi_i^o \sin\lambda_i^o & -R\cos\phi_i^o \\ -R\sin\phi_i^o & R\cos\lambda_i^o & 0 \end{pmatrix} \begin{pmatrix} \Delta\varepsilon_x \\ \Delta\varepsilon_y \\ \Delta\varepsilon_z \end{pmatrix}. \tag{2.39}$$

To find the expression which corresponds to (2.3) (or (2.15)), we first need to know the relation between $(\Delta\varepsilon_x, \Delta\varepsilon_y, \Delta\varepsilon_z)^t$ and $(\Delta\phi_r, \Delta\lambda_r, \Delta A_{rs})^t$. This is given by

$$\begin{pmatrix} \Delta\varepsilon_x \\ \Delta\varepsilon_y \\ \Delta\varepsilon_z \end{pmatrix} = \begin{pmatrix} -\sin\lambda_r^o & 0 & \cos\phi_r^o\cos\lambda_r^o \\ \cos\lambda_r^o & 0 & \cos\phi_r^o\sin\lambda_r^o \\ 0 & -1 & \sin\phi_r^o \end{pmatrix} \begin{pmatrix} \Delta\phi_r \\ \Delta\lambda_r \\ \Delta A_{rs} \end{pmatrix} . \qquad (2.40)$$

Substitution of (2.40) into (2.39) then gives

$$(R\cos\phi_i^o\Delta\lambda_i^{(1)} \;,\; R\Delta\phi_i^{(1)})^t = (R\cos\phi_i^o\Delta\lambda_i^{(2)} \;,\; R\Delta\phi_i^{(2)})^t +$$

$$\begin{pmatrix} \cos\phi_i^o & : & \\ \hline \cos\phi_r^o & : & \\ 0 & \cos(\lambda_i^o-\lambda_r^o) & \end{pmatrix} \begin{pmatrix} \sin\phi_i^o\sin(\lambda_i^o-\lambda_r^o) & : & R\sin\phi_i^o\cos\phi_r^o\cos(\lambda_i^o-\lambda_r^o)-R\cos\phi_i^o\sin\phi_r^o \\ & : & \\ & : & -R\cos\phi_r^o\sin(\lambda_i^o-\lambda_r^o) \end{pmatrix} \begin{pmatrix} R\cos\phi_r^o\Delta\lambda_r \\ R\Delta\phi_r \\ \Delta A_{rs} \end{pmatrix} \qquad (2.41)$$

The spherical analogon of (2.3) (or (2.15)) then finally follows from substituting

$$-\sin\frac{l_{ir}}{R}\sin A_{ir} = \sin(\tfrac{1}{2}\pi-\phi_r)\sin(\lambda_i-\lambda_r)$$
$$= \cos\phi_r\sin(\lambda_i-\lambda_r) \qquad (2.42.a)$$

and

$$\sin\frac{l_{ir}}{R}\cos(2\pi-A_{ir}) = \sin(\tfrac{1}{2}\pi-\phi_i)\cos(\tfrac{1}{2}\pi-\phi_r)-\cos(\tfrac{1}{2}\pi-\phi_i)\sin(\tfrac{1}{2}\pi-\phi_r)\cos(\lambda_i-\lambda_r)$$
$$= \cos\phi_i\sin\phi_r - \sin\phi_i\cos\phi_r\cos(\lambda_i-\lambda_r) \qquad (2.42.b)$$

into (2.41):

$$(R\cos\phi_i^o\Delta\lambda_i^{(1)} \;,\; R\Delta\phi_i^{(1)})^t = (R\cos\phi_i^o\Delta\lambda_i^{(2)} \;,\; R\Delta\phi_i^{(2)})^t +$$

$$+\begin{pmatrix} \cos\phi_i^o & : & \\ \hline \cos\phi_r^o & : & \\ 0 & \cos(\lambda_i^o-\lambda_r^o) & \end{pmatrix} \begin{pmatrix} \sin\phi_i^o\sin(\lambda_i^o-\lambda_r^o) & : & -R\sin\frac{l_{ir}^o}{R}\cos A_{ir}^o \\ & : & \\ & : & R\sin\frac{l_{ir}^o}{R}\sin A_{ir}^o \end{pmatrix} \begin{pmatrix} R\cos\phi_r^o\Delta\lambda_r \\ R\Delta\phi_r \\ \Delta A_{rs} \end{pmatrix} . \qquad (2.43)$$

(2.42.a) and (2.42.b) follow from applying the sin-rule sin a/sin A = sin b/sin B and the so-called five-elements' rule sin c cos a - cos c sin a cos B = cos A cos b (see figure 18) of spherical geometry.

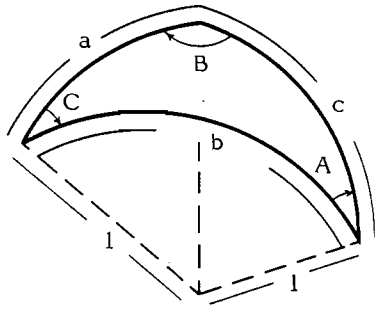


figure 18

Expression (2.43) shows that not surprisingly the spherical model admits a maximal number of three degrees of freedom, all of which are of the rotational type. Hence, we find that theoretically speaking the scale of a spherical network is estimable. Even if only angles are measured. Those who are familiar with global aspects of differential geometry know this of course already from the Gauss-Bonnet formula. When applied to the sphere, this formula says that for a triangular region bounded by three geodesics the sum of the spherical triangle's interior angles minus π equals the ratio of the area enclosed by the triangle and the radius of the sphere (see e.g. Stoker, 1969). We are here thus confronted with a situation where angles alone suffice to determine scale. But still, although scale is theoretically estimable, one can expect, as was made clear in the foregoing introductory discussion, that if the spherical curvature is small enough scale will only be very poorly estimable. And indeed it turns out that for practical spherical networks, scale can be considered as non-estimable. See for instance (Molenaar, 1980a,p.20) or the earlier cited references.

In the same manner it is concluded in these publications that the scale, orientation and position of practical ellipsoidal networks, can considered to be non-estimable. To support these findings we will now show analytically, that the geodetic coordinates lack precision if only the longitudinal degree of freedom of the ellipsoidal model is taken care of. For this purpose consider expression (2.43). The three columns of the matrix on the right-hand side of (2.43) span the nullspace of the design matrix of a spherical triangulation network, whereas the first column vector provides a basis of the nullspace of an ellipsoidal network's design matrix. Thus, if the eccentricity factor e^2 is small enough one can expect that both the second and third column vector of (2.43) get almost annihilated by the ellipsoidal network's design matrix. Hence, we can use one of these vectors, say

$$z = \begin{bmatrix} \vdots \\ \sin\phi_i^o \sin(\lambda_i^o - \lambda_r^o) \\ \vdots \\ \cos(\lambda_i^o - \lambda_r^o) \\ \vdots \end{bmatrix}, \qquad (2.44)$$

to obtain an estimate of the lowerbound (2.30) via (2.28).

Let us consider as an example an ellipsoidal trilateration network. According to (Helmert, 1880, p. 282) the ellipsoidal distance observation equation reads as:

$$\Delta l_{ij} = -\sin\bar{A}_{ij}^o (N_i^o \cos\phi_i^o \Delta\lambda_i) - \cos\bar{A}_{ij}^o (M_i^o \Delta\phi_i) - \sin\bar{A}_{ji}^o (N_j^o \cos\phi_j^o \Delta\lambda_j) - \cos\bar{A}_{ji}^o (M_j^o \Delta\phi_j), \qquad (2.45)$$

where $\bar{A}_{ij}$ denotes the ellipsoidal geodesic azimuth from P_i to P_j. We will abbreviate (2.45) as

$$\Delta l_{ij} = a_k^t x \quad , \tag{2.46}$$

where $a_k = (\ldots\ -\sin\bar{A}^o_{ij},\ -\cos\bar{A}^o_{ij},\ -\sin\bar{A}^o_{ji},\ -\cos\bar{A}^o_{ji},\ \ldots)^t$

is the kth rowvector of the ellipsoidal network's design matrix and

$$x = (\ldots N^o_i \cos\phi^o_i \Delta\lambda_i,\ M^o_i \Delta\phi_i, \ldots)^t .$$

Using (2.44) we get

$$\begin{aligned} \nabla_k = a_k^t z = &-\sin\bar{A}^o_{ij} \sin\phi^o_i \sin(\lambda^o_i - \lambda^o_r) - \cos\bar{A}^o_{ij} \cos(\lambda^o_i - \lambda^o_r) \\ &-\sin\bar{A}^o_{ji} \sin\phi^o_j \sin(\lambda^o_j - \lambda^o_r) - \cos\bar{A}^o_{ji} \cos(\lambda^o_j - \lambda^o_r) \quad . \end{aligned} \tag{2.47}$$

It will be clear that if the network is situated on a sphere, then $\nabla_k = 0$. Let us therefore identify geodesic coordinates with spherical coordinates. With

$$\bar{A}_{ij} = A_{ij} + \Delta A_{ij} \quad \text{and} \quad \bar{A}_{ji} = A_{ji} + \Delta A_{ji},$$

where A_{ij} denotes the spherical azimuth between the points P'_i and P'_j, which are obtained from identifying geodesic coordinates with spherical coordinates, and linear approximations like

$$\sin \bar{A}_{ij} = \sin A_{ij} + \cos A_{ij} \Delta A_{ij},$$

we can rewrite (2.47) as

$$\begin{aligned} \nabla_k = &\left(\sin A^o_{ij} \cos(\lambda^o_i - \lambda^o_r) - \cos A^o_{ij} \sin\phi^o_i \sin(\lambda^o_i - \lambda^o_r)\right)\Delta A_{ij} + \\ &+ \left(\sin A^o_{ji} \cos(\lambda^o_j - \lambda^o_r) - \cos A^o_{ji} \sin\phi^o_j \sin(\lambda^o_j - \lambda^o_r)\right)\Delta A_{ji} \quad . \end{aligned} \tag{2.48}$$

Repeated application of the sine-rule and five-elements' rule of spherical geometry and

$$|\Delta A_{ij}| \simeq |\Delta A_{ji}| \leq \frac{1}{2} e^2 \quad \text{(see Helmert, 1880, p. 289),}$$

then finally gives

$$|\nabla_k| \leq \frac{1}{2} \frac{l_{ij}}{R} \frac{l_{ir}}{R} e^2 \quad . \tag{2.49}$$

From this estimate and (2.30) thus indeed follows that in case of practical ellipsoidal networks (l_{ij} = 64 km, R = 6400 km, $\sigma = \frac{1}{2}\, 10^{-5} . l_{ij}$, $e^2 = 1/300$) geodetic coordinates will lack precision if only the longitudinal degree of freedom is taken care of.

2.3. Three dimensional networks

Now that we have considered the inverse mapping problem in two dimensions it is not too difficult to generalize to three dimensions.

We will first assume that only angles and distance ratios are measured in the three dimensional geodetic network. The generalization of (2.1) to three dimensions becomes then rather straightforward. To see this, observe that we can write (2.1) as

$$\begin{pmatrix} x_i \\ y_i \\ 0 \end{pmatrix} = \begin{pmatrix} x_r \\ y_r \\ 0 \end{pmatrix} + \begin{pmatrix} l_{rs}\sin A_{rs} \\ l_{rs}\cos A_{rs} \\ 0 \end{pmatrix} - \frac{l_{si}}{l_{sr}}\cos\alpha_{rsi}\begin{pmatrix} l_{rs}\sin A_{rs} \\ l_{rs}\cos A_{rs} \\ 0 \end{pmatrix} - \frac{l_{si}}{l_{sr}}\sin\alpha_{rsi}\begin{pmatrix} 0 & 1 & 0 \\ -1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}\begin{pmatrix} l_{rs}\sin A_{rs} \\ l_{rs}\cos A_{rs} \\ 0 \end{pmatrix} \tag{2.50}$$

where the action of the matrix

$$\begin{pmatrix} 0 & 1 & 0 \\ -1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

equals the action of

$$\begin{pmatrix} 0 \\ 0 \\ -1 \end{pmatrix} \mathbf{x} \ , \tag{2.51}$$

with "**x**" denoting the vector- or cross product.

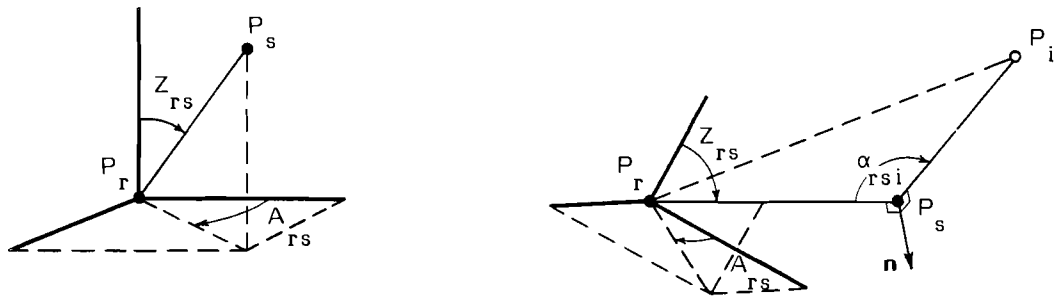
With (2.51), expression (2.50) therefore suggests the following generalization to three dimensions:

$$\begin{pmatrix} x_i \\ y_i \\ z_i \end{pmatrix} = \begin{pmatrix} x_r \\ y_r \\ z_r \end{pmatrix} + \begin{pmatrix} l_{rs}\sin Z_{rs}\sin A_{rs} \\ l_{rs}\sin Z_{rs}\cos A_{rs} \\ l_{rs}\cos Z_{rs} \end{pmatrix} - \frac{l_{si}}{l_{rs}}\cos\alpha_{rsi}\begin{pmatrix} l_{rs}\sin Z_{rs}\sin A_{rs} \\ l_{rs}\sin Z_{rs}\cos A_{rs} \\ l_{rs}\cos Z_{rs} \end{pmatrix}$$

$$- \frac{l_{si}}{l_{sr}}\sin\alpha_{rsi}\begin{pmatrix} n_1 \\ n_2 \\ n_3 \end{pmatrix} \mathbf{x} \begin{pmatrix} l_{rs}\sin Z_{rs}\sin A_{rs} \\ l_{rs}\sin Z_{rs}\cos A_{rs} \\ l_{rs}\cos Z_{rs} \end{pmatrix}, \tag{2.52}$$

where Z_{rs} denotes the vertical angle of the line P_rP_s (see figure 19.a) and $\mathbf{n} = (n_1 n_2 n_3)^t$ is the unit normal of the plane through the points P_r, P_s and P_i (see figure 19.b) defined as

$$\begin{pmatrix} n_1 \\ n_2 \\ n_3 \end{pmatrix} = \frac{1}{l_{sr}\,l_{si}\sin\alpha_{rsi}}\begin{pmatrix} x_{sr} \\ y_{sr} \\ z_{sr} \end{pmatrix} \mathbf{x} \begin{pmatrix} x_{si} \\ y_{si} \\ z_{si} \end{pmatrix}. \tag{2.53}$$



(a) (b)

figure 19

We thus see that one way of introducing coordinates for three dimensional networks of the angular type is by starting to fix the two points P_r and P_s. This would then take care of six degrees of freedom. Namely, three translational degrees of freedom, two rotational degrees of freedom and one freedom of scale. The remaining rotational degree of freedom, namely rotation of the network around the line P_rP_s, is then taken care of by fixing the direction of the unit normal $\mathbf{n}$ in the plane perpendicular to the line P_rP_s. The so defined coordinate system thus corresponds to fixing two points P_r and P_s, and the plane through these two points and a third point, P_t say. Following (Baarda, 1979) we will denote this S-system as the (r,s;t)-system. The $(S^{\perp}_{(r,s;t)})^t$- matrix by which the (r,s;t)-system is defined then follows from the restrictions

$$\left.\begin{aligned} \Delta x_r = \Delta y_r = \Delta z_r &= 0 \\ \Delta x_s = \Delta y_s = \Delta z_s &= 0 \\ n^o_1 \Delta x_t + n^o_2 \Delta y_t + n^o_3 \Delta z_t &= 0 \end{aligned}\right\}, \tag{2.54}$$

where $\mathbf{n}^o = (n^o_1 n^o_2 n^o_3)^t$ can be computed from (2.53) for i = t using approximate values. With

$$R(V^{\perp}) = R\left(\begin{pmatrix} \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 0 & 0 & -z^o_i & y^o_i & x^o_i \\ 0 & 1 & 0 & z^o_i & 0 & -x^o_i & y^o_i \\ 0 & 0 & 1 & -y^o_i & x^o_i & 0 & z^o_i \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \end{pmatrix}\right), \tag{2.55}$$

which follows from the three dimensional differential similarity transformation (2.31), straightforward application of (2.14) then gives

$$\begin{pmatrix} \Delta x_i \\ \Delta y_i \\ \Delta z_i \end{pmatrix}^{(r,s;t)} = \begin{pmatrix} \Delta x_{ri} \\ \Delta y_{ri} \\ \Delta z_{ri} \end{pmatrix} + \frac{1}{(l^o_{rs})^2} \left\{ \begin{pmatrix} n^o_1 \\ n^o_2 \\ n^o_3 \end{pmatrix} \times \begin{pmatrix} x^o_{ri} \\ y^o_{ri} \\ z^o_{ri} \end{pmatrix} \right\} \left\{ \begin{pmatrix} n^o_1 \\ n^o_2 \\ n^o_3 \end{pmatrix} \times \begin{pmatrix} x^o_{rs} \\ y^o_{rs} \\ z^o_{rs} \end{pmatrix} \right\}^t + \begin{pmatrix} x^o_{ri} \\ y^o_{ri} \\ z^o_{ri} \end{pmatrix} \begin{pmatrix} x^o_{rs} \\ y^o_{rs} \\ z^o_{rs} \end{pmatrix}^t \left\{ \begin{pmatrix} \Delta x_{sr} \\ \Delta y_{sr} \\ \Delta z_{sr} \end{pmatrix} + \right.$$

$$+\frac{1}{l^o_{rs}\,l^o_{rt}\sin\alpha^o_{trs}}\left(\begin{pmatrix}x^o_{st}\\y^o_{st}\\z^o_{st}\end{pmatrix}\times\begin{pmatrix}x^o_{ri}\\y^o_{ri}\\z^o_{ri}\end{pmatrix}\right)\begin{pmatrix}n^o_1\\n^o_2\\n^o_3\end{pmatrix}^t\begin{pmatrix}\Delta x_r\\\Delta y_r\\\Delta z_r\end{pmatrix}+\frac{1}{l^o_{rs}\,l^o_{st}\sin\alpha^o_{rst}}\left(\begin{pmatrix}x^o_{tr}\\y^o_{tr}\\z^o_{tr}\end{pmatrix}\times\begin{pmatrix}x^o_{ri}\\y^o_{ri}\\z^o_{ri}\end{pmatrix}\right)\begin{pmatrix}n^o_1\\n^o_2\\n^o_3\end{pmatrix}^t\begin{pmatrix}\Delta x_s\\\Delta y_s\\\Delta z_s\end{pmatrix}$$

$$+\frac{1}{l^o_{st}\,l^o_{rt}\sin\alpha^o_{str}}\left(\begin{pmatrix}x^o_{rs}\\y^o_{rs}\\z^o_{rs}\end{pmatrix}\times\begin{pmatrix}x^o_{ri}\\y^o_{ri}\\z^o_{ri}\end{pmatrix}\right)\begin{pmatrix}n^o_1\\n^o_2\\n^o_3\end{pmatrix}^t\begin{pmatrix}\Delta x_t\\\Delta y_t\\\Delta z_t\end{pmatrix}. \qquad (2.56)$$

Expression (2.56) can be considered as the natural generalization of (2.9). Namely, if we restrict our attention in (2.56) to the Δx, Δy- parts of the points P_i, P_r and P_s and take $z^o_i = 0 \quad \forall i = 1,\ldots,$ and also $n^o_1 = n^o_2 = 0$, $n^o_3 = -1$, we obtain (2.9) again.

In deriving the three dimensional *S*-transformation (2.56) we assumed that only angles and distance ratios were observed. But this assumption is generally only valid in local three dimensional surveys (e.g. construction works). In large three dimensional networks, one will usually have besides the angles and distance ratios also direction measurements like astronomical azimuth, latitude and longitude at ones disposal. It is likely then that one is not only interested in the (cartesian) coordinates describing the network's configuration but also in the orientation (and possibly scale) parameters describing fundamental directions like local verticals and the earth's average rotation axis. It seems therefore that for large three dimensional networks transformations like (2.56), which only transform coordinates (and their co-variances) do not really suffice. And this becomes even more apparent if one thinks of connecting such networks. For large networks we therefore need *S*-transformations that also transform orientation (and scale) parameters.

Now before deriving such *S*-transformations let us first draw a parallel with the two dimensional planar case. Since in practice the observation equations are usually written down in terms of directions r_{ij} and pseudo-distances l_{ij} instead of in terms of angles and distance ratios, the parameter vector x of the linear model $\tilde{y} = Ax$ will contain besides the coordinate increments also orientation- and scale unknowns. Hence, the linear model of two dimensional planar networks will in practice be of the same form as that of large three dimensional networks:

$$\tilde{y} = (A_1 A_2)\begin{pmatrix}x_1\\x_2\end{pmatrix}, \qquad (2.57)$$

with, x_1: coordinate unknowns; x_2: orientation- and/or scale unknowns.

Thus also in case of two dimensional networks one can in principle decide to involve the orientation- and scale unknowns in the many *S*-systems possible. Of course in practice one will not do so, since in two dimensional planar networks these unknowns are generally of no particular interest. But still, let us, for the sake of comparison between the two- and three dimensional case, pursue the idea of involving these unknowns in the many *S*-systems possible.

Consider for this purpose a two dimensional planar network with direction- and pseudo-distance measurements r_{ij} and l_{ij}. In figure 20 a part of such a network is drawn. The theodolite frames in points P_r and P_i are shown by dashed lines and the directions $P_rP_{\bar{r}}$, $P_iP_{\bar{i}}$ are the directions of zero reading.

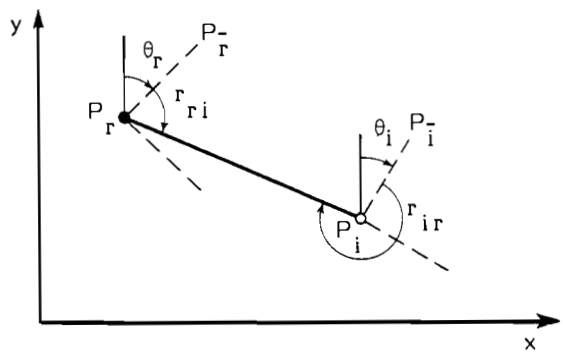


figure 20

Analogous to (2.1) we can then write

$$\left.\begin{aligned} x_i &= x_r + \kappa_r l_{ri} \sin(\theta_r + r_{ri}) \\ y_i &= y_r + \kappa_r l_{ri} \cos(\theta_r + r_{ri}) \\ \theta_i &= \theta_r + r_{ri} - r_{ir} + \pi \\ \ln\kappa_i &= \ln\kappa_r + \ln l_{ri} - \ln l_{ir} \end{aligned}\right\} \qquad (2.58)$$

And linearization gives

$$\begin{pmatrix} \Delta x_i \\ \Delta y_i \\ \Delta\theta_i \\ \Delta\ln\kappa_i \end{pmatrix} = \begin{pmatrix} y^o_{ri} & x^o_{ri} & 0 & 0 \\ -x^o_{ri} & y^o_{ri} & 0 & 0 \\ 1 & 0 & -1 & 0 \\ 0 & 1 & 0 & -1 \end{pmatrix} \begin{pmatrix} \Delta r_{ri} \\ \Delta\ln l_{ri} \\ \Delta r_{ir} \\ \Delta\ln l_{ir} \end{pmatrix} + \begin{pmatrix} 1 & 0 & y^o_{ri} & x^o_{ri} \\ 0 & 1 & -x^o_{ri} & y^o_{ri} \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \Delta x_r \\ \Delta y_r \\ \Delta\theta_r \\ \Delta\ln\kappa_r \end{pmatrix} . \qquad (2.59)$$

Hence, if the unknowns in the linear model (2.57) are ordered like $x^t = (x_1^t \; x_2^t) = (\ldots \Delta x_i \, \Delta y_i \ldots \Delta\theta_i, \; \Delta\ln\kappa_i \ldots)$, its nullspace would read as

$$R(V^\perp) = R\begin{pmatrix} \cdot & \cdot & \cdot & \cdot \\ 1 & 0 & y^o_i & x^o_i \\ 0 & 1 & -x^o_i & y^o_i \\ \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ \cdot & \cdot & \cdot & \cdot \end{pmatrix} . \qquad (2.60)$$

A legitimate choice for defining an *S*-system would therefore be

$$\Delta x_r = \Delta y_r = \Delta\theta_r = \Delta\ln\kappa_r = 0 . \qquad (2.61)$$

That is, instead of fixing coordinates like we did in (2.4) we may just as well fix one network point, one direction of zero-reading and one scale parameter. The corresponding *S*-transformation then follows from (2.59) as

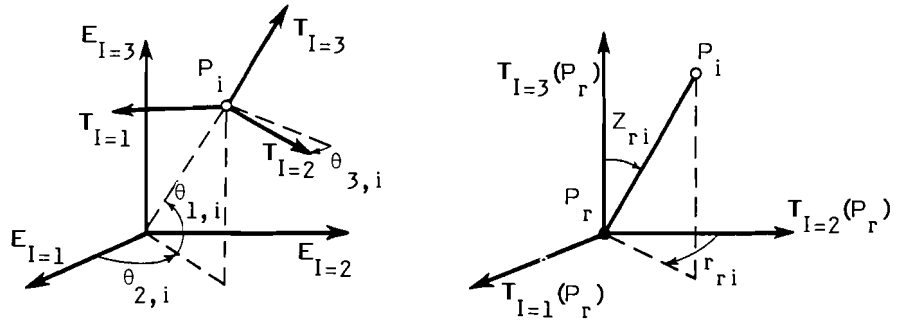
$$\begin{pmatrix} \Delta x_i \\ \Delta y_i \\ \Delta\theta_i \\ \Delta \ln\kappa_i \end{pmatrix}^{(r)} = \begin{pmatrix} \Delta x_i \\ \Delta y_i \\ \Delta\theta_i \\ \Delta \ln\kappa_i \end{pmatrix} - \begin{pmatrix} 1 & 0 & y^o_{ri} & x^o_{ri} \\ 0 & 1 & -x^o_{ri} & y^o_{ri} \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \Delta x_r \\ \Delta y_r \\ \Delta\theta_r \\ \Delta \ln\kappa_r \end{pmatrix}, \qquad (2.62)$$

where the upperindex (r) indicates that these parameters are defined through the restrictions (2.61). Note by the way that once one includes orientation- and scale parameters, one actually extends the notion of network configuration to cover both the point-configuration and attitudes of the theodolite frames. And in fact the direction- and pseudo-distance observables r_{ij} and l_{ij} are then interpretable as angles and distance ratios. They become the invariants of transformation (2.62).

Now let us return to the three dimensional case and generalize the foregoing to three dimensions. We will start by assuming that only horizontal- and vertical direction measurements r_{ij} and Z_{ij}, and pseudo-distance measurements l_{ij} are available. We consider the following two types of righthanded orthonormal triads (see figure 21).

1^o The reference frame E_I , $I = 1,2,3$;
It is to this reference frame that the coordinates x_i, y_i, z_i refer, i.e. the position vector of point P_i, denoted by $\mathbf{X}(P_i) = X^I(P_i)E_I$, has with respect to the frame E_I the components $X^{I=1}(P_i) = x_i$, $X^{I=2}(P_i) = y_i$, $X^{I=3}(P_i) = z_i$.

2^o The theodolite frame $T_I(P_i)$, $I = 1,2,3$, in point P_i;
$T_{I=3}$ points upwards in the direction of the theodolite's first axis,
$T_{I=2}$ points in the direction of zero reading, and
$T_{I=1}$ completes the right-handed system.



(a) figure 21 (b)

The relation between the two frames E_I and $T_I(P_i)$ is given by

$$\begin{pmatrix} T_{I=1}(P_i) \\ T_{I=2}(P_i) \\ T_{I=3}(P_i) \end{pmatrix} = R(\theta_{1,i},\theta_{2,i},\theta_{3,i}) \begin{pmatrix} E_{I=1} \\ E_{I=2} \\ E_{I=3} \end{pmatrix} = R(\theta_{3,i})R(\theta_{1,i},\theta_{2,i}) \begin{pmatrix} E_{I=1} \\ E_{I=2} \\ E_{I=3} \end{pmatrix}, \qquad (2.63)$$

where

$$R_3(\theta_{3,i}) = \begin{pmatrix} \cos\theta_{3,i} & -\sin\theta_{3,i} & 0 \\ \sin\theta_{3,i} & \cos\theta_{3,i} & 0 \\ 0 & 0 & 1 \end{pmatrix} ,$$

and

$$R(\theta_{1,i},\theta_{2,i}) = \begin{pmatrix} -\sin\theta_{2,i} & \cos\theta_{2,i} & 0 \\ -\sin\theta_{1,i}\cos\theta_{2,i} & -\sin\theta_{1,i}\sin\theta_{2,i} & \cos\theta_{1,i} \\ \cos\theta_{1,i}\cos\theta_{2,i} & \cos\theta_{1,i}\sin\theta_{2,i} & \sin\theta_{1,i} \end{pmatrix} .$$

Furthermore, we have for the difference vector $X(P_rP_i) = X(P_i) - X(P_r)$ between the two points P_i and P_r that,

$$X(P_rP_i) = (\kappa_r l_{ri}\sin Z_{ri}\sin r_{ri}, \kappa_r l_{ri}\sin Z_{ri}\cos r_{ri}, \kappa_r l_{ri}\cos Z_{ri}) \begin{pmatrix} T_{I=1}(P_r) \\ T_{I=2}(P_r) \\ T_{I=3}(P_r) \end{pmatrix} \tag{2.64}$$

where κ_r is a scale factor.
From (2.63), (2.64) and $X(P_rP_i) = X^I(P_i)E_I - X^I(P_r)E_I$ follows then that

$$\begin{pmatrix} x_i \\ y_i \\ z_i \end{pmatrix} = \begin{pmatrix} x_r \\ y_r \\ z_r \end{pmatrix} + \begin{pmatrix} -\sin\theta_{2,r} & -\sin\theta_{1,r}\cos\theta_{2,r} & \cos\theta_{1,r}\cos\theta_{2,r} \\ \cos\theta_{2,r} & -\sin\theta_{1,r}\sin\theta_{2,r} & \cos\theta_{1,r}\sin\theta_{2,r} \\ 0 & \cos\theta_{1,r} & \sin\theta_{1,r} \end{pmatrix} \begin{pmatrix} \kappa_r l_{ri}\sin Z_{ri}\sin(\theta_{3,r}+r_{ri}) \\ \kappa_r l_{ri}\sin Z_{ri}\cos(\theta_{3,r}+r_{ri}) \\ \kappa_r l_{ri}\cos Z_{ri} \end{pmatrix} , \tag{2.65}$$

which shows that one can start computing coordinates once the seven parameters $x_r, y_r, z_r, \theta_{1,r}, \theta_{2,r}, \theta_{3,r}$ and κ_r are fixed. Hence, a legitimate choice for defining an S-system would be

$$\Delta x_r = \Delta y_r = \Delta z_r = \Delta\theta_{1,r} = \Delta\theta_{2,r} = \Delta\theta_{3,r} = \Delta\ln\kappa_r = 0 . \tag{2.66}$$

Since (2.65) generalizes the first two equations of (2.58), linearization of (2.65) would give us the three dimensional analogue of the first two equations in (2.59). But this is of course only half of the story. We also need to know how the last two equations of (2.59) read in three dimensions. For scale this is trivial:

$$\Delta\ln\kappa_i = \Delta\ln l_{ri} - \Delta\ln l_{ir} + \Delta\ln\kappa_r . \tag{2.67}$$

To find the corresponding transformation for the orientational parameters though, we need to know how the orientational parameters $\theta_{1,i}, \theta_{2,i}, \theta_{3,i}$ in point P_i are affected by differential changes in the seven parameters $x_r, y_r, z_r, \theta_{1,r}, \theta_{2,r}, \theta_{3,r}$ and κ_r. Since we can rule out differential changes in the scale- and translational parameters, this leaves us with the problem of finding a differential relation which expresses the $\Delta\theta_{1,i}, \Delta\theta_{2,i}, \Delta\theta_{3,i}$ in terms of the observables and the parameters $\Delta\theta_{1,r}, \Delta\theta_{2,r}, \Delta\theta_{3,r}$.
Let us assume that the non-linear relation reads

$$\begin{pmatrix} T_{I=1}(P_i) \\ T_{I=2}(P_i) \\ T_{I=3}(P_i) \end{pmatrix} = K \begin{pmatrix} T_{I=1}(P_r) \\ T_{I=2}(P_r) \\ T_{I=3}(P_r) \end{pmatrix} , \tag{2.68}$$

where matrix K only contains functions of the observables.
With (2.63), it then follows from (2.68) that

$$R(\theta_{1,i},\theta_{2,i},\theta_{3,i}) = K.R(\theta_{1,r},\theta_{2,r},\theta_{3,r}).$$

Linearization gives,

$$\Delta R(\theta_{1,i},\theta_{2,i},\theta_{3,i}) = \Delta K \cdot R(\theta^o_{1,r},\theta^o_{2,r},\theta^o_{3,r}) + K^o.\Delta R(\theta_{1,r},\theta_{2,r},\theta_{3,r})$$

$$= \Delta K.R(\theta^o_{1,r},\theta^o_{2,r},\theta^o_{3,r}) + R(\theta^o_{1,i},\theta^o_{2,i},\theta^o_{3,i})R^{-1}(\theta^o_{1,r},\theta^o_{2,r},\theta^o_{3,r}).\Delta R(\theta_{1,r},\theta_{2,r},\theta_{3,r})$$

(2.69)

Since the first term on the right-hand side of (2.69) only contains observables, it is the second term we are really interested in. In components (2.69) reads then

$$(\Delta\theta_{1,i},\Delta\theta_{2,i},\Delta\theta_{3,i})^t = (\text{ observables })^t +$$

$$\begin{bmatrix} \cos(\theta^o_{2,i}-\theta^o_{2,r}) & 0 & -\cos\theta^o_{1,r}\sin(\theta^o_{2,i}-\theta^o_{2,r}) \\ \tan\theta^o_{1,i}\sin(\theta^o_{2,i}-\theta^o_{2,r}) & 1 & (-\sin\theta^o_{1,r}+\tan\theta^o_{1,i}\cos\theta^o_{1,r}\cos(\theta^o_{2,i}-\theta^o_{2,r})) \\ \cos^{-1}\theta^o_{1,i}\sin(\theta^o_{2,i}-\theta^o_{2,r}) & 0 & \cos^{-1}\theta^o_{1,i}\cos\theta^o_{1,r}\cos(\theta^o_{2,i}-\theta^o_{2,r}) \end{bmatrix} \begin{bmatrix} \Delta\theta_{1,r} \\ \Delta\theta_{2,r} \\ \Delta\theta_{3,r} \end{bmatrix}.$$

(2.70)

We are now in the position to collect our results. From (2.70), (2.67) and the linearized expression of (2.65) follows the three dimensional analogue of (2.59) as

$$(\ldots\Delta x_i,\Delta y_i,\Delta z_i,\ldots,\Delta\theta_{1,i},\Delta\theta_{2,i},\Delta\theta_{3,i},\Delta\ln\kappa_i,\ldots)^t = (\text{ observables })^t +$$

$$\begin{bmatrix} \cdots & \cdots & \cdots \\ 1 & 0 & -z^o_{ri}\cos\theta^o_{2,r} & 0 & -y^o_{ri} & (-z^o_{ri}\cos\theta^o_{1,r}\sin\theta^o_{2,r}+y^o_{ri}\sin\theta^o_{1,r}) \\ 0 & 1 & -z^o_{ri}\sin\theta^o_{2,r} & 0 & x^o_{ri} & (z^o_{ri}\cos\theta^o_{2,r}\cos\theta^o_{2,r}-x^o_{ri}\sin\theta^o_{1,r}) \\ 0 & 0 & (y^o_{ri}\sin\theta^o_{2,r}+x^o_{ri}\cos\theta^o_{2,r}) & 0 & (-y^o_{ri}\cos\theta^o_{1,r}\cos\theta^o_{2,r}+x^o_{ri}\cos\theta^o_{1,r}\sin\theta^o_{2,r}) \\ \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots \\ 0 & 0 & \cos(\theta^o_{2,i}-\theta^o_{2,r}) & 0 & -\cos\theta^o_{1,r}\sin(\theta^o_{2,i}-\theta^o_{2,r}) \\ 0 & 0 & \tan\theta^o_{1,i}\sin(\theta^o_{2,i}-\theta^o_{2,r}) & 1 & (-\sin\theta^o_{1,r}+\tan\theta^o_{1,i}\cos\theta^o_{1,r}\cos(\theta^o_{2,i}-\theta^o_{2,r})) \\ 0 & 0 & \cos^{-1}\theta^o_{1,i}\sin(\theta^o_{2,i}-\theta^o_{2,r}) & 0 & \cos^{-1}\theta^o_{1,i}\cos\theta^o_{1,r}\cos(\theta^o_{2,i}-\theta^o_{2,r}) \\ 0 & 0 & 0 & 0 & 0 & 1 \\ \cdots & \cdots & \cdots \end{bmatrix} \begin{bmatrix} \Delta x_r \\ \Delta y_r \\ \Delta z_r \\ \Delta\theta_{1,r} \\ \Delta\theta_{2,r} \\ \Delta\theta_{3,r} \\ \Delta\ln\kappa_r \end{bmatrix}$$

(2.71)

Thus if the unknowns in the linear model of the three-dimensional network are ordered like $(\ldots \Delta x_i \Delta y_i \Delta z_i \ldots \Delta\theta_{1,i} \Delta\theta_{2,i} \Delta\theta_{3,i} \Delta \ln\kappa_i \ldots)$, the linear model's nullspace would be spanned by the seven columns of the matrix on the right-hand side of (2.71). From this the with choice (2.66) corresponding S-transformation easily follows.

Note that so far we made no reference to the gravity field, i.e. the theodolite frames are allowed to assume any arbitrary atittude in space. Of course it is likely then, like it was in the two-dimensional case, that one has no special interest in computing the orientation- and scale unknowns. In such cases one would probably reduce these unknowns from the model, which would leave one with only coordinates. And then transformations like (2.56) will do.

Let us now assume that in addition to the horizontal- and vertical direction measurements r_{ij} and Z_{ij}, and pseudo-distance measurements l_{ij}, we also have the disposal of astronomical latitude Φ_i, longitude Λ_i and azimuth A_{ij}. We then need to introduce two new orthonormal triads:

3^o The earth-fixed frame ${}^*E_I$, $I=1,2,3$;
- ${}^*E_{I=3}$ points towards the average terrestrial pole (CIO),
- ${}^*E_{I=1}$ points towards the line of intersection of the plane of the average terrestrial equator and the plane containing the Greenwich vertical and parallel to ${}^*E_{I=3}$,
- ${}^*E_{I=2}$ completes the righthanded system.

4^o The local astronomical frame ${}^*T_I(P_i)$, $I=1,2,3$, in point P_i;
- ${}^*T_{I=3}$ points towards the local astronomical zenith,
- ${}^*T_{I=2}$ points towards north,
- ${}^*T_{I=1}$ points towards east.

If we assume that the theodolite frames are levelled, then the following relations between the four triads E_I, ${}^*E_I$, $T_I(P_i)$ and ${}^*T_I(P_i)$ hold:

$$\left.\begin{aligned} R(A_{ij}) \begin{pmatrix} {}^*T_{I=1}(P_i) \\ {}^*T_{I=2}(P_i) \\ {}^*T_{I=3}(P_i) \end{pmatrix} &= R(r_{ij}) \begin{pmatrix} T_{I=1}(P_i) \\ T_{I=2}(P_i) \\ T_{I=3}(P_i) \end{pmatrix} ; \quad \begin{pmatrix} {}^*T_{I=1}(P_i) \\ {}^*T_{I=2}(P_i) \\ {}^*T_{I=3}(P_i) \end{pmatrix} = R(\Phi_i, \Lambda_i) \begin{pmatrix} {}^*E_{I=1} \\ {}^*E_{I=2} \\ {}^*E_{I=3} \end{pmatrix} \\ \begin{pmatrix} T_{I=1}(P_i) \\ T_{I=2}(P_i) \\ T_{I=3}(P_i) \end{pmatrix} &= R(\theta_{3,i}) R(\theta_{1,i}, \theta_{2,i}) \begin{pmatrix} E_{I=1} \\ E_{I=2} \\ E_{I=3} \end{pmatrix} ; \quad \begin{pmatrix} {}^*E_{I=1} \\ {}^*E_{I=2} \\ {}^*E_{I=3} \end{pmatrix} = \bar{R}(\alpha,\beta,\gamma) \begin{pmatrix} E_{I=1} \\ E_{I=2} \\ E_{I=3} \end{pmatrix} \end{aligned}\right\} \quad (2.72)$$

where $\bar{R}(\alpha,\beta,\gamma) = \begin{pmatrix} 1 & \gamma & -\beta \\ -\gamma & 1 & \alpha \\ \beta & -\alpha & 1 \end{pmatrix}$, and α, β and γ are small rotation angles.

From (2.72) follows that

$$\bar{R}(\alpha,\beta,\gamma) = R^t(\Phi_r,\Lambda_r)R^t(A_{ri} - r_{ri})R(\theta_{3,r})R(\theta_{1,r},\theta_{2,r}) \quad . \tag{2.73}$$

Linearization, under the assumption that $\alpha^o = \beta^o = \gamma^o = 0$, gives

$$\begin{pmatrix} \Delta\alpha \\ \Delta\beta \\ \Delta\gamma \end{pmatrix} = \begin{pmatrix} \text{observables} \end{pmatrix} + \begin{pmatrix} \sin\Lambda_r^o & 0 & -\cos\Phi_r^o\cos\Lambda_r^o \\ -\cos\Lambda_r^o & 0 & -\cos\Phi_r^o\sin\Lambda_r^o \\ 0 & 1 & -\sin\Phi_r^o \end{pmatrix} \begin{pmatrix} \Delta\theta_{1,r} \\ \Delta\theta_{2,r} \\ \Delta\theta_{3,r} \end{pmatrix} \quad . \tag{2.74}$$

Since $\alpha^o = \beta^o = \gamma^o = 0$ we can replace $\theta_{1,i}^o$ and $\theta_{2,i}^o$ in (2.71) by Φ_i^o and Λ_i^o. With (2.74) we then find that for large three dimensional networks in which also astronomical latitude, longitude and azimuth are measured, (2.71) generalizes to

$$(..,\Delta x_i,\Delta y_i,\Delta z_i,..,\Delta\theta_{1,i},\Delta\theta_{2,i},\Delta\theta_{3,i},\Delta\ln\kappa_i,..\Delta\alpha,\Delta\beta,\Delta\gamma)^t =$$

$$(..\Delta x_i^{(r)},\Delta y_i^{(r)},\Delta z_i^{(r)},..\Delta\theta_{1,i}^{(r)},\Delta\theta_{2,i}^{(r)},\Delta\theta_{3,i}^{(r)},\Delta\ln\kappa_i^{(r)},..\Delta\alpha^{(r)},\Delta\beta^{(r)},\Delta\gamma^{(r)})^t +$$

$$\begin{pmatrix}
\cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\
1\ 0\ 0 & -z_{ri}^o\cos\Lambda_r^o & -y_{ri}^o & (-z_{ri}^o\cos\Phi_r^o\sin\Lambda_r^o + y_{ri}^o\sin\Phi_r^o) & x_{ri}^o & \Delta x_r \\
0\ 1\ 0 & -z_{ri}^o\sin\Lambda_r^o & x_{ri}^o & (z_{ri}^o\cos\Phi_r^o\cos\Lambda_r^o - x_{ri}^o\sin\Phi_r^o) & y_{ri}^o & \Delta y_r \\
0\ 0\ 1 & (y_{ri}^o\sin\Lambda_r^o + x_{ri}^o\cos\Lambda_r^o) & 0 & (-y_{ri}^o\cos\Phi_r^o\cos\Lambda_r^o + x_{ri}^o\cos\Phi_r^o\sin\Lambda_r^o) & z_{ri}^o & \Delta z_r \\
\cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\
0\ 0\ 0 & \cos(\Lambda_i^o - \Lambda_r^o) & 0 & -\cos\Phi_r^o\sin(\Lambda_i^o - \Lambda_r^o) & 0 & \Delta\theta_{1,r} \\
0\ 0\ 0 & \tan\Phi_i^o\sin(\Lambda_i^o - \Lambda_r^o) & 1 & (-\sin\Phi_r^o + \tan\Phi_i^o\cos\Phi_r^o\cos(\Lambda_i^o - \Lambda_r^o)) & 0 & \Delta\theta_{2,r} \\
0\ 0\ 0 & \cos^{-1}\Phi_i^o\sin(\Lambda_i^o - \Lambda_r^o) & 0 & \cos^{-1}\Phi_i^o\cos\Phi_r^o\cos(\Lambda_i^o - \Lambda_r^o) & 0 & \Delta\theta_{3,r} \\
\cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\
0\ 0\ 0 & 0 & 0 & 0 & 1 & \Delta\ln\kappa_r \\
\cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\
0\ 0\ 0 & \sin\Lambda_r^o & 0 & -\cos\Phi_r^o\cos\Lambda_r^o & 0 & \\
0\ 0\ 0 & -\cos\Lambda_r^o & 0 & -\cos\Phi_r^o\sin\Lambda_r^o & 0 & \\
0\ 0\ 0 & 0 & 1 & -\sin\Phi_r^o & 0 &
\end{pmatrix} \tag{2.75}$$

where we have denoted the first column vector on the right-hand side of (2.71) in which it says "observables", by

$$(\ldots\ \Delta x_i^{(r)}, \Delta y_i^{(r)}, \Delta z_i^{(r)} \ldots \Delta\theta_{1,i}^{(r)}, \Delta\theta_{2,i}^{(r)}, \Delta\theta_{3,i}^{(r)}, \Delta \ln\kappa_i^{(r)} \ldots \Delta\alpha^{(r)}, \Delta\beta^{(r)}, \Delta\gamma^{(r)})^t .$$

When viewing (2.75) one may wonder why there are still seven degrees of freedom. Aren't the Φ_i, Λ_i and A_{ij} supposed to take care of the rotational degrees of freedom? The reason for this apparent discrepancy is of course that the network's point configuration and fundamental directions are described with coordinates referring to the frame E_I, which is essentially an arbitrary one. We have chosen for this approach because it enables us to describe the most general situation, i.e. it allows us to introduce any reference system we like. That is, we do not restrict ourselves beforehand to those reference systems which might be the obvious ones to choose because of the available Φ_i's, Λ_i's and A_{ij}. But, would one aspire after this more conventional *S*-system definition, then decomposition formula (2.75) is easily modified. To see this, let us consider the two dimensional situation. Assume that azimuths A_{ij}, horizontal directions r_{ij} and distances l_{ij} are observed. By taking the general case of describing the network in an arbitrary system (see figure 22) we get from linearizing

$$\left.\begin{aligned} x_i &= x_r + \kappa l_{ri} \sin(A_{ri} - \alpha) \\ y_i &= y_r + \kappa l_{ri} \cos(A_{ri} - \alpha) \\ \theta_i &= A_{ri} - r_{ir} + \pi - \alpha \\ \ln\kappa &= \ln\kappa \\ \alpha &= \alpha \end{aligned}\right\} , \qquad (2.76)$$

that

$$\begin{pmatrix} \Delta x_i \\ \Delta y_i \\ \Delta\theta_i \\ \Delta \ln\kappa \\ \Delta\alpha \end{pmatrix} = \begin{pmatrix} \Delta x_i \\ \Delta y_i \\ \Delta\theta_i \\ \Delta \ln\kappa = 0 \\ \Delta\alpha = 0 \end{pmatrix}^{(r,//)} + \begin{pmatrix} 1 & 0 & -y^o_{ri} & x^o_{ri} \\ 0 & 1 & x^o_{ri} & y^o_{ri} \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} \Delta x_r \\ \Delta y_r \\ \Delta\alpha \\ \Delta \ln\kappa \end{pmatrix} , \qquad (2.77)$$

where the upper indices (r,//) indicate that these coordinates are computed in the *S*-system which is defined through fixing the point P_r $(\Delta x_r = \Delta y_r = 0)$, the scale parameter $(\Delta \ln\kappa = 0)$ and the orientation parallel (if $\alpha^o = 0$) to the north direction $(\Delta\alpha = 0)$.
From decomposing (2.77) like

$$
\left.
\begin{aligned}
\text{(a)}\quad
\begin{pmatrix} \Delta x_i \\ \Delta y_i \\ \Delta\theta_i \\ \Delta \ln\kappa \\ \Delta\alpha \end{pmatrix}
&= \begin{pmatrix} \Delta x_i \\ \Delta y_i \\ \Delta\theta_i \\ \Delta \ln\kappa = 0 \\ \Delta\alpha = 0 \end{pmatrix}^{(//)}
+ \begin{pmatrix} -y^o_{ri} & x^o_{ri} \\ x^o_{ri} & y^o_{ri} \\ -1 & 0 \\ 0 & 1 \\ 1 & 0 \end{pmatrix}
\begin{pmatrix} \Delta\alpha \\ \Delta \ln\kappa \end{pmatrix}, \\
\text{(b)}\quad
\begin{pmatrix} \Delta x_i \\ \Delta y_i \\ \Delta\theta_i \\ \Delta \ln\kappa = 0 \\ \Delta\alpha = 0 \end{pmatrix}^{(//)}
&= \begin{pmatrix} \Delta x_i \\ \Delta y_i \\ \Delta\theta_i \\ \Delta \ln\kappa = 0 \\ \Delta\alpha = 0 \end{pmatrix}^{(r,//)}
+ \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{pmatrix}
\begin{pmatrix} \Delta x_r \\ \Delta y_r \end{pmatrix},
\end{aligned}
\right\} \qquad (2.78)
$$

follows that the reference systems one usually considers when azimuths and distances are observed, are of the (//)-type. They are defined through $\Delta\alpha = 0$, $\Delta \ln\kappa = 0$. Thus, although it is usually not explicitly stated, the conventional S-system chosen when azimuths and distances are measured, is:

$$\Delta x_r = \Delta y_r = \Delta \ln\kappa = \Delta\alpha = 0 \ . \qquad (2.79)$$

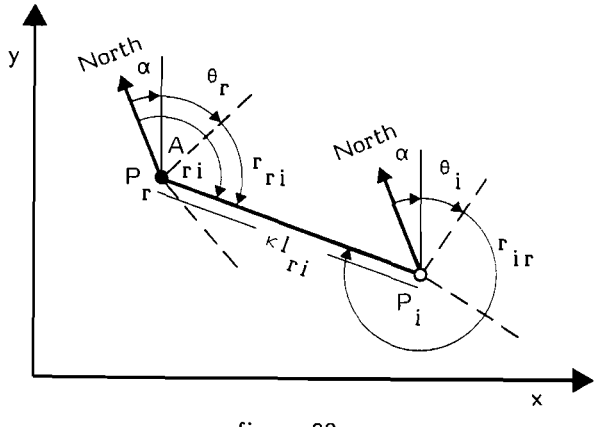


figure 22

In three dimensions (2.79) generalizes to

$$\Delta x_r = \Delta y_r = \Delta z_r = \Delta\alpha = \Delta\beta = \Delta\gamma = \Delta \ln\kappa = 0, \qquad (2.80)$$

and it will now be clear that the usual phrase "astronomical latitude, longitude and azimuth take care of the rotational degrees of freedom" essentially means that one has fixed the orientation of the reference system through $\Delta\alpha = \Delta\beta = \Delta\gamma = 0$.

From (2.75) follows that the with S-system (2.80) corresponding decomposition is given by:

$$
\begin{pmatrix} \Delta x_i \\ \Delta y_i \\ \Delta z_i \\ \vdots \\ \Delta\theta_{1,i} \\ \Delta\theta_{2,i} \\ \Delta\theta_{3,i} \\ \vdots \\ \Delta \ln\kappa \\ \Delta\alpha \\ \Delta\beta \\ \Delta\gamma \end{pmatrix} = \begin{pmatrix} \Delta x_i \\ \Delta y_i \\ \Delta z_i \\ \vdots \\ \Delta\theta_{1,i} \\ \Delta\theta_{2,i} \\ \Delta\theta_{3,i} \\ \vdots \\ \Delta \ln\kappa = 0 \\ \Delta\alpha = 0 \\ \Delta\beta = 0 \\ \Delta\gamma = 0 \end{pmatrix}^{(r,//)} + \begin{pmatrix} 1 & 0 & 0 & 0 & z^o_{ri} & -y^o_{ri} & x^o_{ri} \\ 0 & 1 & 0 & -z^o_{ri} & 0 & x^o_{ri} & y^o_{ri} \\ 0 & 0 & 1 & y^o_{ri} & -x^o_{ri} & 0 & z^o_{ri} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \sin\Lambda^o_i & -\cos\Lambda^o_i & 0 & 0 \\ 0 & 0 & 0 & -\tan\Phi^o_i \cos\Lambda^o_i & -\tan\Phi^o_i \sin\Lambda^o_i & 1 & 0 \\ 0 & 0 & 0 & -\cos^{-1}\Phi^o_i \cos\Lambda^o_i & -\cos^{-1}\Phi^o_i \sin\Lambda^o_i & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} \Delta x_r \\ \Delta y_r \\ \Delta z_r \\ \Delta\alpha \\ \Delta\beta \\ \Delta\gamma \\ \Delta \ln\kappa \end{pmatrix} \tag{2.81}
$$

The corresponding *S*-transformation is then easily found from bringing the second term on the right-hand side of (2.81) to the left-hand side (see also Teunissen, 1984a). Note that since $\Delta\alpha^{(r,//)} = \Delta\beta^{(r,//)} = \Delta\gamma^{(r,//)} = 0$, one can replace $\Delta\theta^{(r,//)}_{1,i}$ and $\Delta\theta^{(r,//)}_{2,i}$ in (2.81) by respectively $\Delta\Phi^{(r,//)}_i$ and $\Delta\Lambda^{(r,//)}_i$.

Instead of (2.80) one could of course also consider still other types of *S*-system definitions. One could for instance take the restrictions given by (2.54). The orientation of the earth-fixed frame ${}^*E_I$ and the directions of the local verticals are then given by respectively $\Delta\alpha^{(r,s;t)}$, $\Delta\beta^{(r,s;t)}$, $\Delta\gamma^{(r,s;t)}$ and $\Delta\theta^{(r,s;t)}_{1,i}$, $\Delta\theta^{(r,s;t)}_{2,i}$. And if one replaces the cartesian coordinates in e.g. (2.75) by geodetic coordinates and the direction unknowns $\Delta\theta_{1,i}$, $\Delta\theta_{2,i}$ by the deflection of the vertical components ξ_i, η_i through using

$$
\xi_i = \Delta\theta_{1,i} - \Delta\phi_i, \quad \eta_i = (\Delta\theta_{2,i} - \Delta\lambda_i) \cos\Phi^o_i,
$$

one can show that also the following sets of restrictions are legitimate choices for defining an *S*-system:

$$
\left.\begin{array}{ll}
\text{(a)} & \xi_r = \eta_r = \Delta A_{rs} = 0, \qquad \Delta\phi_r = \Delta\lambda_r = \Delta h_r = 0, \ \Delta \ln\kappa = 0 \\
 & (\hat{=}: \ \Delta\theta_{1,r} = \Delta\theta_{2,r} = \Delta A_{rs} = 0, \ \Delta\phi_r = \Delta\lambda_r = \Delta h_r = 0, \ \Delta \ln\kappa = 0) \\
\text{(b)} & \Delta\alpha = \Delta\beta = \Delta\gamma = 0, \qquad \xi_r = \eta_r = \Delta h_r = 0, \ \Delta \ln\kappa = 0 \\
\text{(c)} & \Delta\alpha = \Delta\beta = \Delta A_{rs} = 0, \qquad \Delta\phi_r = \Delta\lambda_r = \Delta h_r = 0, \ \Delta \ln\kappa = 0
\end{array}\right\} \tag{2.82}
$$

(see also Strang v. Hees, 1977; Yeremeyev and Yurkina, 1969). And in this way many more sets of

necessary and sufficient restrictions can be found. Note that also the geodetic coordinates should be given an upperindex referring to the S-system through which they are defined.

In principle of course there is no need for introducing deflection of the vertical components. For computing three dimensional networks one can just as well do without them. Due, however, to the fact that many existing large networks lack the necessary zenithdistances one has preferred in the past the classical method of reductions to a reference ellipsoid and computation by means of ellipsoidal quantities to the more theoretically attractive spatial triangulations of Bruns and Hotine (see e.g. Hotine, 1969; Torge and Wenzel, 1978; Engler et al. 1982). Instead of solving the height problem by using zenith distances one resorts to the astrogeodetic (or gravimetric) method. The problem of the network computation is then split into two nearly independent problems, namely the

$$\text{(a)} \quad \phi_i, \lambda_i \text{- problem, and the} \quad \text{(b)} \quad \xi_i, \eta_i, h_i \text{- problem} \; .$$

The procedure followed is in short the following (see also Heiskanen and Moritz, 1967). One starts by defining a three dimensional S-system (geodetic datum). Usually one takes the datum given by (2.82.b) or (2.82.c). Using the approximate information available on $\{\phi_i^o, \lambda_i^o, h_i^o, \Phi_i^o, \Lambda_i^o\}$ one then reduces the observed angles and distances to the ellipsoid and computes on it the geodetic coordinate increments $\Delta\phi_i, \Delta\lambda_i$. After having solved for (a), one enters the solution of (b) where new heights and new deflections of the vertical need to be determined based on the new ellipsoidal values of ϕ_i and λ_i. With these new values the whole procedure is repeated. One can consider this iteration procedure as a block Gauss-Seidel type of iteration where a linear system

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

is solved iteratively as

$$\begin{aligned} x_1^{(k+1)} &= A_{11}^{-1}(y_1 - A_{12}x_2^{(k)}) \\ x_2^{(k+1)} &= A_{22}^{-1}(y_2 - A_{21}x_1^{(k+1)}) \; . \end{aligned}$$

A practical point of concern is, however, the reduction procedure. In many cases the necessary gravity field information, needed to perform a proper reduction of the observational data, is lacking (see e.g. Meissl, 1973; Teunissen, 1982, 1983). But if the necessary gravity field information is available, the classical method of reduction to the ellipsoid can be seen to be formally equivalent to the truly three dimensional method and both methods, if applied correctly, will give the same results (Wolf, 1963a; Levallois, 1960). Hence, the final iterated solution of the classical method for the network's shape will be free form any deterministic effects of the arbitrarily introduced datum. The intermediate solutions of the iteration procedure, however, do theoretically depend on the choice of datum. It is gratifying to know therefore, as has been shown in subsection 2.2, that these effects are practically negligible.

3. (Free)networks and their connection

3.1. Types of networks considered

Now that we have given representations of $Nu(\mathbf{A})$ in various situations we can start discussing the problem of connecting geodetic networks.

In principle this problem is not too difficult. Essential is to know the type of information the two networks have in common. Based on this information one can then formulate the appropriate model and perform the adjustment.

As to the methods of connecting geodetic networks one can distinguish between three solution strategies. Two of them need the parameters, describing the two separate adjusted networks, while the third method starts from the assumption that the original observation equations (or rather the reduced normal equations) are still available.

In the first method (method I) use is made of condition equations. The idea is to eliminate first all non-common information from the two sets of parameters describing the two separate adjusted networks. This can be done by means of an appropriate S-transformation. The so transformed parameters are then finally used on an equal footing in the method of condition equations.

It is curious that this method has found so little attention in the literature. We only know of a few areas where it is applied (see e.g. Baarda, 1973; or Van Mierlo, 1978). An explanation could perhaps be the general aversion one has for the method of condition equations since it is known to be cumbersome in computation. However, for our present application of connecting networks this argument does not hold. On the contrary, the method can in many cases be very tractable indeed.

The second method (method II) is essentially the counterpart of the above mentioned method. In this method one starts by determining the transformation parameters. This is done by means of a least-squares adjustment. After the adjustment one then applies the transformation parameters to obtain the final estimates of the parameters describing the two connected networks.

Method II seems to be very popular with those working on the problem of connecting satellite networks with terrestrial networks (see e.g. Peterson, 1974). A serious shortcoming of most discussions on this method is, however, that often the starting assumptions are not explicitly formulated. As we will see this may avenge itself on the general applicability of the method and also may affect the interpretability of the transformation parameters.

Finally the third method (method III) makes use of the so-called Helmert blocking procedure. It is therefore essentially a phased type of adjustment, applied to the original models of the two overlapping networks (e.g. Wolf, 1978).

Usually when one applies this method one starts from the principle that both the reduced normals are regular, thereby suggesting that the two overlapping networks have no degrees of freedom at all. For a general application of the method, this is of course a too restrictive assumption to start with. We will therefore have to show how the method applies in the general case.

From the above few remarks it will be clear that we feel that a truly general discussion of the problem of connecting geodetic networks has not yet been given in the literature. Either the

assumptions are too restrictive to render a general application of the methods possible or they are not too precisely formulated.

For a proper course of things let us therefore start by stating our basic

Assumptions

First consider the original models. We assume that the first network is described by the linear(ized) model

$$\underset{m\times 1}{\Delta\tilde{y}} = \underset{m\times n}{(A_1} \vdots \underset{m\times n_2}{A_2)} \underset{(n+n_2)\times 1}{\begin{pmatrix}\Delta x_1\\ \Delta x_2\end{pmatrix}}, \quad Q_y, \text{ with dim. } Nu(A_1:A_2) = q, \tag{3.1.1.a}$$

and the second by

$$\underset{\bar{m}\times 1}{\Delta\bar{\tilde{y}}} = \underset{\bar{m}\times n}{(\bar{A}_1} \vdots \underset{\bar{m}\times n_3}{\bar{A}_3)} \underset{(n+n_3)\times 1}{\begin{pmatrix}\Delta\bar{x}_1\\ \Delta\bar{x}_3\end{pmatrix}}, \quad \underset{\bar{y}}{Q}, \text{ with dim. } Nu(\bar{A}_1:\bar{A}_3) = \bar{q} \geq q. \tag{3.1.1.b}$$

We further assume that the second network, apart from some additional degrees of freedom, has the same type of degrees of freedom as the first network. This means that we assume the nullspace of (3.1.1.a)'s normal reduced for Δx_2 to be a proper subspace of the nullspace of (3.1.1.b)'s normal reduced for $\Delta\bar{x}_3$, i.e.

$$Nu(P_2A_1) \subset Nu(\bar{P}_3\bar{A}_1), \tag{3.1.1.c}$$

with the projectors $P_2 = I - A_2(A_2^tQ_y^{-1}A_2)^{-1}A_2^tQ_y^{-1}$ and $\bar{P}_3 = I - \bar{A}_3(\bar{A}_3^t\bar{Q}_y^{-1}\bar{A}_3)^{-1}\bar{A}_3^t\bar{Q}_y^{-1}$.

And finally we assume that

$$\underset{(n+n_2+n_3)\times 1}{\begin{pmatrix}\Delta x_1\\ \Delta x_2\\ \Delta x_3\end{pmatrix}} = \underset{(n+n_2+n_3)\times 1}{\begin{pmatrix}\Delta\bar{x}_1\\ \Delta\bar{x}_2\\ \Delta\bar{x}_3\end{pmatrix}} + \underset{(n+n_2+n_3)\times r}{\begin{pmatrix}V_1^{\perp}\\ V_2^{\perp}\\ V_3^{\perp}\end{pmatrix}} \underset{r\times 1}{\Delta p}, \text{ with } r \geq \bar{q}, \text{ and} \tag{3.1.1.d}$$

$$Nu(\bar{P}_3\bar{A}_1) \subset R(V_1^{\perp}) \ . \tag{3.1.1.e}$$

Since some of the derivations and formulae in the next section become quite elaborate, we will use from time to time the following

Example

as reference to exemplify our results:

The first network can be thought of as being a planar network determined from distance -,

astronomical azimuth - and angle measurements. And the second network can considered to be planar with magnetic compass readings and angle observations only.

If the parameters $(\Delta x_1, \Delta x_2)$ and $(\Delta \bar{x}_1, \Delta \bar{x}_3)$ are assumed to contain cartesian coordinate increments only, then

$$Nu(A_1 \vdots A_2) = R\begin{pmatrix} \cdot & \cdot \\ \cdot & \cdot \\ 1 & 0 \\ 0 & 1 \\ \cdot & \cdot \\ \cdot & \cdot \end{pmatrix}, \text{ with } q = 2 \text{ and } Nu(\bar{A}_1 \vdots \bar{A}_3) = R\begin{pmatrix} \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ 1 & 0 & x_o^i \\ 0 & 1 & y_o^i \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \end{pmatrix}, \text{ with } q = 3.$$

The second network has namely apart from the two translational degrees of freedom also an additional freedom of scale.

Furthermore, transformation (3.1.1.d) would then be characterized by

$$R\begin{pmatrix} V_1^{\perp} \\ V_2^{\perp} \\ V_3^{\perp} \end{pmatrix} = R\begin{pmatrix} \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ y_o^i & 1 & 0 & x_o^i \\ -x_o^i & 0 & 1 & y_o^i \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \end{pmatrix}, \text{ with } r = 4$$

and the nullspaces of the reduced normals by

$$Nu(P_2 A_1) = R\begin{pmatrix} \cdot & \cdot \\ \cdot & \cdot \\ 1 & 0 \\ 0 & 1 \\ \cdot & \cdot \\ \cdot & \cdot \end{pmatrix}, \quad Nu(\bar{P}_3 \bar{A}_1) = R\begin{pmatrix} \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ 1 & 0 & x_o^i \\ 0 & 1 & y_o^i \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \end{pmatrix}.$$

Finally, using the decomposition

$$R(V_1^{\perp}) = R((V_1^{\perp})_1) \oplus R((V_1^{\perp})_2) = R((V_1^{\perp})_1) \oplus Nu(\bar{P}_3 \bar{A}_1),$$

with

$$R((V_1^{\perp})_1) = R\begin{pmatrix} \cdot \\ \cdot \\ y_o^i \\ -x_o^i \\ \cdot \\ \cdot \end{pmatrix} \text{ and } R((V_1^{\perp})_2) = R\begin{pmatrix} \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ 1 & 0 & x_o^i \\ 0 & 1 & y_o^i \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \end{pmatrix},$$

we can identify the Δp_1 parameter of $\Delta p = (\Delta p_1^t \; \Delta p_2^t)^t$ as a rotation angle and the Δp_2 parameters as respectively two translational and one scale parameter.

3.2. Three alternatives

Since the above mentioned first two methods are closely related we will discuss them together.

Method I and II

Both methods are applicable if the parameters, describing the two separate adjusted networks and their covariances are available. Thus we assume given (see figure 23):

$$\begin{bmatrix} \Delta\hat{x}_1^{(s)} \\ \Delta\hat{x}_2^{(s)} \end{bmatrix}, \quad \begin{bmatrix} Q_{\hat{x}_1^{(s)}} & Q_{\hat{x}_1^{(s)},\hat{x}_2^{(s)}} \\ Q_{\hat{x}_2^{(s)},\hat{x}_1^{(s)}} & Q_{\hat{x}_2^{(s)}} \end{bmatrix} \text{ and } \begin{bmatrix} \Delta\hat{\bar{x}}_1^{(\bar{s})} \\ \Delta\hat{\bar{x}}_2^{(\bar{s})} \end{bmatrix}, \quad \begin{bmatrix} Q_{\hat{\bar{x}}_1^{(\bar{s})}} & Q_{\hat{\bar{x}}_1^{(\bar{s})},\hat{\bar{x}}_3^{(\bar{s})}} \\ Q_{\hat{\bar{x}}_3^{(\bar{s})},\hat{\bar{x}}_1^{(\bar{s})}} & Q_{\hat{\bar{x}}_3^{(\bar{s})}} \end{bmatrix}$$

with (3.2.1)

$S = R(S)$ complementary to $Nu(A_1 : A_2)$ and $\bar{S} = R(\bar{S})$ complementary to $Nu(\bar{A}_1 : \bar{A}_3)$.

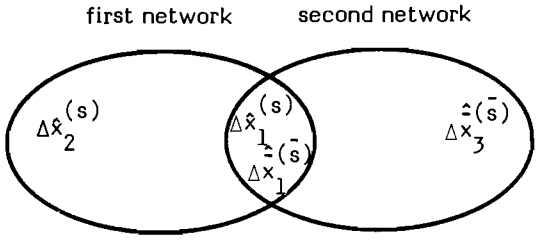


figure 23

Our goal is now, to solve for the transformation parameters $\Delta\hat{\hat{p}}$ and the increments $(\Delta\hat{x}_1^{(s)}, \Delta\hat{x}_2^{(s)}, \Delta\hat{x}_3^{(s)})$. Here we implicitly assume that we wish our results to be expressed in the same coordinate system as that of the first network. For our example in subsection 3.1 this means that we wish our results to take the scale and orientation of the first network. This is a sensible choice since the first network contains by assumption more information than the second $[\, Nu(P_2 A_1) \subset Nu(\bar{P}_3 \bar{A}_1)]$. But if one so desires one could also proceed otherwise, viz. by adopting the orientation of the second network.

We believe, that for **explanatory** purposes method I best shows the principles involved in connecting networks. Let us therefore first, before we proceed with the actual solution strategies of the two methods, consider the following simple but general enough situation. We assume to have measured two overlapping planar networks. And furthermore we assume that for both networks we have the disposal of distance -, azimuth and angle observations. When adjusting the two networks separately we thus need to take care of the in both cases existing translational degrees of freedom. But as we know from the previous section this can be done in very many ways. The simplest way being to fix just one network point. Having done this we thus finally end up with two sets of coordinates each

describing one of the two separate adjusted networks. How are we now to compare these two coordinate sets? Not by blithely comparing the coordinates of corresponding networkpoints for these were introduced in a rather arbitrary way. In general namely, the two fixed networkpoints will be different ones. In fact, even if one would have fixed the same networkpoint in both networks, one still should exercise great care. This is because the numerical values assigned to the fixed point need not be identical for both networks. Now if we disregard this possibility for the moment and assume that the same set of approximate coordinates are used for linearizing the observation equations of both networks, we would have the inequality

$$\Delta \tilde{\hat{x}}_1^{(s)} \neq \Delta \tilde{\bar{x}}_1^{(\bar{s})} ,$$

if

$$S \neq \bar{S} .$$

That is, the two sets of adjusted Δx_1- parameters cannot be compared directly. But we know already from the previous section that one can easily take care of this discrepancy by applying the appropriate S-transformation. This S-transformation should enable us then to compare corresponding coordinate differences.

Now let us change the situation slightly and assume that the azimuth measurements of the first network are of the astronomical type and those of the second network follow from magnetic compass readings. Then we would have

$$\Delta \tilde{\hat{x}}_1^{(s)} \neq \Delta \tilde{\bar{x}}_1^{(\bar{s})},$$

even if

$$S = \bar{S} .$$

The reason being of course that the first network is orientated with respect to astronomical north and the second with respect to magnetic north. Thus the only information the two networks have in common is of the distance- and angular type. But again we can take care of this discrepancy by using the appropriate S-transformation, namely one that eliminates the azimuthal information from both networks.

Finally we complicate the situation a bit further by assuming that the second network lacks distance measurements, i.e. lacks scale. In this case we are in the situation as described by the example of the previous subsection 3.1., because both networks then still have their translational degrees of freedom but now the second network also has an additional freedom of scale. In this case we thus certainly will have the inequality

$$\Delta \tilde{\hat{x}}_1^{(s)} \neq \Delta \tilde{\bar{x}}_1^{(\bar{s})},$$

irrespective the choices made for

$$S \text{ and } \bar{S} .$$

But as will be clear now, one can again overcome this discrepancy by using the appropriate S-

transformation, namely one which reduces both networks to ones of the angular type.
Summarizing, we can conclude from the above discussion that although the causes for the incompatibility of $\Delta\tilde{\underline{x}}_1^{(s)}$ and $\Delta\tilde{\tilde{\underline{x}}}_1^{(\bar{s})}$ may be different, one can always find the appropriate *S*-transformation to eliminate this discrepancy. And in view of our general assumptions (3.1.1) it follows that an appropriate *S*-transformation would be:

$$P_{R(\bar{\bar{S}}),\,R(V_1^{\perp})} = \bar{\bar{S}}(V_1^t\bar{\bar{S}})^{-1}V_1^t = I_n - V_1^{\perp}\left((\bar{\bar{S}}^{\perp})^t V_1^{\perp}\right)^{-1}(\bar{\bar{S}}^{\perp})^t \quad . \tag{3.2.2}$$

This would give us then

$$P_{R(\bar{\bar{S}}),\,R(V_1^{\perp})}(\Delta\tilde{\underline{x}}_1^{(s)} - \Delta\tilde{\tilde{\underline{x}}}_1^{(\bar{s})}) = 0 \quad , \tag{3.2.3}$$

or equivalently

$$V_1^t(\Delta\tilde{\underline{x}}_1^{(s)} - \Delta\tilde{\tilde{\underline{x}}}_1^{(\bar{s})}) = 0 \quad . \tag{3.2.4}$$

If the situation as sketched in the example of subsection 3.1 applies, (3.2.3) reads in cartesian coordinates as

$$(\ldots\Delta\tilde{x}_{1i}^{(\bar{\bar{s}})} - \Delta\tilde{\tilde{x}}_{1i}^{(\bar{\bar{s}})},\ \Delta\tilde{y}_{1i}^{(\bar{\bar{s}})} - \Delta\tilde{\tilde{y}}_{1i}^{(\bar{\bar{s}})},\ldots)^t = 0^t \quad , \quad i = 1,\ldots,\tfrac{1}{2}n. \tag{3.2.3'}$$

The equivalent formulation (3.2.4) represents then an independent set of n-4 angular condition equations

$$\Delta\tilde{\alpha}_{ijk} - \Delta\tilde{\tilde{\alpha}}_{ijk} = 0 \quad , \tag{3.2.4'}$$

or a set of n-4 linear equations which is in one-to-one correspondence to such a set of n-4 angular condition equations.

Some authors have expressed their hesitation towards the above described procedure for using *S*-transformations. They argue that by using an *S*-transformation which eliminates e.g. the available azimuthal and scale information, one eliminates information which is important in its own right. This, however, is in our opinion a missappreciation of the concept of *S*-transformations. The *S*-transformation is in the first instance only applied to obtain the equality (3.2.3) or (3.2.4), on which then the adjustment for connecting both networks is based. **After** the adjustment one can then always, if so desired, transform the adjusted coordinates back to one of the original coordinate-systems. In the above example for instance one can always transform back to the system of the first network, the one that contains scale- **and** orientation information.

Now let us consider the actual solution strategies of the two methods I and II. We will start with method I.
Although it is customary in the literature to start from modelformulation (3.2.3), we, for reasons yet to be explained, will start from modelformulation (3.2.4). Straightforward application of the least-

squares algoritm for the method of condition equations gives then

$$\begin{pmatrix} \Delta\hat{\underset{\sim}{x}}_1^{(s)} \\ \Delta\hat{\underset{\sim}{x}}_2^{(s)} \\ \hat{\underset{\sim}{x}}_1^{(s)} \\ \hat{\underset{\sim}{x}}_3^{(s)} \end{pmatrix} = \begin{pmatrix} \Delta\underset{\sim}{x}_1^{(s)} \\ \Delta\underset{\sim}{x}_2^{(s)} \\ \underset{\sim}{x}_1^{(s)} \\ \underset{\sim}{x}_3^{(s)} \end{pmatrix} - \begin{pmatrix} Q_{\underset{\sim}{x}_1^{(s)}} \\ Q_{\underset{\sim}{x}_2^{(s)}, \underset{\sim}{x}_1^{(s)}} \\ -Q_{\underset{\sim}{x}_1^{(s)}} \\ -Q_{\underset{\sim}{x}_3^{(s)}, \underset{\sim}{x}_1^{(s)}} \end{pmatrix} V_1 \left(V_1^t (Q_{\underset{\sim}{x}_1^{(s)}} + Q_{\hat{\underset{\sim}{x}}_1^{(\bar{s})}}) V_1 \right)^{-1} V_1^t (\Delta\underset{\sim}{x}_1^{(s)} - \Delta\hat{\underset{\sim}{x}}_1^{(\bar{s})}) \; . \tag{3.2.5}$$

This formulation of the least-squares solution of method I is however not yet in concurrence with the formulation one usually finds in the literature (see e.g. Baarda,1973,p.125 or Van Mierlo,1978,p.9-26). We therefore have to rewrite (3.2.5) a bit. For this purpose take the following abbreviation

$$A := P_{R(\bar{\bar{S}}), R(V_1^{\perp})} \left(Q_{\underset{\sim}{x}_1^{(s)}} + Q_{\hat{\underset{\sim}{x}}_1^{(\bar{s})}} \right) P^t_{R(\bar{\bar{S}}), R(V_1^{\perp})} \; . \tag{3.2.6}$$

Since $R(A) = R(\bar{\bar{S}})$ it follows, if B denotes an arbitrary inverse of A, that AB is a projector which projects onto $R(\bar{\bar{S}})$ and along a complementary subspace.
Hence

$$AB.P_{R(\bar{\bar{S}}), R(V_1^{\perp})} = P_{R(\bar{\bar{S}}), R(V_1^{\perp})} \; .$$

From premultiplying this expression with $V_1 \left(V_1^t (Q_{\underset{\sim}{x}_1^{(s)}} + Q_{\hat{\underset{\sim}{x}}_1^{(\bar{s})}}) V_1 \right)^{-1} V_1^t$ follows then

$$V_1 \left(V_1^t (Q_{\underset{\sim}{x}_1^{(s)}} + Q_{\hat{\underset{\sim}{x}}_1^{(\bar{s})}}) V_1 \right)^{-1} V_1^t . AB . P_{R(\bar{\bar{S}}), R(V_1^{\perp})} = V_1 \left(V_1^t (Q_{\underset{\sim}{x}_1^{(s)}} + Q_{\hat{\underset{\sim}{x}}_1^{(\bar{s})}}) V_1 \right)^{-1} V_1^t . P_{R(\bar{\bar{S}}), R(V_1^{\perp})}$$

or

$$P^t_{R(\bar{\bar{S}}), R(V_1^{\perp})} . B . P_{R(\bar{\bar{S}}), R(V_1^{\perp})} = V_1 \left(V_1^t (Q_{\underset{\sim}{x}_1^{(s)}} + Q_{\hat{\underset{\sim}{x}}_1^{(\bar{s})}}) V_1 \right)^{-1} V_1^t \; . \tag{3.2.7}$$

Hence, if we use the customary notation

$$\hat{d}^{(\bar{\bar{s}})} := P_{R(\bar{\bar{S}}), R(V_1^{\perp})} (\Delta\underset{\sim}{x}_1^{(s)} - \Delta\hat{\underset{\sim}{x}}_1^{(\bar{s})})$$

and

$$Q^-_{\hat{d}^{(\bar{\bar{s}})}} := B,$$

we can rewrite (3.2.5) as

$$\begin{pmatrix} \Delta\hat{\underline{x}}_1^{(s)} \\ \Delta\hat{\underline{x}}_2^{(s)} \\ \Delta\hat{\underline{\underline{x}}}_1^{(\bar{s})} \\ \Delta\hat{\underline{\underline{x}}}_3^{(\bar{s})} \end{pmatrix} = \begin{pmatrix} \Delta\underline{x}_1^{(s)} \\ \Delta\underline{x}_2^{(s)} \\ \Delta\underline{\underline{x}}_1^{(\bar{s})} \\ \Delta\underline{\underline{x}}_3^{(\bar{s})} \end{pmatrix} - \begin{pmatrix} Q_{\underline{x}_1^{(s)}, \hat{d}^{(\bar{\bar{s}})}} \\ Q_{\underline{x}_2^{(s)}, \hat{d}^{(\bar{\bar{s}})}} \\ Q_{\underline{\underline{x}}_1^{(\bar{s})}, \hat{d}^{(\bar{\bar{s}})}} \\ Q_{\underline{\underline{x}}_3^{(\bar{s})}, \hat{d}^{(\bar{\bar{s}})}} \end{pmatrix} Q^{-}_{\hat{d}^{(\bar{\bar{s}})}} \hat{d}^{(\bar{\bar{s}})} \quad . \tag{3.2.8}$$

To finally transform the adjusted parameters $(\Delta\hat{\underline{\underline{x}}}_1^{(\bar{s})}, \Delta\hat{\underline{\underline{x}}}_3^{(\bar{s})})$ to the coordinate system of the first network, we need to determine the transformation parameters $\Delta\hat{\underline{\underline{p}}}$. From (3.1.1.d) follows that

$$\Delta p = \left((\bar{\bar{S}}^{\perp})^t V_1^{\perp} \right)^{-1} (\bar{\bar{S}}^{\perp})^t (\Delta x_1^{(s)} - \Delta\bar{x}_1^{(\bar{s})}),$$

with $R(\bar{\bar{S}})$ complementary to $R(V_1^{\perp})$. Hence the transformation parameters are easily found through

$$\Delta\hat{\underline{\underline{p}}} = \left((\bar{\bar{S}}^{\perp})^t V_1^{\perp} \right)^{-1} (\bar{\bar{S}}^{\perp})^t (\Delta\hat{\underline{x}}_1^{(s)} - \Delta\hat{\underline{\underline{x}}}_1^{(\bar{s})}). \tag{3.2.9}$$

Summarizing, we can thus write the solution of method I as:

(a)
$$\begin{pmatrix} \Delta\hat{\underline{x}}_1^{(s)} \\ \Delta\hat{\underline{x}}_2^{(s)} \\ \Delta\hat{\underline{\underline{x}}}_1^{(\bar{s})} \\ \Delta\hat{\underline{\underline{x}}}_3^{(\bar{s})} \end{pmatrix} = \begin{pmatrix} \Delta\underline{x}_1^{(s)} \\ \Delta\underline{x}_2^{(s)} \\ \Delta\underline{\underline{x}}_1^{(\bar{s})} \\ \Delta\underline{\underline{x}}_3^{(\bar{s})} \end{pmatrix} - \begin{pmatrix} Q_{\underline{x}_1^{(s)}, \hat{d}^{(\bar{\bar{s}})}} \\ Q_{\underline{x}_2^{(s)}, \hat{d}^{(\bar{\bar{s}})}} \\ Q_{\underline{\underline{x}}_1^{(\bar{s})}, \hat{d}^{(\bar{\bar{s}})}} \\ Q_{\underline{\underline{x}}_3^{(\bar{s})}, \hat{d}^{(\bar{\bar{s}})}} \end{pmatrix} Q^{-}_{\hat{d}^{(\bar{\bar{s}})}} \hat{d}^{(\bar{\bar{s}})}, \text{ with}$$

$$\hat{d}^{(\bar{\bar{s}})} = P_{R(\bar{\bar{S}}), R(V_1^{\perp})} (\Delta\hat{\underline{x}}_1^{(s)} - \Delta\hat{\underline{\underline{x}}}_1^{(\bar{s})}), \quad \text{and}$$

$Q^{-}_{\hat{d}^{(\bar{\bar{s}})}}$ an arbitrary inverse of $Q_{\hat{d}^{(\bar{\bar{s}})}}$.

(b)
$$\Delta\hat{\underline{\underline{p}}} = \left((\bar{\bar{S}}^{\perp})^t V_1^{\perp} \right)^{-1} (\bar{\bar{S}}^{\perp})^t (\Delta\hat{\underline{x}}_1^{(s)} - \Delta\hat{\underline{\underline{x}}}_1^{(\bar{s})}),$$

(c)
$$\begin{pmatrix} \Delta\hat{\underline{x}}_1^{(s)} \\ \Delta\hat{\underline{x}}_3^{(s)} \end{pmatrix} = \begin{pmatrix} \Delta\hat{\underline{\underline{x}}}_1^{(\bar{s})} \\ \Delta\hat{\underline{\underline{x}}}_3^{(\bar{s})} \end{pmatrix} + \begin{pmatrix} V_1^{\perp} \\ V_3^{\perp} \end{pmatrix} \Delta\hat{\underline{\underline{p}}}.$$

(3.2.10)

This is also the solution one can find in (Baarda, 1973) although there the result is derived under the more restrictive assumptions that $Nu(P_2A_1) = Nu(\bar{P}_3\bar{A}_1) = R(V_1^{\perp})$.

When comparing (3.2.10.a) with (3.2.5) one may wonder which formulation is the more attractive computationwise. Formulation (3.2.10.a) suggests the customary practice of first appying an S-transformation, namely (3.2.2), and then computing the inverse $Q_{\hat{d}}^{-}(\bar{s})$. A more direct way is however suggested by (3.2.7) and the method of prolongation discussed in sections 4 and 5 of chapter II.

Note, that in the special case of $Nu(P_2A_1) = Nu(\bar{P}_3\bar{A}_1) = R(V_1^{\perp})$, $n_2 = n_3 = 0$ and $S = \bar{S}$, $V_1\{V_1^t(Q_{\hat{x}_1}(s) + Q_{\hat{x}_1}(\bar{s}))V_1\}^{-1}V_1^t$ is a symmetric minimum rank inverse of $Q_{\hat{x}_1}(s) + Q_{\hat{x}_1}(\bar{s})$. With our expression (4.5) of chapter II follows then that (3.2.7) can be computed from

$$\begin{pmatrix} Q_{\hat{x}_1}(s) + Q_{\hat{x}_1}(\bar{s}) & V_1^{\perp} \\ (V_1^{\perp})^t & 0 \end{pmatrix}^{-1} = \begin{pmatrix} V_1\{V_1^t(Q_{\hat{x}_1}(s) + Q_{\hat{x}_1}(\bar{s}))V_1\}^{-1}V_1^t & V_1^{\perp}((V_1^{\perp})^tV_1^{\perp})^{-1} \\ ((V_1^{\perp})^tV_1^{\perp})^{-1}(V_1^{\perp})^t & 0 \end{pmatrix} . \tag{3.2.11}$$

In the general case that $Nu(P_2A_1) \subset Nu(\bar{P}_3\bar{A}_1) \subset R(V_1^{\perp})$, (3.2.7) will cease to be a minimum rank inverse of $Q_{\hat{x}_1}(s) + Q_{\hat{x}_1}(\bar{s})$. Instead it becomes a constrained inverse of $Q_{\hat{x}_1}(s) + Q_{\hat{x}_1}(\bar{s})$. With (5.21) of chapter II follows then

$$\begin{pmatrix} Q_{\hat{x}_1}(s) + Q_{\hat{x}_1}(\bar{s}) & V_1^{\perp} \\ (V_1^{\perp})^t & 0 \end{pmatrix}^{-1} = \begin{pmatrix} V_1\{V_1^t(Q_{\hat{x}_1}(s) + Q_{\hat{x}_1}(\bar{s}))V_1\}^{-1}V_1^t & * * * \\ * * * & * * * \end{pmatrix} . \tag{3.2.11'}$$

Thus, since a representation of $R(V_1^{\perp})$ is usually readily available, we see that instead of (3.2.10.a) one can also use formulation (3.2.5) with (3.2.7) computed via (3.2.11') (or (3.2.11)).

Now let us consider method II. Its model formulation is the parametric counterpart of (3.2.4) and reads as

$$\Delta\tilde{\hat{x}}_1(s) - \Delta\tilde{\hat{x}}_1(\bar{s}) = V_1^{\perp}\Delta p . \tag{3.2.12}$$

Usually this model will constitute the differential similarity transformation

$$\begin{pmatrix} \vdots \\ \Delta x_i \\ \Delta y_i \\ \Delta z_i \\ \vdots \end{pmatrix}(s) - \begin{pmatrix} \vdots \\ \Delta x_i \\ \Delta y_i \\ \Delta z_i \\ \vdots \end{pmatrix}(\bar{s}) = \begin{pmatrix} \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 0 & 0 & -z_i^o & y_i^o & x_i^o \\ 0 & 1 & 0 & z_i^o & 0 & -x_i^o & y_i^o \\ 0 & 0 & 1 & -y_i^o & x_i^o & 0 & z_i^o \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \end{pmatrix} \begin{pmatrix} \Delta t^x \\ \Delta t^y \\ \Delta t^z \\ \Delta\varepsilon^x \\ \Delta\varepsilon^y \\ \Delta\kappa^z \end{pmatrix} ,$$

e.g. when combining doppler networks with terrestrial networks (Peterson, 1974). However, since the common unknowns of the two overlapping networks need not be restricted to coordinates, relation (3.2.12) could be a kind of modified differential similarity transformation such as for instance (2.81). In fact, relation (3.2.12) need not be restricted to the differential similarity transformation at all. It could for instance also include additional "transformation" parameters which describe projected geophysical hypotheses in a deformation analysis. Or it could include, say, a refraction model.

When we solve for (3.2.12) we immediately notice a difficulty which is often overlooked in the literature. Namely, that the covariance sum $Q_{\hat{x}_1}(s) + Q_{\hat{\bar{x}}_1}(\bar{s})$ can turn out to be singular. Assume for instance that $S = R(S)$ is complementary to $Nu(P_2\bar{A}_1)$, $\bar{S} = R(\bar{S})$ is complementary to $Nu(\bar{P}_3\bar{A}_1)$ and that $\bar{S} \subset S$. Then $Nu(Q_{\hat{x}_1}(s) + Q_{\hat{\bar{x}}_1}(\bar{s})) \neq \{0\}$ and no ordinary inverse of $Q_{\hat{x}_1}(s) + Q_{\hat{\bar{x}}_1}(\bar{s})$ will exist. One could of course ask oneselves then whether it is possible to take a generalized inverse of $Q_{\hat{x}_1}(s) + Q_{\hat{\bar{x}}_1}(\bar{s})$. In some cases this is possible. We will refrain however from further elaboration on this point, since if one really insists on using (3.2.12), one can either transform one of the covariance matrices by means of an appropriate S-transformation so that the sum $Q_{\hat{x}_1}(s) + Q_{\hat{\bar{x}}_1}(\bar{s})$ becomes regular again, or, what is more practical, add the matrix $(V_1^{\perp})(V_1^{\perp})^t$ to $Q_{\hat{x}_1}(s) + Q_{\hat{\bar{x}}_1}(\bar{s})$. The solution of (3.2.12) follows then from straightforward application of the least-squares algorithm. To show the close relationship with solution (3.2.10) we will make use of a slight detour.

First consider the transformation parameters. With the aid of (3.2.5) we can write (3.2.9) as

$$\Delta\hat{\hat{p}} = \left[(\bar{\bar{S}}^{\perp})^t V_1^{\perp} \right]^{-1} (\bar{\bar{S}}^{\perp})^t \left[I - (Q_{\hat{x}_1}(s) + Q_{\hat{\bar{x}}_1}(\bar{s}))V_1 \left[V_1^t (Q_{\hat{x}_1}(s) + Q_{\hat{\bar{x}}_1}(\bar{s}))V_1 \right]^{-1} V_1^t \right] (\Delta\hat{x}_1^{(s)} - \Delta\hat{\bar{x}}_1(\bar{s})) \tag{3.2.13}$$

And since we have the projector identity

$$I - (Q_{\hat{x}_1}(s) + Q_{\hat{\bar{x}}_1}(\bar{s}))V_1 \left[V_1^t (Q_{\hat{x}_1}(s) + Q_{\hat{\bar{x}}_1}(\bar{s}))V_1 \right]^{-1} V_1^t =$$

$$= V_1^{\perp} \left[(V_1^{\perp})^t \left[Q_{\hat{x}_1}(s) + Q_{\hat{\bar{x}}_1}(\bar{s}) + (V_1^{\perp})(V_1^{\perp})^t \right]^{-1} V_1^{\perp} \right]^{-1} (V_1^{\perp})^t \left[Q_{\hat{x}_1}(s) + Q_{\hat{\bar{x}}_1}(\bar{s}) + (V_1^{\perp})(V_1^{\perp})^t \right]^{-1},$$

it follows from (3.2.13) that

$$\Delta\hat{\hat{p}} = \left[(V_1^{\perp})^t \left[Q_{\hat{x}_1}(s) + Q_{\hat{\bar{x}}_1}(\bar{s}) + (V_1^{\perp})(V_1^{\perp})^t \right]^{-1} V_1^{\perp} \right]^{-1} (V_1^{\perp})^t \left[Q_{\hat{x}_1}(s) + Q_{\hat{\bar{x}}_1}(\bar{s}) + (V_1^{\perp})(V_1^{\perp})^t \right]^{-1} (\Delta\hat{x}_1^{(s)} - \Delta\hat{\bar{x}}_1(\bar{s})) \tag{3.2.14.a}$$

In a similar way one can prove that

$$\begin{bmatrix} \Delta\hat{\underset{\sim}{x}}_1^{(s)} \\ \Delta\hat{\underset{\sim}{x}}_2^{(s)} \\ \Delta\hat{\underset{\sim}{\hat{x}}}_1^{(\bar{s})} \\ \Delta\hat{\underset{\sim}{\hat{x}}}_3^{(\bar{s})} \end{bmatrix} = \begin{bmatrix} \Delta\underset{\sim}{x}_1^{(s)} \\ \Delta\underset{\sim}{x}_2^{(s)} \\ \Delta\underset{\sim}{\hat{x}}_1^{(\bar{s})} \\ \Delta\underset{\sim}{\hat{x}}_3^{(\bar{s})} \end{bmatrix} - \begin{bmatrix} Q_{\underset{\sim}{x}_1^{(s)}} \\ Q_{\underset{\sim}{x}_2^{(s)},\underset{\sim}{x}_1^{(s)}} \\ -Q_{\underset{\sim}{\hat{x}}_1^{(\bar{s})}} \\ -Q_{\underset{\sim}{\hat{x}}_3^{(\bar{s})},\underset{\sim}{\hat{x}}_1^{(\bar{s})}} \end{bmatrix} \left[Q_{\underset{\sim}{x}_1^{(s)}} + Q_{\underset{\sim}{\hat{x}}_1^{(\bar{s})}} + (V_1^{\perp})(V_1^{\perp})^t \right]^{-1} (\Delta\underset{\sim}{x}_1^{(s)} - \Delta\underset{\sim}{\hat{x}}_1^{(\bar{s})} - V_1^{\perp}\Delta\hat{p}) \qquad (3.2.14.b)$$

Summarizing, we can thus write the solution of method II as:

$$\text{(a)}\quad \Delta\hat{p} = \left[(V_1^{\perp})^t \left[Q_{\underset{\sim}{x}_1^{(s)}} + Q_{\underset{\sim}{\hat{x}}_1^{(\bar{s})}} + (V_1^{\perp})(V_1^{\perp})^t \right]^{-1} V_1^{\perp} \right]^{-1} (V_1^{\perp})^t \left[Q_{\underset{\sim}{x}_1^{(s)}} + Q_{\underset{\sim}{\hat{x}}_1^{(\bar{s})}} + (V_1^{\perp})(V_1^{\perp})^t \right]^{-1} (\Delta\underset{\sim}{x}_1^{(s)} - \Delta\underset{\sim}{\hat{x}}_1^{(\bar{s})})$$

$$\text{(b)}\quad \begin{bmatrix} \Delta\hat{\underset{\sim}{x}}_1^{(s)} \\ \Delta\hat{\underset{\sim}{x}}_2^{(s)} \\ \Delta\hat{\underset{\sim}{\hat{x}}}_1^{(\bar{s})} \\ \Delta\hat{\underset{\sim}{\hat{x}}}_3^{(\bar{s})} \end{bmatrix} = \begin{bmatrix} \Delta\underset{\sim}{x}_1^{(s)} \\ \Delta\underset{\sim}{x}_2^{(s)} \\ \Delta\underset{\sim}{\hat{x}}_1^{(\bar{s})} \\ \Delta\underset{\sim}{\hat{x}}_3^{(\bar{s})} \end{bmatrix} - \begin{bmatrix} Q_{\underset{\sim}{x}_1^{(s)}} \\ Q_{\underset{\sim}{x}_2^{(s)},\underset{\sim}{x}_1^{(s)}} \\ -Q_{\underset{\sim}{\hat{x}}_1^{(\bar{s})}} \\ -Q_{\underset{\sim}{\hat{x}}_3^{(\bar{s})},\underset{\sim}{\hat{x}}_1^{(\bar{s})}} \end{bmatrix} \left[Q_{\underset{\sim}{x}_1^{(s)}} + Q_{\underset{\sim}{\hat{x}}_1^{(\bar{s})}} + (V_1^{\perp})(V_1^{\perp})^t \right]^{-1} (\Delta\underset{\sim}{x}_1^{(s)} - \Delta\underset{\sim}{\hat{x}}_1^{(\bar{s})} - V_1^{\perp}\Delta\hat{p})$$

$$\text{(c)}\quad \begin{bmatrix} \Delta\hat{\underset{\sim}{x}}_1^{(s)} \\ \Delta\hat{\underset{\sim}{x}}_3^{(s)} \end{bmatrix} = \begin{bmatrix} \Delta\hat{\underset{\sim}{\hat{x}}}_1^{(\bar{s})} \\ \Delta\hat{\underset{\sim}{\hat{x}}}_3^{(\bar{s})} \end{bmatrix} + \begin{bmatrix} V_1^{\perp} \\ V_3^{\perp} \end{bmatrix} \Delta\hat{p}$$

(3.2.15)

For the special case that $Q_{\underset{\sim}{x}_1^{(s)}} + Q_{\underset{\sim}{\hat{x}}_1^{(\bar{s})}}$ itself is regular, (3.2.15) without the additional term $(V_1^{\perp})(V_1^{\perp})^t$ is the solution one usually finds cited in the literature (e.g. Adam et al., 1982). However, the necessary relation with $Nu(P_2 A_1)$ and $Nu(\bar{P}_3 \bar{A}_1)$ is usually not made.
Note by the way that $\left[Q_{\underset{\sim}{x}_1^{(s)}} + Q_{\underset{\sim}{\hat{x}}_1^{(\bar{s})}} + (V_1^{\perp})(V_1^{\perp})^t \right]$ is a symmetric maximum rank inverse of (3.2.7).

For those who are used to thinking in terms of *S*-systems, it may come as a surprise that one is allowed to simply add the covariance maps $Q_{\underset{\sim}{x}_1^{(s)}}$ and $Q_{\underset{\sim}{\hat{x}}_1^{(\bar{s})}}$ of coordinates defined in **different** *S*-systems. *The reason is that the transformation parameters* Δp *in model formulation* (3.2.12) already take care of the possible discrepancy between the two *S*-systems.

This brings us to another important point, namely that of the interpretability of the transformation parameters $\Delta\hat{\tilde{p}}$. A shallow study of (3.2.15) might convince us that all transformation parameters are estimable and that one is allowed, in the context of testing alternative hypotheses, to test whether some or all of the transformation parameters are significant or not. Here, however, one should exercise great care. In particular one should be aware that one can not test whether an arbitrary linear function of the transformation parameters, $\underset{1xr}{c^t}\ \underset{rx1}{\Delta p}$ say, is zero or not, i.e.:

$$H_o: \Delta\hat{\tilde{x}}_1^{(s)} - \Delta\hat{\bar{\tilde{x}}}_1^{(\bar{s})} = V_1^{\perp}\Delta p, \qquad c^t\Delta p = 0,$$

against

$$H_A: \Delta\hat{\tilde{x}}_1^{(s)} - \Delta\hat{\bar{\tilde{x}}}_1^{(\bar{s})} = V_1^{\perp}\Delta p, \qquad c^t\Delta p \neq 0.$$

The reason is that, in the general case we are considering here, one cannot treat all transformation parameters on an equal footing. In case of our example of subsection 3.1, for instance, only the orientational parameter Δp_1 is eligible for a test like above.

Finally we like to point out the great resemblance between (3.2.10) and (3.2.15). The two methods only differ in their order of computing the transformation parameters $\Delta\hat{\tilde{p}}$ and increments $(\Delta\hat{\tilde{x}}_1^{(s)},\ \Delta\hat{\tilde{x}}_2^{(s)},\ \Delta\hat{\bar{\tilde{x}}}_1^{(\bar{s})},\ \Delta\hat{\bar{\tilde{x}}}_3^{(\bar{s})})$. Hence, in principle no preference can be given to either method, unless one chooses on the basis of computational convenience. One can argue namely that method I is to be preferred since it only needs the inverse of the covariance matrix of the difference vector $\hat{d}^{(\bar{s})}$ or (3.2.7), whereas method II needs the inverse of both $Q_{\hat{x}_1}(s) + Q_{\hat{x}_1}^{\hat{}}(\bar{s}) + (V_1^{\perp})(V_1^{\perp})^t$ and $(V_1^{\perp})^t[Q_{\hat{x}_1}(s) + Q_{\hat{x}_1}^{\hat{}}(\bar{s}) + (V_1^{\perp})(V_1^{\perp})^t]^{-1}(V_1^{\perp})$.

Let us now consider

method III

The Helmert-blocking method is essentially a phased type of adjustment applied to a second standard problem formulation. Instead of performing the adjustment in one step, the original set of observation equations is divided into two groups, each describing one of the two overlapping networks. After having formed the corresponding normalsystems one then reduces to obtain the reduced normals pertaining to the common unknowns of the two networks. Through inversion of the sum of these reduced normals one solves for the final adjusted values of the common unknowns. The remaining unknowns are found by means of back-substitution (e.g. Wolf, 1978).

If we reduce for the Δx_2 -parameters, (3.1.1.a)'s normalsystem becomes

$$\begin{bmatrix} (P_2A_1)^tQ_y^{-1}(P_2A_1) & 0 \\ A_2^tQ_y^{-1}A_1 & A_2^tQ_y^{-1}A_2 \end{bmatrix}\begin{bmatrix}\Delta x_1 \\ \Delta x_2\end{bmatrix} = \begin{bmatrix} (P_2A_1)^tQ_y^{-1}\Delta y \\ A_2^tQ_y^{-1}\Delta y \end{bmatrix}.$$

Hence as a solution of (3.1.1.a) we have

$$
\left.
\begin{aligned}
1^{o}\quad & \Delta\hat{x}_1^{(s)} = S\{S^t(P_2A_1)^tQ_y^{-1}(P_2A_1)S\}^{-1}S^t(P_2A_1)^tQ_y^{-1}\Delta y \\
2^{o}\quad & \Delta\hat{x}_2^{(s)} = (A_2^tQ_y^{-1}A_2)^{-1}A_2^tQ_y^{-1}(\Delta y - A_1\Delta\hat{x}_1^{(s)}), \text{ with} \\
& P_2 = I - A_2(A_2^tQ_y^{-1}A_2)^{-1}A_2^tQ_y^{-1}, \text{ and} \\
& S = \mathcal{R}(S) \text{ complementary to } \mathcal{N}u(P_2A_1)
\end{aligned}
\right\} \quad (3.2.16.a)
$$

In a similar way we find for (3.1.1.b) the solution

$$
\left.
\begin{aligned}
1^{o}\quad & \Delta\hat{\bar{x}}_1^{(\bar{s})} = \bar{S}\{\bar{S}^t(\bar{P}_3\bar{A}_1)^tQ_{\bar{y}}^{-1}(\bar{P}_3\bar{A}_1)\bar{S}\}^{-1}\bar{S}^t(\bar{P}_3\bar{A}_1)^tQ_{\bar{y}}^{-1}\Delta\bar{y} \\
2^{o}\quad & \Delta\hat{\bar{x}}_3^{(\bar{s})} = (\bar{A}_3^tQ_{\bar{y}}^{-1}\bar{A}_3)^{-1}\bar{A}_3^tQ_{\bar{y}}^{-1}(\Delta\bar{y} - \bar{A}_1\Delta\hat{\bar{x}}_1^{(\bar{s})}), \text{ with} \\
& \bar{P}_3 = I - \bar{A}_3(\bar{A}_3^tQ_{\bar{y}}^{-1}\bar{A}_3)^{-1}\bar{A}_3^tQ_{\bar{y}}^{-1}, \text{ and} \\
& \bar{S} = \mathcal{R}(\bar{S}) \text{ complementary to } \mathcal{N}u(\bar{P}_3\bar{A}_1)
\end{aligned}
\right\} \quad (3.2.16.b)
$$

Thus the reduced normals of (3.1.1.a) and (3.1.1.b) pertaining to the common unknowns are respectively $N_1 = (P_2A_1)^tQ_y^{-1}(P_2A_1)$ and $\bar{N}_1 = (\bar{P}_3\bar{A}_1)^tQ_{\bar{y}}^{-1}(\bar{P}_3\bar{A}_1)$. However, in view of our assumptions (3.1.1) we cannot simply add them together yet. What we need is a slight modification of one of the two reduced normals N_1 and $\bar{N}_1$, such that relation (3.1.1.d) is taken care of. That is, if

$$
\mathcal{R}(V_1^{\perp}) = \mathcal{R}((V_1^{\perp})_1) \oplus \mathcal{R}((V_1^{\perp})_2) = \mathcal{R}((V_1^{\perp})_1) \oplus \mathcal{N}u(\bar{P}_3\bar{A}_1),
$$

and

$$
\mathcal{R}(V_1^{\perp}) = \mathcal{R}((V_1^{\perp})_3) \oplus \mathcal{R}((V_1^{\perp})_4) = \mathcal{R}((V_1^{\perp})_3) \oplus \mathcal{N}u(P_2A_1),
$$

we either need to modify N_1 with the aid of $\mathcal{R}((V_1^{\perp})_3)$ to an N_1' with $\mathcal{N}u(N_1') = \mathcal{R}(V_1^{\perp})$, or $\bar{N}_1$ with the aid of $\mathcal{R}((V_1^{\perp})_1)$ to an $\bar{\bar{N}}_1$ with $\mathcal{N}u(\bar{\bar{N}}_1) = \mathcal{R}(V_1^{\perp})$.

For our example of subsection 3.1 such a modification of N_1 would mean that we eliminate the scale- and orientational information of the first network. And likewise, elimination of the orientational information of the second network would correspond to modifying $\bar{N}_1$ to an $\bar{\bar{N}}_1$ with $\mathcal{N}u(\bar{\bar{N}}_1) = \mathcal{R}(V_1^{\perp})$.

Since by assumption the first network contains more information than the second, we will opt for modifying $\bar{N}_1$. For our example this means that we eliminate the orientation of the second network in favour of the astronomical orientation of the first network.

The modified reduced normal $\bar{\bar{N}}_1$ we are looking for will thus be the reduced normal of the relaxed model

$$
\underset{\bar{m}\times 1}{\Delta\tilde{\bar{y}}} = \underset{\bar{m}\times n}{(\bar{A}_1 \vdots} \underset{\bar{m}\times n_3}{\bar{A}_3 \vdots} \underset{\bar{m}\times(r-\bar{q})}{-\bar{A}_1(V_1^{\perp})_1 - \bar{A}_3(V_3^{\perp})_1)} \begin{pmatrix} \Delta x_1 \\ \Delta x_3 \\ \Delta p_1 \end{pmatrix} \underset{=}{\text{call}} \underset{(n+n_3+r-\bar{q})\times 1}{(\bar{A}_1 \vdots \bar{\bar{A}}_3)} \begin{pmatrix} \Delta x_1 \\ \Delta v \end{pmatrix} . \quad (3.2.17)
$$

And since the solution of this relaxed model reads as

$$
\left.
\begin{aligned}
&1^{o}\quad \Delta\hat{x}_1^{(\bar{\bar{s}})} = \bar{\bar{S}}\{\bar{\bar{S}}^t(\bar{\bar{P}}_3\bar{A}_1)^tQ_{\bar{y}}^{-1}(\bar{\bar{P}}_3\bar{A}_1)\bar{\bar{S}}\}^{-1}\bar{\bar{S}}^t(\bar{\bar{P}}_3\bar{A}_1)^tQ_{\bar{y}}^{-1}\Delta\bar{y},\\
&2^{o}\quad \Delta\hat{v}^{(\bar{\bar{s}})} = (\bar{\bar{A}}_3^tQ_{\bar{y}}^{-1}\bar{\bar{A}}_3)^{-1}\bar{\bar{A}}_3^tQ_{\bar{y}}^{-1}(\Delta\bar{y}-A_1\Delta\hat{x}_1^{(\bar{\bar{s}})}),\ \text{with}\\
&\bar{\bar{P}}_3 = I - \bar{\bar{A}}_3(\bar{\bar{A}}_3^tQ_{\bar{y}}^{-1}\bar{\bar{A}}_3)^{-1}\bar{\bar{A}}_3^tQ_{\bar{y}}^{-1},\ \text{and}\\
&\bar{\bar{S}} = R(\bar{\bar{S}})\ \text{complementary to}\ Nu(\bar{\bar{P}}_3\bar{A}_1) = R(V_1^{\perp}) \supset Nu(\bar{P}_3\bar{A}_1)
\end{aligned}
\right\} \qquad (3.2.18)
$$

the reduced normal we are looking for is given by $\bar{N}_1 = (\bar{\bar{P}}_3\bar{A}_1)^tQ_{\bar{y}}^{-1}(\bar{\bar{P}}_3\bar{A}_1)$. Note that since (3.2.18) is merely obtained from relaxing (3.1.1b) to (3.2.17) the two solutions (3.2.16.b) and (3.2.18) will be related by an appropriate S-transformation. We have for instance

$$\Delta\hat{x}_1^{(\bar{\bar{s}})} = P_{\bar{\bar{S}},\,Nu(\bar{\bar{P}}_3\bar{A}_1)}\,\Delta\hat{\hat{x}}_1^{(\bar{s})}.$$

Now that we have the appropriate reduced normals N_1 and $\bar{N}_1$ we can proceed with the Helmert-blocking method and add the two reduced normals to solve for the common unknowns. The remaining unknowns are then found through back-substitution.
All in all the final solutions reads as:

$$\Delta\hat{\hat{x}}_1^{(s)} = S\{S^t((P_2A_1)^tQ_y^{-1}(P_2A_1)+(\bar{\bar{P}}_3\bar{A}_1)^tQ_{\bar{y}}^{-1}(\bar{\bar{P}}_3\bar{A}_1))S\}^{-1}S^t((P_2A_2)^tQ_y^{-1}\Delta y+(\bar{\bar{P}}_3\bar{A}_1)^tQ_{\bar{y}}^{-1}\Delta\bar{y}),$$

$$\Delta\hat{\hat{x}}_2^{(s)} = (A_2^tQ_y^{-1}A_2)^{-1}A_2^tQ_y^{-1}(\Delta y - A_1\Delta\hat{\hat{x}}_1^{(s)}),$$

$$\Delta\hat{\hat{v}}^{(s)} = (\bar{\bar{A}}_3^tQ_{\bar{y}}^{-1}\bar{\bar{A}}_3)^{-1}\bar{\bar{A}}_3^tQ_{\bar{y}}^{-1}(\Delta\bar{y} - \bar{A}_1\Delta\hat{\hat{x}}_1^{(s)}),\ \text{which can be decomposed by means of}$$

$$\Delta v = (\Delta x_2^t\Delta p_1^t)^t \ \text{and}\ \bar{\bar{A}}_3 = (\bar{A}_3 \vdots - \bar{A}_1(V_1^{\perp})_1 - \bar{A}_3(V_3^{\perp})_1)\ \text{into}$$

$$\Delta\hat{\hat{p}}_1 = -\{(V_1^{\perp})_1^t(\bar{\bar{P}}_3\bar{A}_1)^tQ_{\bar{y}}^{-1}(\bar{\bar{P}}_3\bar{A}_1)(V_1^{\perp})_1\}^{-1}(V_1^{\perp})_1^t(\bar{\bar{P}}_3\bar{A}_1)^tQ_{\bar{y}}^{-1}(\Delta\bar{y} - \bar{A}_1\Delta\hat{\hat{x}}_1^{(s)}),$$

$$\Delta\hat{\hat{x}}_3^{(s)} = (\bar{A}_3^tQ_{\bar{y}}^{-1}\bar{A}_3)^{-1}\bar{A}_3^tQ_{\bar{y}}^{-1}\{I - (\bar{A}_1(V_1^{\perp})_1+\bar{A}_3(V_3^{\perp})_1)\cdot$$
$$\cdot\{(V_1^{\perp})_1^t(\bar{\bar{P}}_3\bar{A}_1)^tQ_{\bar{y}}^{-1}(\bar{\bar{P}}_3\bar{A}_1)(V_1^{\perp})_1\}^{-1}(V_1^{\perp})_1^t(\bar{\bar{P}}_3\bar{A}_1)^tQ_{\bar{y}}^{-1}\}(\Delta\bar{y} - \bar{A}_1\Delta\hat{\hat{x}}_1^{(s)})$$

$$= (\bar{A}_3Q_{\bar{y}}^{-1}\bar{A}_3)^{-1}\bar{A}_3^tQ_{\bar{y}}^{-1}\{\Delta\bar{y} - \bar{A}_1(\Delta\hat{\hat{x}}_1^{(s)}-(V_1^{\perp})_1\Delta\hat{\hat{p}}_1)\}+(V_3^{\perp})_1\Delta\hat{\hat{p}}_1,$$

with $S = R(S)$ complementary to $Nu(P_2A_1)$. (3.2.19)

Thus if we take the customary abbreviations

$$N_1 = (P_2A_1)^t Q_y^{-1}(P_2A_1),\ \bar{N}_1 = (\bar{P}_3\bar{A}_1)^t \bar{Q}_y^{-1}(\bar{P}_3\bar{A}_1),\ \bar{\bar{N}}_1 = (\bar{\bar{P}}_3\bar{A}_1)^t \bar{\bar{Q}}_y^{-1}(\bar{\bar{P}}_3\bar{A}_1),$$

and (3.2.20)

$$\Delta n_1 = (P_2A_1)^t Q_y^{-1}\Delta y,\quad \Delta \bar{n}_1 = (\bar{P}_3\bar{A}_1)^t \bar{Q}_y^{-1}\Delta \bar{y},\quad \Delta \bar{\bar{n}}_1 = (\bar{\bar{P}}_3\bar{A}_1)^t \bar{\bar{Q}}_y^{-1}\Delta \bar{\bar{y}}$$

we can summarize the general procedure of method III as:

a) Reduce the normal systems of the two original models (3.1.1.a) and (3.1.1.b) to the common unknowns: $N_1\Delta x_1 = \Delta n_1$ and $\bar{N}_1\Delta \bar{x}_1 = \Delta \bar{n}_1$.

b) Relax the reduced normal system of the second network with the aid of $(V_1^{\perp})_1$:

$$\begin{bmatrix} \bar{N}_1 & -\bar{N}_1(V_1^{\perp})_1 \\ -(V_1^{\perp})_1^t\bar{N}_1 & (V_1^{\perp})_1^t\bar{N}_1(V_1^{\perp})_1 \end{bmatrix}\begin{bmatrix} \Delta x_1 \\ \Delta p_1 \end{bmatrix} = \begin{bmatrix} \Delta \bar{n}_1 \\ -(V_1^{\perp})^t\Delta \bar{n}_1 \end{bmatrix}.$$

c) Add the reduced normal system of the first network:

$$\begin{bmatrix} N_1+\bar{N}_1 & -\bar{N}_1(V_1^{\perp})_1 \\ -(V_1^{\perp})_1^t\bar{N}_1 & (V_1^{\perp})_1^t\bar{N}_1(V_1^{\perp})_1 \end{bmatrix}\begin{bmatrix} \Delta x_1 \\ \Delta p_1 \end{bmatrix} = \begin{bmatrix} \Delta n_1+\Delta \bar{n}_1 \\ -(V_1^{\perp})_1^t\Delta \bar{n}_1 \end{bmatrix}.$$

d) By means of further reduction one gets

$$\begin{bmatrix} N_1+\bar{\bar{N}}_1 & 0 \\ -(V_1^{\perp})_1^t\bar{N}_1 & (V_1^{\perp})_1^t\bar{N}_1(V_1^{\perp})_1 \end{bmatrix}\begin{bmatrix} \Delta x_1 \\ \Delta p_1 \end{bmatrix} = \begin{bmatrix} \Delta n_1+\Delta \bar{\bar{n}}_1 \\ -(V_1^{\perp})_1^t\Delta \bar{n}_1 \end{bmatrix},$$

with the solution:

$$\Delta \hat{\hat{x}}_1^{(s)} = S\{S^t(N_1+\bar{\bar{N}}_1)S\}^{-1}S^t(\Delta n_1 + \Delta \bar{\bar{n}}_1)$$

$$\Delta \hat{\hat{p}}_1 = -\{(V_1^{\perp})_1^t\bar{N}_1(V_1^{\perp})_1\}^{-1}(V_1^{\perp})_1^t(\Delta \bar{n}_1 - \bar{N}_1\Delta \hat{\hat{x}}_1^{(s)}).$$

e) The remaining unknowns are found through:

$$\Delta \hat{\hat{x}}_2^{(s)} = (A_2^tQ_y^{-1}A_2)^{-1}A_2^tQ_y^{-1}(\Delta y - A_1\Delta \hat{\hat{x}}_1^{(s)})$$

$$\Delta \hat{\hat{x}}_3^{(s)} = (\bar{A}_3^t\bar{Q}_y^{-1}\bar{A}_3)^{-1}\bar{A}_3^t\bar{Q}_y^{-1}[\Delta \bar{y} - \bar{A}_1(\Delta \hat{\hat{x}}_1^{(s)} - (V_1^{\perp})_1\Delta \hat{\hat{p}}_1)] + (V_3^{\perp})_1\Delta \hat{\hat{p}}_1.$$

(3.2.21)

In the above approach to the Helmert blocking procedure we have seen that, as a consequence of our general assumptions (3.1.1), the reduced normals N_1 and $\bar{N}_1$ are singular. Hence, in general one can not start from the principle that both reduced normals are regular, unless 1^o there are no degrees of freedom involved, which is highly unlikely, or 2^o one assumes that the degrees of freedom are already been taken care of **before** applying the Helmert blocking procedure. The reduced normals will namely be regular if for instance the *S*-systems of both networks are defined a priori in their non-overlapping parts.

The question that remains to be answered is then, whether one can still apply the procedure as outlined in (3.2.21). With some slight modifications we will see that the answer is in the affirmative. The important difference with (3.2.31) is however that we shall need additional transformation parameters to take care of the a priori *S*-system definition.

Let us start with the two solutions one gets when the *S*-systems are defined in the two non-overlapping parts of the two networks.

For the first network one would get instead of (3.2.16.a), the solution

$$
\begin{aligned}
1^o \quad & \Delta\hat{x}_1^{(s_2)} = \left((P_2''A_1)^t Q_y^{-1}(P_2''A_1)\right)^{-1}(P_2''A_1)^t Q_y^{-1}\Delta y \ , \\
2^o \quad & \Delta\hat{x}_2^{(s_2)} = S_2\left(S_2^t A_2^t Q_y^{-1} A_2 S_2\right)^{-1} S_2^t A_2^t Q_y^{-1}(\Delta y - A_1\Delta\hat{x}_1^{(s_2)}), \text{ with} \qquad (3.2.22.a) \\
& P_2'' = I - A_2 S_2 (S_2^t A_2^t Q_y^{-1} A_2 S_2)^{-1} S_2^t A_2^t Q_y^{-1}, \text{ and} \\
& S_2 = R(S_2) \text{ complementary to } Nu(P_1 A_2) = R((I - A_1(A_1^t Q_y^{-1} A_1)^{-1} A_1^t Q_y^{-1})A_2).
\end{aligned}
$$

and for the second network,

$$
\begin{aligned}
1^o \quad & \Delta\hat{\bar{x}}_1^{(\bar{s}_3)} = \left((\bar{P}_3''\bar{A}_1)^t \bar{Q}_y^{-1}(\bar{P}_3''\bar{A}_1)\right)^{-1}(\bar{P}_3''\bar{A}_1)^t \bar{Q}_y^{-1}\Delta\bar{y}, \\
2^o \quad & \Delta\hat{\bar{x}}_3^{(\bar{s}_3)} = \bar{S}_3\left(\bar{S}_3^t \bar{A}_3^t \bar{Q}_y^{-1} \bar{A}_3 \bar{S}_3\right)^{-1} \bar{S}_3^t \bar{A}_3^t \bar{Q}_y^{-1}(\Delta\bar{y} - \bar{A}_1\Delta\hat{\bar{x}}_1^{(\bar{s}_3)}), \text{ with} \qquad (3.2.22.b) \\
& \bar{P}_3'' = I - \bar{A}_3\bar{S}_3(\bar{S}_3^t \bar{A}_3^t \bar{Q}_y^{-1} \bar{A}_3 \bar{S}_3)^{-1} \bar{S}_3^t \bar{A}_3^t \bar{Q}_y^{-1}, \text{ and} \\
& \bar{S}_3 = R(\bar{S}_3) \text{ complementary to } Nu(\bar{P}_1\bar{A}_3) = R((I - \bar{A}_1(\bar{A}_1^t \bar{Q}_y^{-1}\bar{A}_1)^{-1}\bar{A}_1^t \bar{Q}_y^{-1})\bar{A}_3).
\end{aligned}
$$

These two solutions are easily verified by transforming with the appropriate *S*-transformations the two solutions (3.2.16.a) and (3.2.16.b).

For the Helmert blocking procedure we have in the above case the disposal of the reduced normal systems

$$
N_1''\Delta x_1 = \Delta n_1'' \quad \text{and} \qquad \bar{N}_1''\Delta\bar{x}_1 = \Delta\bar{n}_1'', \qquad (3.2.23.a)
$$

with

$$
N_1'' = (P_2''A_1)^t Q_y^{-1}(P_2''A_1), \quad \bar{N}_1'' = (\bar{P}_3''\bar{A}_1)^t \bar{Q}_y^{-1}(\bar{P}_3''\bar{A}_1),
$$

and

$$
\Delta n_1'' = (P_2''A_1)^t Q_y^{-1}\Delta y \quad , \quad \Delta\bar{n}_1'' = (\bar{P}_3''\bar{A}_1)^t \bar{Q}_y^{-1}\Delta\bar{y}.
$$

But as before we cannot simply add the two reduced normals N_1'' and $\bar{N}_1''$ to solve for the common unknowns. What we need is a modification of $\bar{N}_1''$. First to get rid of the a priori $\bar{S}_3$- system definition. This will give us $\bar{N}_1$ back. And secondly to incorporate $(V_1^{\perp})_1$ to get $\bar{\bar{N}}_1$. Therefore instead of (3.2.21.b), the relaxed normal system needed, reads

$$\begin{pmatrix} \bar{N}_1'' & -\bar{N}_1''(V_1^{\perp})_1 & -\bar{N}_1''(V_1^{\perp})_2 \\ -(V_1^{\perp})_1^t\bar{N}_1'' & (V_1^{\perp})_1^t\bar{N}_1''(V_1^{\perp})_1 & (V_1^{\perp})_1^t\bar{N}_1''(V_1^{\perp})_2 \\ -(V_1^{\perp})_2^t\bar{N}_1'' & (V_1^{\perp})_2^t\bar{N}_1''(V_1^{\perp})_1 & (V_1^{\perp})_2^t\bar{N}_1''(V_1^{\perp})_2 \end{pmatrix} \begin{pmatrix} \Delta x_1 \\ \Delta p_1 \\ \Delta p_2 \end{pmatrix} = \begin{pmatrix} \Delta\bar{n}_1'' \\ -(V_1^{\perp})_1^t\Delta\bar{n}_1'' \\ -(V_1^{\perp})_2^t\Delta\bar{n}_1'' \end{pmatrix},$$

or

$$\begin{pmatrix} \bar{N}_1'' & -\bar{N}_1''(V_1^{\perp}) \\ -(V_1^{\perp})^t\bar{N}_1'' & (V_1^{\perp})^t\bar{N}_1''(V_1^{\perp}) \end{pmatrix} \begin{pmatrix} \Delta x_1 \\ \Delta p \end{pmatrix} = \begin{pmatrix} \Delta\bar{n}_1'' \\ -(V_1^{\perp})^t\Delta\bar{n}_1'' \end{pmatrix}. \qquad (3.2.23.b)$$

Note that, contrary to (3.2.21), additional transformation parameters Δp_2 are needed to take care of the a priori $\bar{S}_3$- system definition.
By adding $N_1''\Delta x_1 = \Delta n_1''$ to (3.2.23.b) we get

$$\begin{pmatrix} N_1'' + \bar{N}_1'' & -\bar{N}_1''(V_1^{\perp}) \\ -(V_1^{\perp})^t\bar{N}_1'' & (V_1^{\perp})^t\bar{N}_1''(V_1^{\perp}) \end{pmatrix} \begin{pmatrix} \Delta x_1 \\ \Delta p \end{pmatrix} = \begin{pmatrix} \Delta n_1'' + \Delta\bar{n}_1'' \\ -(V_1^{\perp})^t\Delta\bar{n}_1'' \end{pmatrix}, \qquad (3.2.23.c)$$

and after some reduction steps we obtain

$$\begin{pmatrix} N_1'' + \bar{\bar{N}}_1 & 0 & 0 \\ -(V_1^{\perp})^t\bar{N}_1 & (V_1^{\perp})_1^t\bar{N}_1(V_1^{\perp})_1 & 0 \\ -(V_1^{\perp})_2^t\bar{N}_1'' & (V_1^{\perp})_2^t\bar{N}_1''(V_1^{\perp})_1 & (V_1^{\perp})_2^t\bar{N}_1''(V_1^{\perp})_2 \end{pmatrix} \begin{pmatrix} \Delta x_1 \\ \Delta p_1 \\ \Delta p_2 \end{pmatrix} = \begin{pmatrix} \Delta n_1'' + \Delta\bar{\bar{n}}_1 \\ -(V_1^{\perp})_1^t\Delta\bar{n}_1 \\ -(V_1^{\perp})_2^t\Delta\bar{n}_1'' \end{pmatrix}. \qquad (3.2.23.d)$$

In a similar way as in (3.2.21) we then find the final solution as

$$\Delta\hat{\hat{x}}_1^{(s_2)} = (N_1'' + \bar{\bar{N}}_1)^{-1}(\Delta n_1'' + \Delta\bar{\bar{n}}_1)$$

$$= \left((P_2''A_1)^tQ_y^{-1}(P_2''A_1) + (\bar{\bar{P}}_3\bar{A}_1)^tQ_{\bar{y}}^{-1}(\bar{\bar{P}}_3\bar{A}_1)\right)^{-1}\left((P_2''A_1)^tQ_y^{-1}\Delta y + (\bar{\bar{P}}_3\bar{A}_1)^tQ_{\bar{y}}^{-1}\Delta\bar{y}\right),$$

$$\Delta\hat{p}_1 = -\left((V_1^{\perp})_1^t\bar{N}_1(V_1^{\perp})_1\right)^{-1}(V_1^{\perp})_1^t\left(\Delta\bar{n}_1 - \bar{N}_1\Delta\hat{\hat{x}}_1^{(s_2)}\right)$$

$$= -\left((V_1^{\perp})_1^t(\bar{P}_3\bar{A}_1)^tQ_{\bar{y}}^{-1}(\bar{P}_3\bar{A}_1)(V_1^{\perp})_1\right)^{-1}(V_1^{\perp})_1^t(\bar{P}_3\bar{A}_1)^tQ_{\bar{y}}^{-1}\left(\Delta\bar{y} - \bar{A}_1\Delta\hat{\hat{x}}_1^{(s_2)}\right),$$

$$\Delta\hat{\bar{p}}_2 = -\left((V_1^{\perp})_2^t \tilde{\bar{N}}_1 (V_1^{\perp})_2\right)^{-1} (V_1^{\perp})_2^t \left(\Delta\tilde{\bar{n}}_1 - \tilde{\bar{N}}_1 \Delta\hat{\hat{x}}_1^{(s_2)} + \tilde{\bar{N}}_1 (V_1^{\perp})_1 \Delta\hat{\bar{p}}_1\right)$$

$$= -\left((V_1^{\perp})_2^t (\tilde{\bar{P}}_3 \bar{A}_1)^t Q_y^{-1} (\tilde{\bar{P}}_3 \bar{A}_1)(V_1^{\perp})_2\right)^{-1} (V_1^{\perp})_2^t (\tilde{\bar{P}}_3 \bar{A}_1)^t Q_y^{-1} \left(\Delta\bar{y} - \bar{A}_1 (\Delta\hat{\hat{x}}_1^{(s_2)} - (V_1^{\perp})_1 \Delta\hat{\bar{p}}_1)\right),$$

$$\Delta\hat{\hat{x}}_2^{(s_2)} = S_2 \left(S_2^t A_2^t Q_y^{-1} A_2 S_2\right)^{-1} S_2^t A_2^t Q_y^{-1} \left(\Delta y - A_1 \Delta\hat{\hat{x}}_1^{(s_2)}\right),$$

$$\Delta\hat{\hat{x}}_3^{(s_2)} = \bar{S}_3 \left(\bar{S}_3^t \bar{A}_3^t Q_y^{-1} \bar{A}_3 \bar{S}_3\right)^{-1} \bar{S}_3^t \bar{A}_3^t Q_y^{-1} \left(\Delta\bar{y} - \bar{A}_1 \left(\Delta\hat{\hat{x}}_1^{(s_2)} - V_1^{\perp} \Delta\hat{\bar{p}}\right)\right) + V_3^{\perp} \Delta\hat{\bar{p}} \ .$$

(3.2.23.e)

We thus see that also in case the S-systems of both networks are defined a priori, one can apply the procedure as outlined in (3.2.21). The important difference is however, that in the above case additional transformation parameters Δp_2 are needed which, contrary to Δp_1, will **not** be invariant to the choice of S-systems. This emphasizes once more our earlier remark about the interpretability of the transformation parameters.

Note that solution (3.2.23.e) is essentially the same as solution (3.2.19) or (3.2.21). One can verify this by showing that $(\Delta\hat{\hat{x}}_1^{(s_2)}, \Delta\hat{\hat{x}}_2^{(s_2)}, \Delta\hat{\hat{x}}_3^{(s_2)})$ transforms with an appropriate S-transformation to $(\Delta\hat{\hat{x}}_1^{(s)}, \Delta\hat{\hat{x}}_2^{(s)}, \Delta\hat{\hat{x}}_3^{(s)})$.

In this section we have seen how the three customary methods for connecting geodetic networks generalize if one starts from the general assumptions (3.1.1).

As to the first two methods, it is interesting to remark that in the geodetic literature one usually assumes either one of the following two attitudes when discussing the problem of connecting geodetic networks: Either one places the whole discussion in the context of free networks, thereby suggesting that free networks are really something special and that they should not be confused, let alone be compared with "ordinary" networks. Or, one assumes the attitude that the coordinates of the two overlapping networks merely differ by a similarity transformation, which is easily taken care of by estimating the transformation parameters in a least-squares sense. Both attitudes are however needlessly too restrictive. Although in the first approach one is normally very careful in stating what type of networks are involved, one usually starts from the too restrictive assumptions that $\mathcal{N}u(P_2 A_1) = \mathcal{N}u(\bar{P}_3 \bar{A}_1) = \mathcal{R}(V_1^{\perp})$. In the second case, however, one often neglects to state the basic starting assumptions. It is namely not enough to say that the two coordinate sets differ by a similarity transformation. Important is, to know what type of networks are involved. Only then will one be able to identify which of the transformation parameters are estimable.

When reviewing the relevant geodetic literature, it is also interesting to note that those who assume the above mentioned first attitude usually end up with the method of condition equations as solution strategy, whereas those who assume the second attitude usually find themselves formulating the problem in such a form that first the transformation parameters need estimation. But both methods are of course equally applicable in principle. In fact, the aversion which is generally felt towards the method of condition equations, does not apply in case of connecting networks, since one can argue

that method I is more tractable computationwise than method II. In some cases at least.
As to the third method, we showed how one should go about when the S-systems are defined either before or after the merging of the two reduced normals. Here also the fact that in general not all transformation parameters can be treated on an equal footing, became apparant.
Some authors have proposed in the context of method III to give weights to some of the transformation parameters. They argue that in case of for instance two networks which both are known to contain orientational information, this is a way of deciding how much of the orientational information of both networks is carried over to the final solution. This in itself is true of course, but we do not think that in general this is an advisable way to go about, since it has an element of arbitrariness in it. So far namely, no objective criterium has been proposed on the basis of which to decide to follow such a procedure. It seems therefore more advisable to decide on the basis of statistical tests whether or not the two networks significantly differ in their orientation.
As a final remark we mention that in this chapter we have adopted the customary assumption that the coordinate systems in which the two networks are described differ only differentially. If this is not the case then one has to recourse to either a preliminary transformation which make the two networks coincide approximately or to an iteration. In the next chapter we will see that in some cases one can do without an iteration and formulate an **exact non-linear solution.**

IV. GEOMETRY OF NON-LINEAR ADJUSTMENT

1. General problem statement

In the previous chapters we were primarily concerned with the **linear** model

$$\tilde{\mathbf{y}} \in \bar{N} \subset M, \qquad \bar{N} = \mathbf{A}N \ . \tag{1.1}$$

As a general solution of the linear unbiased estimation problem we found that the actual adjustment problem was solved by

$$1^{o}. \qquad \hat{\mathbf{y}}^{(c)} = \mathbf{P}_{R(\mathbf{A}),C^{o}}\, \mathbf{y}_{s}, \qquad M = R(\mathbf{A}) \oplus C^{o},$$

and the actual inverse linear mapping problem by

$$2^{o}. \qquad \hat{\mathbf{x}}^{(s,c)} = \mathbf{P}_{S,Nu(\mathbf{A})}\, \mathbf{B}\hat{\mathbf{y}}^{(c)}, \qquad N = S \oplus Nu(\mathbf{A}),$$

where $\mathbf{B}: M \to N$ is allowed to be any arbitrary inverse of the linear map $\mathbf{A}: N \to M$.
In this chapter we take up the study of **non-linear adjustment.** A problem which heretofore has almost been avoided in the geodetic literature. To this end we replace the linear map $\mathbf{A}$ by a non-linear map $\mathbf{y}: N \to M$. Instead of the linear model (1.1) we then have the **non-linear** model

$$\tilde{\mathbf{y}} \in \bar{N} \subset M, \qquad \bar{N} = \mathbf{y}(N) \ . \tag{1.2}$$

It seems natural now to extend our results of the linear theory to the companion problem of non-linear operators. But unfortunately one can very seldom extend the elegant formulations and solution techniques from linear to non-linear situations.
In correspondence with the linear theory the problem of non-linear adjustment can roughly be divided into (a) the problem of finding the estimates $\hat{\mathbf{y}}$ and $\hat{\mathbf{x}}$, and (b) the problem of finding the statistical properties of the estimators involved. In order to keep our non-linear adjustment problem surmountable we will restrict ourselves to least-squares estimation and also we assume for the moment that map $\mathbf{y}$ is injective. Our non-linear least-squares adjustment problem reads then

$$\min_{\mathbf{x}\in N}.2\, E(\mathbf{x}) = \min_{\mathbf{y}\in\bar{N}=\mathbf{y}(N)}. \langle \mathbf{y}_{s}-\mathbf{y}, \mathbf{y}_{s}-\mathbf{y} \rangle_{M} = \langle \mathbf{y}_{s}-\hat{\mathbf{y}}, \mathbf{y}_{s}-\hat{\mathbf{y}} \rangle_{M}, \tag{1.3}$$

(the factor 2 is merely inserted for convenience).
In order to solve for $\hat{\mathbf{y}}$ and $\hat{\mathbf{x}}$ we need non-linear maps $\mathbf{P}: M \to \bar{N}$ and $\mathbf{y}^{-1}: M \to N$ such that

$$1^{o} \quad \hat{y} = P(y_s),$$

and

$$2^{o} \quad \hat{x} = y^{-1}(\hat{y}), \quad \text{with} \quad y^{-1} \circ y = \text{identity}.$$

Due, however, to the non-linearity of map y it is very seldom that one can find closed expressions for the maps P and y^{-1} (there are exceptions!). In practice one will therefore have to recourse to methods which are iterative in nature. One starts with a given point $x_o \in N$, the initial guess, and proceeds to generate a sequence $x_o, x_1, x_2 \ldots$ which hopefully converges to the point $\hat{x}$. Most methods which are discussed in the literature (see e.g. Ortega and Rheinboldt, 1970) adhere to the following scheme:

$$x^{\beta}_{q+1} = x^{\beta}_{q} + t_q \Delta x^{\beta}_{q}, \qquad \beta = 1, \ldots, n; \quad \text{no summation over q,} \tag{1.4}$$

(i) Set q=0. An initial guess is provided externally,

(ii) Determine an increment vector Δx_q in the direction of the proposed step,

(iii) Determine a scalar t_q such that $||y_s - y(x_{q+1})||_M \leq ||y_s - y(x_q)||_M$, i.e., such that the qth step may considered to be an improvement over the (q-1)th step. The way in which t_q is chosen is known as a line search strategy.

(iv) Test whether the termination criterion is met. If so, accept x_{q+1} as the value of $\hat{x}$. If not, increase q by one and return to (ii).

Generally one can say that the individual methods falling under (1.4) differ in their choice of the increment vector Δx_q and the scalar t_q. The iterative techniques fall roughly into two classes: direct search methods and gradient methods. Direct search methods are those which do not require the explicit evaluation of any partial derivatives of the function E, but instead rely solely on values of the objective function E, plus information gained from the earlier iterations. Gradient methods on the other hand are those which select the direction Δx_q using values of the partial derivatives of the objective function E with respect to the independent variables, as well as values of E itself, together with information gained from earlier iterations. The required derivatives, which for some methods are of order higher than the first, can be obtained either analytically or numerically using some finite difference scheme. This latter approach necessitates extra function evaluations close to the current point x_q, and effectively reduces a gradient method to one of direct search.
We will not attempt to give an exhaustive list of iteration methods which could possibly solve our adjustment problem (1.3). For a comprehensive survey of the various methods we refer the reader to the encyclopaedic work of (Ortega and Rheinboldt, 1970). Instead, we restrict ourselves to that gradient method which seems to be preeminently suited for our least-squares adjustment problem, namely **Gauss' iteration method.** This method can be considered as the natural generalization of the linear case and it is the only method which fully exploits the sum of squares structure of the objective function E.

As to the second problem, namely that of finding the statistical properties of the estimators involved we will not present a complete treatment of the statistical theory dealing with non-linear adjustment. We cannot expect a well working theory for the non-linear model as we know it for the

linear one. The probability distribution of the non-linear estimator for $\tilde{\mathbf{y}}$ for instance, depends on both the non-linear map **P** and on the distribution of the data. Hence, it depends on the "true" values of **x** which are generally unknown. Therefore, even when we can derive a precise formula for the distribution of the estimator, we can evaluate in general only the approximation obtained by substituting the estimated parameter values for the "true" ones.

The plan for this chapter is the following:

As said we will discuss Gauss' iteration method in some detail. We have chosen to make use of differential geometry as a tool for studying Gauss' method. We strongly believe namely that geometry and in particular differential geometry provides us with a better and richer understanding of the complicated problem of non-linear adjustment. Many of the geometric concepts developed in differential geometry turn out to be important indicators, qualitatively as well as quantitatively, of how non-linearity manifests itself in the local behaviour of Gauss' method and in the statistical properties of the estimators. We therefore commence in section 2 with a brief introduction into Riemannian geometry.

In section 3 we consider the problem of univariate non-linear least-squares. That is, we consider the problem of orthogonal projection onto a parametrized space curve. For this purpose we first study the local geometry of a space curve with the aid of the so-called Frenet frame and Frenet formulae. The geometrical impact of the Frenet formulae is that if **T** and **N** are respectively the unit tangent vector and unit normal to a plane curve and s its arclength parameter, than to an accuracy of the order of the second power of small quantities Δs, we have

$$\begin{aligned} \mathbf{T} + \Delta\mathbf{T} &= \cos(k\Delta s)\,\mathbf{T} + \sin(k\Delta s)\,\mathbf{N} \\ \mathbf{N} + \Delta\mathbf{N} &= -\sin(k\Delta s)\,\mathbf{T} + \cos(k\Delta s)\,\mathbf{N}, \end{aligned}$$

i.e., the Frenet formulae embody the fact that the Frenet frame $(\mathbf{T},\mathbf{N})$ undergoes a rotation depending on the curvature k of the plane curve as one moves from the point on the curve corresponding to s to the nearby point corresponding to $s + \Delta s$. It is this observation on which most of our further developments are based.
After having studied the local geometry of a space curve, we show how curvature affects the local behaviour of Gauss' method. The section is closed with some examples and preliminary conclusions.

In section 4 we consider the case of multivariate non-linear least-squares adjustment. That is, we consider the problem of orthogonal projection onto a parametrized submanifold. In order to generalize the results of section 3 we have to find an appropriate generalization to the Frenet formulae. This we find in the so-called Gauss' equation. With the aid of the normal field **B**, which can be considered as the multivariate generalization of the second fundamental tensor **b** known from classical surface theory (see e.g. Stoker, 1969), we then show how the extrinsic curvatures of the submanifold affect the local behaviour of the multivariate Gauss' iteration method. At the end of subsection 4.4 we summarize the more important conclusions. The section is ended with a subsection

in which we show how Gauss' method can be made into a globally convergent iteration method.

In section 5 we start by considering the classical two dimensional Helmert transformation as a typical example of a totally geodesic submanifold, i.e. a manifold for which all extrinsic curvatures are identically zero. Next we show that for a particular class of manifolds, namely **ruled** surfaces, important simplifications of the non-linear least-squares adjustment problem can be obtained through dimensional reduction. Based on this idea we then present a non-linear generalization of the classical two dimensional Helmert transformation, which we call the two dimensional **Symmetric Helmert transformation.** We also give the solution of the two dimensional Symmetric Helmert transformation when a non-trivial rotational invariant covariance structure is pre-supposed. After this we generalize our results to three dimensions. Finally we give some suggestions as to how to estimate the extrinsic curvatures in practice and we estimate the curvature of some simple 2-dimensional geodetic networks.

In the last but one section we briefly discuss some of the consequences of non-linearity for the statistical treatment of an adjustment. We also show how the first moments of the estimators are affected by curvature.

2. A brief introduction into Riemannian geometry

We cannot expect to convey here much of the theory of Riemannian geometry. For a comprehensive treatment of the theory we refer the reader to the relevant mathematical literature (see e.g. Spivak, 1975).
Riemannian geometry is a generalization of metric differential geometry of surfaces. Instead of surfaces one considers n-dimensional Riemannian manifolds. These are obtained from differential manifolds by introducing a Riemannian metric, that is, a metric defined by a quadratic differential form whose coefficients are the components of a two times covariant positive definite symmetric tensor field. The corresponding geometry is called Riemannian geometry.
Surfaces, with their usual metric inherited or induced from the ambient 3-dimensional Euclidean space, are 2-dimensional Riemannian manifolds, and part of our considerations will be a generalization of ideas from the theory of surfaces and curves. However, for n=1 or 2 there are many simplifications that have no counterpart when $n > 2$. Consequently, a number of new facts and concepts will have to be introduced in the following sections.

In this section we only present briefly some of the basic notions of Riemannian geometry. We first consider **manifolds.** An n-dimensional differentiable or smooth manifold can roughly be described as a set of points tied together continuously and differentiably, so that the points in any sufficiently small region can be put into a one-to-one correspondence with an open set of points in $\mathbb{R}^n$. That correspondence furnishes then a coordinate system for the neighbourhood. Moreover the passage from one coordinate system to another is assumed to be smooth in the overlapping region.
The manifold concepts generalizes and includes the special cases of the real line, plane, linear vector

space and surfaces which are studied in the classical theory. The mathematician (see e.g. Hirsch, 1976) usually begins his development of differential topology by introducing some primitive concepts, such as sets and topology of sets, then builds an elaborate framework out of them and uses that framework to define the concept of a differential manifold. For our present application, however, we can ignore most of the topological aspects. They are either very natural, such as continuity and connectedness or highly technical. Moreover, our analysis in subsequent sections will mainly be of a local nature, i.e. differential geometry in the small. For differential geometry in the small one can do without the global considerations in most cases since one assumes that a single coordinate system without singularities covers the portion of the manifold studied.

We have chosen to define manifolds as subsets of some big, ambient space $\mathbb{R}^k$. This has the advantage that manifolds appear as objects already familiar to those who studied the classical theory of surfaces and it also enables us to surpass many of the topological concepts. Suppose that N is a subset of some big, ambient space $\mathbb{R}^k$. Then N is an n-dimensional manifold if it is locally diffeomorphic to $\mathbb{R}^n$; this means that each point $\mathbf{x}$ of N possesses a neighbourhood $V = V' \cap N$, for some open set V' of $\mathbb{R}^k$, which is diffeomorphic to an open set U of $\mathbb{R}^n$. The two sets $U \subset \mathbb{R}^n$ and $V \subset N$ are said to be diffeomorphic if there exists a map $\mathbf{h}: U \to V$ which is one-to-one, onto and smooth in both directions. This diffeomorphism is called a parametrization of the neighbourhood V. Its inverse $\mathbf{h}^{-1}: V \to U$ is called a coordinate system on V. When the map $\mathbf{h}^{-1}$ is written in coordinates, $\mathbf{h}^{-1} = (x^1, \ldots, x^n)$, the n functions x^α, $\alpha = 1, \ldots, n$, are called coordinate functions.

As a simple geodetic example of a manifold, let N be the set of all planar geodetic networks having, say, $\frac{1}{2}n$ number of points. Each planar geodetic network represents then a point $\mathbf{x}$ of N. The most obvious way to give N a manifold structure is then by taking the diffeomorphism $\mathbf{h}^{-1}: N \to \mathbb{R}^n$ as the identity map. The coordinate functions are then the standard cartesian coordinates. However, one could of course also take polar coordinates, cylindrical coordinates, spherical coordinates or any of the other customary curvilinear coordinates provided they are suitable restricted so as to be one-to-one and have non-zero Jacobian determinant.

If two sets O and N both are manifolds and $O \subset N$, then O is said to be a submanifold of N. In particular, any open set of N is a submanifold of N. Assume for instance that O is the set of all planar geodetic networks having $\frac{1}{2}n$ number of points, with the additional restrictions that, say, some distances between some network points are taken to be constant. Then O can be shown to be a submanifold of the above defined N.

Let us consider the linear approximation of a manifold N, i.e. its **tangent space.** The vectors in it are the tangent vectors to N. Let $\mathbf{c}$ be a point on the manifold N and let $\mathbf{c}$ trace out a curve $\mathbf{c}(t)$. In local coordinates the curve is given by $c^\alpha(t) = x^\alpha(\mathbf{c}(t))$, $\alpha = 1, \ldots, n$. The velocity vector to this curve is given by dc^α/dt. It is now established practice in differential geometry to generalize the classical definition of tangent vector, and to consider a differential operator as tangent vector. To do this we take a real-valued function $E(\mathbf{x})$ defined on N and consider its rate of change along the curve $\mathbf{c}(t)$. The rate of change of $E(\mathbf{x})$ in the direction of $\mathbf{c}(t)$ is dE/dt. In local coordinates this becomes $\partial_\alpha E \, dc^\alpha/dt$ (here we have abbreviated $\partial E/\partial x^\alpha$ by $\partial_\alpha E$). In other words dE/dt is obtained by applying the differential operator $\mathbf{T} = dc^\alpha/dt \, \boldsymbol{\partial}_\alpha$ to E. It is $\mathbf{T}$ which we now define as the tangent to N at $\mathbf{c}$ in the directions given by the curve $\mathbf{c}(t)$. If we apply $\mathbf{T}$ to the

local coordinate functions x^α we obtain the traditional velocity vector, i.e. $T(x^\alpha) = dc^\beta/dt\, \partial_\beta x^\alpha = dc^\alpha/dt$. So, a tangent vector **T** is now a differential operator of the form $\mathbf{T} = T^\alpha \partial_\alpha$. The space of all possible tangents at a point **c** is called the tangent space of N at **c** and is written as $T_c N$. In terms of local coordinates the differential operators ∂_α, $\alpha = 1, \ldots, n$, form a basis of $T_c N$. If the components T^α are smooth functions, then $\mathbf{T} = T^\alpha(x)\partial_\alpha$ is called a vector field on N.

In addition to partial differentiation, a second differential operator is commonly introduced on a manifold. This is the operator of **covariant differentiation.** It is closely related to the concept of connections. The subject begins by observing that the tangent spaces $T_x N$, $T_{x'} N$ at two neighbouring points **x** and **x'** change as one moves from **x** to **x'**. A connection is essentially a structure which endows one with the ability to compare two such tangent spaces at a pair of inifinitesimally separated points. The connection is given by defining what is called parallel transport or parallel translation in N. Consider $T_x N$ and $T_{x'} N$ and any curve, **c** say, joining **x** to **x'**. Let **T** be a tangent to the curve **c** at **x**, then **T** is said to be parallely transported along the curve **c** if **T** is pushed from **x** to **x'** in such a way to always remain parallel to itself. If t is the parameter of the curve then the covariant derivative of **T** is the rate of change of **T** with respect to t. This covariant derivative will differ from the ordinary partial derivative, the quantity that measures this difference is the connection.

Let **X** and **Y** be vector fields on N. The covariant derivative of **Y** with respect to **X** is then denoted by $\nabla_X Y$ and it is a vector field on N. The application of the operator ∇ is defined to be linear in both its arguments and must satisfy the chain rule $\nabla_X(fY) = X(f)Y + f\nabla_X Y$, where f is any real-valued smooth function on N. With the local coordinate expressions $\mathbf{X} = X^\alpha \partial_\alpha$, $\mathbf{Y} = Y^\alpha \partial_\alpha$ we therefore get

$$\nabla_X Y = \nabla_X(Y^\beta \partial_\beta) = X(Y^\beta)\partial_\beta + Y^\beta \nabla_X \partial_\beta = X^\alpha \partial_\alpha (Y^\beta)\partial_\beta + X^\alpha Y^\beta \nabla_{\partial_\alpha} \partial_\beta \quad , (2.1)$$

which shows that $\nabla_X Y$ is totally specified once $\nabla_{\partial_\alpha} \partial_\beta$ is given. It is customary to express these vectors fields in the coordinate fields ∂_γ as

$$\nabla_{\partial_\alpha} \partial_\beta = \Gamma^\gamma_{\alpha\beta}\, \partial_\gamma \,, \qquad \alpha, \beta, \gamma = 1, \ldots, n. \tag{2.2}$$

The n^3 real-valued smooth functions $\Gamma^\gamma_{\alpha\beta}$ determine the connection and are called the connection coefficients.

Let $\mathbf{c}(t)$ be a curve in N. A vector field **X** on N is then said to be a parallel vector field along the curve $\mathbf{c}(t)$, if its covariant derivative with respect to the direction $\mathbf{T} = dc^\alpha/dt\, \partial_\alpha$ is identically zero, i.e.,

$$\nabla_T X = 0. \tag{2.3}$$

There are special types of curves $\mathbf{c}(t)$ which are so-called self parallel. That is, parallel transport from t to t' takes the velocity vector at $\mathbf{c}(t)$ into the velocity vector at $\mathbf{c}(t')$. These curves are called geodesics. Since the covariant derivative $\nabla_T T$ measures the rate of change of **T** in the

direction $\mathbf{T}$ under parallel transport, an equation describing the above definition of a geodesic is simply

$$\nabla_{\mathbf{T}}\mathbf{T} = \mathbf{0}, \tag{2.4}$$

where $\mathbf{T}$ is the velocity vector of $\mathbf{c}(t)$. With $\mathbf{T} = dc^{\alpha}/dt\ \partial_{\alpha}$, (2.1) and (2.2), (2.4) becomes in local coordinates

$$\frac{d^2c^{\alpha}}{dt^2} + \Gamma^{\alpha}_{\beta\gamma}\,\frac{dc^{\beta}}{dt}\,\frac{dc^{\gamma}}{dt} = 0. \tag{2.4'}$$

So far we have equipped the manifold N only with a connection given by the defining equation (2.2). We will now give it some additional structure. Assume given a smooth real-valued, symmetric and positive-definite bi-linear map $\langle .,. \rangle_{xN} : T_xN \times T_xN \to \mathbb{R}$. A manifold equipped with such a bi-linear map is called a **Riemannian manifold.** The bi-linear map $\langle .,. \rangle_{xN}$ is called the metric tensor and in local coordinates it is given by the smooth functions $g_{\alpha\beta}(\mathbf{x}) = \langle \partial_{\alpha}, \partial_{\beta} \rangle_{xN}$.
There is a unique symmetric connection on a Riemannian manifold such that parallel translation preserves the Riemannian metric. It is called the Levi-Civita or Riemannian connection. It is that unique connection satisfying

$$\begin{aligned} &\text{a)} \quad \nabla_X Y - \nabla_Y X = X\,Y - Y\,X \\ &\text{b)} \quad Z\langle X,Y\rangle_N = \langle \nabla_Z X, Y\rangle_N + \langle X, \nabla_Z Y\rangle_N , \end{aligned} \tag{2.5}$$

for any vector fields $\mathbf{X},\mathbf{Y}$ and $\mathbf{Z}$ on N. A connection satisfying (2.5.a) is said to be symmetric or torsionfree, and a connection satisfying (2.5.b) is said to be metric.

Up till now we have considered only one manifold N. Let us now consider two manifolds N and M, and a smooth injective map $\mathbf{y}$ between them, i.e. $\mathbf{y}: N \to M$. Then the image $\bar{N} = \mathbf{y}(N) \subset M$ defines a submanifold of M.
The map $\mathbf{y}$ provides a way of mapping vectors on N into vectors on M. The image of T_xN under $\mathbf{y}$ is a tangent space of $\bar{N}$ at $\mathbf{y(x)}$, and is denoted by $T_{\mathbf{y}(\mathbf{x})}\bar{N}$. This map between tangent spaces induced by $\mathbf{y}$ is written $\mathbf{y}_*: T_xN \to T_{\mathbf{y}(\mathbf{x})}\bar{N}$ and is called the push forward of $\mathbf{y}$. The precise action on a vector $X \in T_xN$ is such that given a function f on M, so that $f(\mathbf{y(x)})$ is a function on N, then $\mathbf{y}_*(X) \in T_{\mathbf{y}(\mathbf{x})}\bar{N}$ is defined by $(\mathbf{y}_*X)_{\mathbf{x}} f = Xf(\mathbf{y}(\mathbf{x}))$. With $X = X^{\alpha}\partial_{\alpha}$ this would give in local coordinates

$$(\mathbf{y}_*X)_{\mathbf{x}} f = X\, f(\mathbf{y}(\mathbf{x})) = X^{\alpha}\partial_i f\ \partial_{\alpha} y^i = X^{\alpha}\partial_{\alpha} y^i\ \partial_i f ,$$

or

$$\mathbf{y}_*(X)_{\mathbf{x}} = X^{\alpha}(\mathbf{x})\ \partial_{\alpha} y^i(\mathbf{x})\ \partial_i ,$$

where y^i, $i=1,\ldots,m$, are the local coordinate functions of M, ∂_i, $i=1,..,m$, the corresponding coordinate vector fields and $y^i(x^{\alpha})$ the coordinization of the map $\mathbf{y}: N \to M$.
Although it is possible to suppress explicit reference to the map $\mathbf{y}$, to identify N with the subset

$\mathbf{y}(N)$ of M and each $T_x N$ with the subspace $\mathbf{y}_*(T_x N)$ of $T_y M$, we will not do so. Recall namely that also in the case of linear maps we are not used to identify the range space with the domain space, although both spaces are isomorphic.

As a closing of this section we define the observation- and parameter space of our adjustment. In our least-squares adjustment context the observation space M is taken to be Euclidean with Euclidean metric $\langle \cdot,\cdot \rangle_M$. The coefficients of the metric are given by the real-valued constants $g_{ij} = \langle \partial_i, \partial_j \rangle_M$. The connection compatible with the Euclidean metric of M will be denoted by D. And since $D_{\partial_i} \partial_j = 0$, $i,j=1,\ldots,m$, we have for any two vector fields $\mathbf{V}$ and $\mathbf{W}$ on M that $D_{\mathbf{V}}\mathbf{W} = V^i \partial_i (W^j) \partial_j$, i.e. the covariant derivative reduces to the ordinary vector derivative. The directional derivative of a function f on M in a direction $\mathbf{V}$ will sometimes be denoted by $D_{\mathbf{V}} f$.
Manifold N will play the role of the parameter space and the non-linear map $\mathbf{y}$ replaces the linear map $\mathbf{A}$ which has been used hitherto. Manifold N will be endowed with a Riemannian metric by pulling the metric of M back by $\mathbf{y}$. That is, given the metric of M we define the metric of N by

$$\langle X,Y \rangle_N = \langle \mathbf{y}_*(X), \mathbf{y}_*(Y) \rangle_M \quad \text{for any} \quad X,Y \in T_x N. \tag{2.6}$$

3. Orthogonal projection onto a parametrized space curve

3.1. Gauss' iteration method

It seems reasonable that we should begin our discussion of non-linear least-squares adjustment with the simplest class of problems, namely those in which manifold N is one dimensional. In case of our least-squares problem

$$\min_{\mathbf{y} \in \bar{N} = \mathbf{y}(N)} \langle \mathbf{y}_s - \mathbf{y}, \mathbf{y}_s - \mathbf{y} \rangle_M = \langle \mathbf{y}_s - \hat{\mathbf{y}}, \mathbf{y}_s - \hat{\mathbf{y}} \rangle_M , \tag{3.1}$$

this means that we need to consider the problem of orthogonally projecting the observation-point $\mathbf{y}_s \in M$ onto a space curve.
Since we like to denote the space curve by $\mathbf{c}(t)$, we replace the map $\mathbf{y}: N \to M$ in this section by the map

$$\mathbf{c}: t \in \mathbb{R} = N \to M . \tag{3.2}$$

Our univariate least-squares adjustment problem reads then

$$\min_{\mathbf{c} \in \bar{N} = \mathbf{c}(N)} \langle \mathbf{y}_s - \mathbf{c}, \mathbf{y}_s - \mathbf{c} \rangle_M = \langle \mathbf{y}_s - \hat{\mathbf{c}}, \mathbf{y}_s - \hat{\mathbf{c}} \rangle_M . \tag{3.3}$$

From geometric reasoning it will be clear that a necessary condition for $\hat{\mathbf{c}}$ to be the least-squares solution of (3.3), is that

$$\langle \mathbf{c}_*(\frac{d}{dt}),\ \mathbf{y}_s - \mathbf{c} \rangle_{\hat{\mathbf{c}},M} = 0, \tag{3.4}$$

must hold, where $\frac{d}{dt}$ is a basis of $T_t\mathbb{R} = T_t N$.
In the linear case it was necessary and also sufficient for the residual vector to be orthogonal to the linear submanifold $\bar{N} = \mathbf{A}N$. In the non-linear case however, it is necessary but not sufficient.
Since the residual vector $\mathbf{y}_s - \hat{\mathbf{c}}$ needs to be orthogonal to the linear tangent space $T_{\hat{\mathbf{c}}}(\mathbf{c}(N)) = T_{\hat{\mathbf{c}}}\bar{N}$ of the non-linear manifold $\bar{N} = \mathbf{c}(N)$ at $\hat{\mathbf{c}}$, we need to know $T_{\hat{\mathbf{c}}}\bar{N}$. But due to the assumed non-linearity of the mapping $\mathbf{c}: N = \mathbb{R} \to M$, the tangent space $T_{\hat{\mathbf{c}}}(\mathbf{c}(N))$ is generally unknown a priori. Hence our minimization problem cannot be solved directly. Expression (3.4) does however suggest a way of solving our adjustment problem. Instead of orthogonally projecting $\mathbf{y}_s$ onto the tangent space $T_{\hat{\mathbf{c}}}\bar{N}$, one can take as a first approximation the orthogonal projection of $\mathbf{y}_s$ onto a nearby tangent space, $T_{\mathbf{c}_q}\bar{N}$ say. Of course then,

$$\langle \mathbf{c}_*(\frac{d}{dt}),\ \mathbf{y}_s - \mathbf{c} \rangle_{\mathbf{c}_q,M} \neq 0. \tag{3.5}$$

But by pulling the non-orthogonality as measured by (3.5) back to the Riemannian manifold N, we get

$$\langle \frac{d}{dt},\ \Delta t_q \rangle_{t_q,N} = \langle \mathbf{c}_*(\frac{d}{dt}),\ \mathbf{y}_s - \mathbf{c} \rangle_{\mathbf{c}_q,M} \text{ , with } \Delta t_q \in T_{t_q}\mathbb{R} = T_{t_q}N \text{ ,} \tag{3.6}$$

which suggests in local coordinates the following iteration procedure:

$$\boxed{\begin{aligned} \Delta t_q &= g(t_q)^{-1}\, \frac{dc^i}{dt}(t_q)\, g_{ij}(y_s^j - c^j(t_q)), \quad i,j = 1,\ldots,m. \\ t_{q+1} &= t_q + \Delta t_q \end{aligned}} \text{ ,} \tag{3.7}$$

where g(t) is the induced metric of $N = \mathbb{R}$.
This is **Gauss' iteration method** and it consists of successively solving a linear least-distance adjustment problem until condition (3.4) is met.

Before we now proceed with studying the local behaviour of Gauss' iteration method (3.7), we will first derive some local geometric properties of the space curve $\mathbf{c}$ itself. An appropriate approach for studying the local geometry of curve $\mathbf{c}$ is by using

3.2. The Frenet frame

With the tangent field (or velocity field if one considers $t \in \mathbb{R}$ to be a time parameter)

$$V = c_* \left(\frac{d}{dt}\right)$$

of curve $c(t)$, we obtain for non-zero velocities the unit tangent field T as

$$T = V/||V||_M .$$

And since $||T||_M = 1$ for all admissible $t \in \mathbb{R}$, we have

$$0 = T\langle T, T\rangle_M = \langle D_T T, T\rangle_M + \langle T, D_T T\rangle_M ,$$

which shows that $D_T T$ is orthogonal to the unit tangent field T. We define the **first curvature** k_1 as

$$k_1 = ||D_T T||_M , \tag{3.8}$$

and when $k_1 > 0$ the **first normal** N_1 by

$$k_1 N_1 = D_T T . \tag{3.9}$$

Geometrically the first curvature k_1 can be seen to determine the rate of change of the direction of the tangent to the curve with respect to its arclength, where arclength is defined as

$$s(t) = \int_{t_o}^{t} ||V(t')||_M \, dt' . \tag{3.10}$$

The curvature k_1 is a property of the curve c and it is invariant to a reparametrization.
From the orthogonality of N_1 and T follows that

$$0 = T\langle N_1, T\rangle_M = \langle D_T N_1, T\rangle_M + \langle N_1, D_T T\rangle_M = \langle D_T N_1, T\rangle_M + k_1 ,$$

which shows that T is orthogonal to $D_T N_1 + k_1 T$. Similarly it follows from $||N_1||_M = 1$ that

$$0 = T\langle N_1, N_1\rangle_M = \langle D_T N_1, N_1\rangle_M + \langle N_1, D_T N_1\rangle_M .$$

Thus $D_T N_1 + k_1 T$ is orthogonal to both N_1 and T. We now define **the second curvature** k_2 as

$$k_2 = ||D_T N_1 + k_1 T||_M , \tag{3.11}$$

and when $k_2 > 0$ the second normal N_2 by

$$k_2 N_2 = D_T N_1 + k_1 T . \tag{3.12}$$

We can proceed in this way to define k_3, N_3 etc. The vectors T, N_1, N_2, ... are called the Frenet vectors and the equations that express the $D_T T$, $D_T N_i$ in terms of the Frenetvectors are called the Frenet equations. For the case $m=3$ they read as

$$\left.\begin{aligned} D_T T &= k_1 N_1 \\ D_T N_1 &= -k_1 T + k_2 N_2 \\ D_T N_2 &= -k_2 N_1 \end{aligned}\right\} \qquad (3.13)$$

In order to find the relative position of the curve **c** with respect to its Frenetframe at some regular point, we can study the projections of the curve onto the planes of the Frenetframe. For convenience we assume that the curve **c** has been parametrized with the arclength parameter s . Now let our point, P say, correspond to the value s = 0 of the arclength parameter. The curve **c**(s) can then be written in the form

$$\mathbf{c}(s) = \mathbf{c}(o) + T_o s + \frac{1}{2}(D_T T)_o s^2 + \frac{1}{6}(D_T^2 T)_o s^3 + o(s^3). \qquad (3.14)$$

The subscript "**o**" denotes that the value is taken at the point corresponding to s = 0. And Landau's o(.) symbol means that $o(s^3)/s^3 \to 0$ if $s \to 0$.
Since

$$D_T T = k_1 N_1,$$

it follows that

$$D_T^2 T = D_T(k_1 N_1) = T(k_1) N_1 + k_1 D_T N_1 = \frac{dk_1}{ds} N_1 + k_1 (-k_1 T + k_2 N_2).$$

Substituting the above two expressions into (3.14) gives then with $k_1' = \frac{dk_1}{ds}$:

$$\mathbf{c}(s)-\mathbf{c}(o) = (s - \tfrac{1}{6}k_1^2(0)s^3)T_o + (\tfrac{1}{2}k_1(0)s^2 + \tfrac{1}{6}k_1'(0)s^3)N_{1,o} + (\tfrac{1}{6}k_1(0)k_2(0)s^3)N_{2,o} + o(s^3).$$

Choose now a special coordinate system in M such that the point P under consideration is the origin and the vectors T_o, $N_{1,o}$ and $N_{2,o}$ are the unit vectors of the first three coordinate axes. In this coordinate system the curve **c**(s) can be represented by the equations

$$\left.\begin{aligned} c^{i=1}(s) &= s - \frac{1}{6}k_1^2(0)s^3 + o(s^3) \\ c^{i=2}(s) &= \frac{1}{2}k_1(0)s^2 + \frac{1}{6}k_1'(0)s^2 + o(s^3) \\ c^{i=3}(s) &= \frac{1}{6}k_1(0)k_2(0)s^3 + o(s^3) \\ c^{i}(s) &= o(s^3), \quad i = 4,\ldots,m. \end{aligned}\right\} \qquad (3.15)$$

These equations are called the **canonical representation** of curve **c**(s) at s = 0, and the leading terms in it conveniently describe the behaviour of **c**(s) near the point corresponding to s = 0.
It will be clear that many curves exist which have up to $o(s^3)$ the same canonical representation as **c**(s). That is, for s small enough these curves behave alike and are thus indistinguishable.
We will now give a characterization of such "kissing" curves and one of them, namely

will be used for a further analysis of Gauss' iteration scheme (3.7).
Consider two curves $\mathbf{c}_1(s_1)$ and $\mathbf{c}_2(s_2)$ with a common point $\mathbf{c}_1(o) = \mathbf{c}_2(o)$. s_1 and s_2 are taken as their natural arclength parameter. Let $\mathbf{c}_1(s_1=h)$ and $\mathbf{c}_2(s_2=h)$ be two points on respective $\mathbf{c}_1(s_1)$ and $\mathbf{c}_2(s_2)$. We say that the two curves have a **contact of order** n if

$$||\mathbf{c}_1(h) - \mathbf{c}_2(h)||_M = o(h^n),$$

but

$$||\mathbf{c}_1(h) - \mathbf{c}_2(h)||_M \neq o(h^{n+1}) \quad \text{as } h \to 0 .$$

From this follows that two curves $\mathbf{c}_1(s_1)$ and $\mathbf{c}_2(s_2)$ have a contact of order n at a regular point corresponding to $s_1 = s_2 = 0$ if and only if

$$c_1^i(0) = c_2^i(0),\ \frac{dc_1^i}{ds_1}(0) = \frac{dc_2^i}{ds_2}(0),\ldots,\ \frac{d^n c_1^i}{ds_1^n}(0) = \frac{d^n c_2^i}{ds_2^n}(0),\ \frac{d^{n+1} c_1^i}{ds_1^{n+1}}(0) \neq \frac{d^{n+1} c_2^i}{ds_2^{n+1}}(0),$$

$i = 1,\ldots,m,$

where the coordinates of the two curves are given with respect to a fixed frame of M. With (3.15) follows then that two curves have a contact of order at least two at a common point P if and only if they have at P a common tangent vector $\mathbf{T}_o$, a common normal $\mathbf{N}_{1,o}$ and moreover, the same curvature $k_1(0)$. All such curves will thus have the same canonical representation

$$\mathbf{c}(s) - \mathbf{c}(o) = s\mathbf{T}_o + \frac{1}{2} k_1(0) s^2 \mathbf{N}_{1,o} + o(s^2). \tag{3.16}$$

And in the above sense of contact such curves can be considered each others best approximation. Now, if we recall our iteration scheme (3.7) we observe that only first order derivative information is used. Hence, for a small enough portion of the curve $\mathbf{c}(s)$ about the least-squares solution $\hat{\mathbf{c}} = \mathbf{c}(o)$, we can replace the space curve $\mathbf{c}(s)$ by

$$\bar{\mathbf{c}}(s) = \mathbf{c}(o) + s\mathbf{T}_o + \frac{1}{2} k_1(0) s^2 \mathbf{N}_{1,o}.$$

In fact, with the same approximation we can replace the space curve $\mathbf{c}(s)$ by the circle

$$\mathbf{C}(s) = \mathbf{c}(o) + k_1(0)^{-1} \sin(k_1(0)s)\mathbf{T}_o + [k_1(0)^{-1} - k_1(0)^{-1}\cos(k_1(0)s)]\mathbf{N}_{1,o} \tag{3.17}$$

This follows from

$$k_1(0)^{-1}\sin(k_1(0)s) = s + o(s^2)$$

and

$$k_1(0)^{-1}\cos(k_1(0)s) = k_1(0)^{-1} - \frac{1}{2} k_1(0) s^2 + o(s^2).$$

Thus we can use the circle C(s) to replace the curve c(s) in a neighborhood of P. The circle C(s) is known as the **osculating (="kissing") circle** of c(s) at $\hat{c} = c(0)$ or the **circle of curvature.**

Note that by replacing c(s) by C(s) we achieve a drastic simplification of our original non-linear least-squares adjustment problem. First of all we achieve a dramatic decrease in dimensionality: $c(s) \in M$, whereas C(s) lies in a two-dimensional plane of M spanned by T_0 and $N_{1,0}$. And secondly we can now exploit the simple geometry of the osculating circle C(s) in order to understand the local behaviour of Gauss' iteration method (3.7).

Consider therefore the situation as sketched in figure 24.

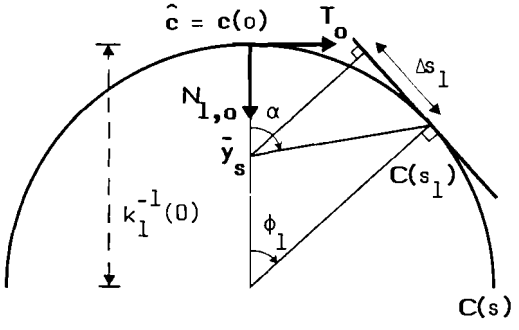


figure 24

$\bar{y}_s$ is the orthogonal projection of the observation point $y_s \in M$ onto the plane spanned by T_0 and $N_{1,0}$, and $C(s_1)$ is the initial guess to start the iteration procedure.

Since the orthogonal projection of y_s onto the tangent of C(s) at $C(s_1)$ gives the same increment Δs_1 as the orthogonal projection of $\bar{y}_s$, we have for our first iteration step

$$\begin{aligned}\left\langle \frac{d}{ds} \, , \, \Delta s_1 \right\rangle_{s_1,N} &= \left\langle C_*\left(\frac{d}{ds}\right), \, y_s - C \right\rangle_{C(s_1),M} \\ &= \left\langle C_*\left(\frac{d}{ds}\right), \, \bar{y}_s - C \right\rangle_{C(s_1),M} \\ &= -\| \bar{y}_s - C(s_1) \|_M \sin(\alpha - \phi_1) \quad ,\end{aligned}$$

or

$$\Delta s_1 = -\| \bar{y}_s - C(s_1) \|_M \sin(\alpha - \phi_1) \quad . \tag{3.18}$$

From the figure also follows that in a sufficiently small neighbourhood of $\hat{c}$,

$$\phi_1 \simeq \tan \phi_1 = \frac{\| \bar{y}_s - C(s_1) \|_M \sin \alpha}{k_1(0)^{-1} - \| \bar{y}_s - \hat{c} \|_M + \| \bar{y}_s - C(s_1) \|_M \cos \alpha} \, ,$$

or

$$\begin{aligned}\phi_1 (k_1(0)^{-1} - \| \bar{y}_s - \hat{c} \|_M) &\simeq \| \bar{y}_s - C(s_1) \|_M \, (\sin \alpha - \phi_1 \cos \alpha) \\ &\simeq \| \bar{y}_s - C(s_1) \|_M \sin(\alpha - \phi_1) \quad .\end{aligned} \tag{3.19}$$

With $\phi_1 = k_1(0)s_1$, combination of (3.18) and (3.19) gives

$$\Delta s_1 \simeq (k_1(0)|| \bar{\mathbf{y}}_s - \hat{\mathbf{c}}||_M - 1)s_1 .$$

And with $s_2 = s_1 + \Delta s_1$, this finally gives the relation

$$\boxed{s_2 = s_1 + \Delta s_1 \simeq k_1(0)|| \bar{\mathbf{y}}_s - \hat{\mathbf{c}}||_M \, s_1 = \langle k_1(0)\mathbf{N}_{1,o},\ \mathbf{y}_s - \hat{\mathbf{c}} \rangle_M \, s_1} \qquad (3.20)$$

From this expression we can now formulate several important conclusions concerning the local behaviour of Gauss' iteration method as applied to the curve $\mathbf{c}(s)$: First of all expression (3.20) tells us that in case $k_1(0) \neq 0$, the local convergence behaviour of Gauss' iteration method as applied to the space curve $\mathbf{c}(s)$, is **linear.** That is, the computed arclength of the curve $\mathbf{c}(s)$ from $\hat{\mathbf{c}}$ to $\mathbf{c}(s_{q+1})$ depends linearly on the computed arclength from $\hat{\mathbf{c}}$ to the point $\mathbf{c}(s_q)$ of the preceding step.
Secondly, a necessary condition for convergence of Gauss' iteration method is that

$$|\langle \mathbf{N}_1, \mathbf{y}_s - \mathbf{c} \rangle_{\hat{\mathbf{c}},M}| < k_1(0)^{-1} . \qquad (3.21)$$

And thirdly, expression (3.20) shows that the local linear convergence behaviour is determined by two terms, namely the first curvature k_1 of the curve $\mathbf{c}(s)$ at $\hat{\mathbf{c}}$ and the projection $\langle \mathbf{N}_1, \mathbf{y}_s - \mathbf{c} \rangle_{\hat{\mathbf{c}},M}$ of the residual vector $\mathbf{y}_s - \hat{\mathbf{c}}$ onto the first normal $\mathbf{N}_1$ at $\hat{\mathbf{c}}$. Thus the smaller the curvature and the smaller the component of $\mathbf{y}_s - \hat{\mathbf{c}}$ in the direction of $\mathbf{N}_1$, the faster Gauss' iteration method as applied to the space curve $\mathbf{c}(s)$ converges.

So far we assumed for convenience that the curve $\mathbf{c}: \mathbb{R} = N \to M$ was parametrized with its arclength parameter s. But in general one would of course have an arbitrary parametrization $\mathbf{c}(t)$, with $t \neq s$. The question that remains is then whether the above given conclusions still hold when $t \neq s$. To study this more general case, it seems appropriate to look for the direct analogon of the Frenet equations (3.13). These are given by the so-called

3.4 One dimensional Gauss- and Weingarten equations

From the definition of the arclength parameter s ,

$$s(t) = \int_{t_o}^{t} ||\mathbf{V}(t')||_M \, dt' , \text{ with } \mathbf{V} = \mathbf{c}_*(\frac{d}{dt}) ,$$

follows that

$$s'(t) = \frac{ds}{dt}(t) = ||\mathbf{V}(t)||_M . \qquad (3.22)$$

We therefore have that

$$D_V V = D_{s'T}(s'T) = s'D_T(s'T) = s'T(s')T + (s')^2 D_T T = s''T + (s')^2 D_T T.$$

And with (3.22) and $D_T T = k_1 N_1$ follows that

$$D_V V = (s')^{-1}(s'')V + (s')^2 k_1 N_1.$$

In a similar way we find that

$$D_V N_1 = (s')D_T N_1 \text{ and } D_V N_2 = (s')D_T N_2.$$

With these last three equations we can now replace (3.13) by

$$\left.\begin{aligned} D_V V &= (s')^{-1}(s'')V + (s')^2 k_1 N_1 \\ D_V N_1 &= -k_1 V + (s')k_2 N_2 \\ D_V N_2 &= -(s')k_2 N_1 \end{aligned}\right\} \qquad (3.23)$$

For m = 3, these equations can be considered as the one-dimensional analogons of the **Gauss- and Weingarten equations.**

3.5 Local convergence behaviour of Gauss' iteration method

Now let us return to our adjustment problem and see how the equations (3.23) come to our use for describing the local properties of iteration scheme (3.7).
First observe that (3.7) can also be written as

$$\Delta t_q = -g(t_q)^{-1} \frac{dE}{dt}(t_q) \quad . \qquad (3.24)$$

Expanding the right-hand side into a Taylor series about the least-squares solution $\hat{t}$ gives then with $\frac{dE}{dt}(\hat{t}) = 0$:

$$\Delta t_q = -g(\hat{t})^{-1} \frac{d^2E}{dt^2}(\hat{t})(t_q - \hat{t}) + \tfrac{1}{2}(2g(\hat{t})^{-2} \frac{dg}{dt}(\hat{t})\frac{d^2E}{dt^2}(\hat{t}) - g(\hat{t})^{-1} \frac{d^3E}{dt^3}(\hat{t}))(t_q - \hat{t})^2 + o((t_q - \hat{t})^2) \qquad (3.25)$$

And with

$$\frac{d^2E}{dt^2}(\hat{t}) = \langle V, V\rangle_{\hat{c},M} - \langle D_V V, y_s - c\rangle_{\hat{c},M} \ ,$$

$$\langle \mathbf{V},\mathbf{V}\rangle_{\hat{\mathbf{c}},M} = g(t) = (s'(t))^2,$$

$$D_{\mathbf{V}}\mathbf{V} = (s')^{-1}(s'')\mathbf{V} + (s')^2 k_1 \mathbf{N}_1, \text{ and}$$

$$t_{q+1} - t_q = \Delta t_q,$$

the above expression (3.25) reduces to

$$\boxed{t_{q+1} - \hat{t} \simeq \langle k_1\mathbf{N}_1, \mathbf{y}_s - \mathbf{c}\rangle_{\hat{\mathbf{c}},M}(t_q - \hat{t})} \quad . \tag{3.26}$$

But this is exactly the result we obtained in (3.20) for the special case $t = s$, $\hat{t} = 0$. Hence, we have as a fourth conclusion that the local linear convergence behaviour of Gauss' iteration method as applied to the space curve $\mathbf{c}(t)$, is **invariant** to any admissible non-linear parameter transformation. It is thus idle hope to think that one can improve the convergence behaviour by changing to a different coordinate system.

Now let us assume that the first curvature k_1 of the space curve $\mathbf{c}(t)$ is identically zero.
Then

$$D_{\mathbf{T}}\mathbf{T} = \mathbf{0} \quad ,$$

which means that the unit tangent vector $\mathbf{T}$ is parallel along the whole curve $\mathbf{c}(t)$. And since M is Euclidean by assumption, this means that the curve $\mathbf{c}(t)$ is a straight line. From (3.25) follows then that

$$t_{q+1} - \hat{t} = \tfrac{1}{2}\left[2g(\hat{t})^{-2}\frac{dg}{dt}(\hat{t})\frac{d^2E}{dt^2}(\hat{t}) - g(t)^{-1}\frac{d^3E}{dt^3}(\hat{t})\right](t_q-\hat{t})^2 + o((t_q-\hat{t})^2). \tag{3.27}$$

And with

$$\frac{d^2E}{dt^2}(\hat{t}) = g(\hat{t}) \quad ,$$

$$g(\hat{t}) = (s'(\hat{t}))^2 \quad , \text{ and} \tag{3.28}$$

$$\frac{d^3E}{dt^3}(\hat{t}) = 3\, s'(\hat{t})s''(\hat{t}) \quad ,$$

for $k_1 = 0$, follows then that

$$\boxed{t_{q+1} - \hat{t} \simeq \tfrac{1}{2}\left[(s'(\hat{t}))^{-1}s''(\hat{t})\right](t_q-\hat{t})^2} \quad . \tag{3.29}$$

Hence, for the case the curve $\mathbf{c}(t)$ is a straight line ($k_1 \equiv 0$), Gauss' iteration scheme (3.7) will have a local **quadratic** convergence behaviour. But how is this possible? Doesn't orthogonal projection onto a straight line correspond to the case of linear least-squares adjustment. And if so, wouldn't that mean that iterations are superfluous? The answer is partly in the affirmative and partly in the negative. It essentially boils down to our earlier remarks made in the previous chapters, namely that adjustment in the general sense should be thought of as being composed of the problem of adjustment in the narrow sense, i.e. the problem of finding an estimate $\hat{y}$ such that $\min_{y \in \bar{N}} \langle y_s - y, y_s - y \rangle_M = \langle y_s - \hat{y}, y_s - \hat{y} \rangle_M$, and the problem of inverse mapping, i.e. the problem of finding the pre-image $\hat{x}$ of $\hat{y}$ under the map $\mathbf{y}: N \to M$. Thus the actual adjustment part, namely that of finding the point $\hat{y}$ in the submanifold $\bar{N}$ of M which has smallest distance to $y_s \in M$, is essentially an observation space oriented problem. In this light we must therefore be more precise as to what we mean by "linear least-squares adjustment". Usually one means by "linear least-squares adjustment" that the coordinate functions $y^i(x^\alpha)$, $i = 1,\ldots,m$, $\alpha = 1,\ldots,n$ of the map $\mathbf{y}$ are linear. We will, however, call a least-squares adjustment problem linear, if the submanifold $\bar{N}$ of the Euclidean observation space M defined by the map $\mathbf{y}: N \to M$, is linear or flat. For our problem of orthogonal projection onto the curve $\mathbf{c}$ this means that the adjustment problem is termed linear if $k_1 = 0$. But it also means that linear least-squares problems may admit non-linear functions $c^i(t)$, $i = 1,\ldots,m$. The non-linearity in $c^i(t)$ is then only caused by the choice of the parameter t. That is, by choosing another parameter it is possible to eliminate the non-linearity in $c^i(t)$. In particular if one takes the arclength parameter s or a linear function thereof as parameter, the functions $c^i(t)$ will become linear. As a consequence we see that the local quadratic convergence factor of (3.29) is not a property of the curve $\mathbf{c}(t)$ itself, but instead depends on its parametrization. In the special case namely of $t = s$, we would have $(s')^{-1}s'' = 0$, i.e. no iteration would be necessary then. Thus we see that with (3.29) we are actually solving for the inverse mapping problem, instead of the actual adjustment problem.

To put the argument geometrically, consider an arbitrary parametrization of the straight line $\mathbf{c}$ such that the parameter t is not a linear function of the arclength s. The length $\|\mathbf{V}(t)\|_M = s'(t)$ of the curve's tangent vector $\mathbf{V}$ changes then when moving along the curve from point to point. Hence, the coordinate expression of the induced metric of N, $g(t) = \langle \mathbf{c}_*(\frac{d}{dt}), \mathbf{c}_*(\frac{d}{dt}) \rangle_M$, will be a function of the parameter t. But this means that when one applies formula (3.7) of Gauss' iteration method one is in fact using two different "yardsticks". One yardstick given by the pulled back metric of the tangent space of the curve $\mathbf{c}$ at point $\mathbf{c}(t_q)$, namely $g(t_q)$, and a second yardstick, namely $g(t)$, the induced metric of the parameter space N itself. And it will be clear that the induced metric $g(t_q)$ of the linear tangent space $T_{t_q} N$ will be constant for the whole space, whereas the induced metric $g(t)$ of N itself changes from point to point. Thus when one computes the tangent vector $\Delta \mathbf{t}_q = \Delta t_q \frac{d}{dt} \in T_{t_q} N$ through

$$\langle \frac{d}{dt}, \Delta \mathbf{t}_q \rangle_{t_q, N} = \langle \mathbf{c}_*(\frac{d}{dt}), y_s - \mathbf{c} \rangle_{\mathbf{c}(t_q), M}$$

and adds its coordinate Δt_q to t_q, to obtain

$$t_{q+1} = t_q + \Delta t_q,$$

one is in fact neglecting that $T_{t_q}N$ and N are endowed with two different metric tensors (see figure 25).

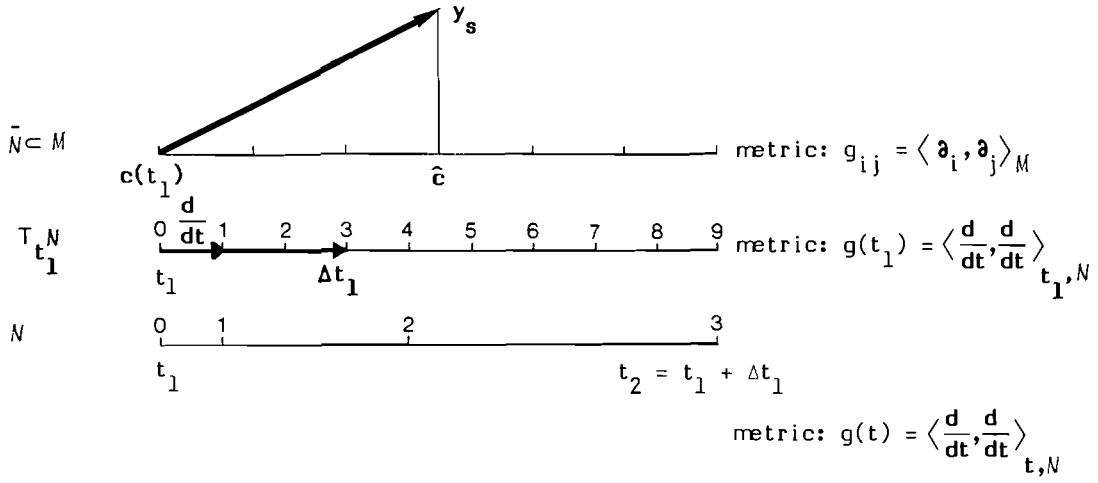


figure 25

And because of this neglectance one is, despite the flatness of $\bar{N}$, still forced to recourse to an iteration to find $\hat{t}$. Note, however, that if one is not interested in $\hat{t}$, but instead is satisfied with $\hat{c}$ no iteration is necessary. From the linearity of the submanifold $\bar{N} = c(N)$ follows namely that

$$\hat{c}^i = c^i(t_q) + \frac{dc^i}{dt}(t_q) g(t_q)^{-1} \frac{dc^j}{dt}(t_q) g_{jk} (y_s^k - c^k(t_q))$$

is independent of the choice for t_q.

Since (3.28) also holds for the case $k_1 \neq 0$ but $\mathbf{y}_s - \hat{\mathbf{c}} = \mathbf{0}$, it follows that we also have the local quadratic convergence rule (3.29) for zero residual vector adjustment problems. This is in fact not very surprising since for both the cases $k_1 = 0$ and $\mathbf{y}_s - \hat{\mathbf{c}} = \mathbf{0}$, we do not need an iteration to solve the actual adjustment problem. In case of $k_1 = 0$ the actual adjustment problem is namely linear and in case of $\mathbf{y}_s - \hat{\mathbf{c}} = \mathbf{0}$ the actual adjustment problem is indeed already solved a priori, since $\mathbf{y}_s = \hat{\mathbf{c}}$. Thus for both the cases $k_1 = 0$ and $\mathbf{y}_s - \hat{\mathbf{c}} = \mathbf{0}$, the iteration is only needed for the inverse mapping problem and not for the actual adjustment problem.

To illustrate the theory developed so far and to demonstrate the various effects mentioned we will now give some examples.

3.6 Examples

Example 1: Orthogonal projection onto the curve O(2).

In this first example we take as non-linear model the two dimensional **Helmert transformation** only admitting a rotation. The non-linear model reads

$$\begin{aligned} \tilde{x}_i &= \bar{x}_i \cos\theta + \bar{y}_i \sin\theta \\ \tilde{y}_i &= -\bar{x}_i \sin\theta + \bar{y}_i \cos\theta \ , \end{aligned} \tag{3.30}$$

where:
- $i = 1,\ldots,n$ = number of network points,
- the tilde "$\sim$" sign stands for the mathematical expectation,
- x_i, y_i are cartesian coordinates of the networkpoints,
- $\bar{x}_i, \bar{y}_i$ are the fixed given coordinates,
- $(x_1, y_1, \ldots, x_n, y_n)^t_s$ is the observation vector, and
- θ is the rotation angle to be estimated.

For the observation space $M = \mathbb{R}^{2n}$ we take the standard metric, i.e.

$$\langle \mathbf{a}_i, \mathbf{a}_j \rangle_M = \delta_{ij}, \tag{3.31}$$

with $\mathbf{a}_i$, $i = 1,\ldots.2n$ the standard basis.
It will be clear that the above model (3.30) determines a curve $\mathbf{c}(\theta)$ in the observation-space M. To solve for (3.30) we therefore need to project the observation vector $(x_1, y_1, \ldots, x_n, y_n)^t_s$ orthogonally onto $\mathbf{c}(\theta)$.
For illustrative purposes we will first derive expressions for the induced metric, the first curvature k_1 of $\mathbf{c}(\theta)$ and the convergence factor cf. of Gauss' iteration method as applied to (3.30). After this, we give the exact non-linear solution to (3.30). And finally we will give an alternative interpretation of model (3.30) by using the manifold structure of the group O(2) of orthogonal matrices of order 2.

Note that we can write model (3.30) in the form of

$$\tilde{\mathbf{y}} = (I^{\frac{1}{2}} \cos\theta)\mathbf{e}_1 + (I^{\frac{1}{2}} \sin\theta)\mathbf{e}_2 \ , \tag{3.32}$$

where:
$$\tilde{\mathbf{y}} = (x_1, y_1, \ldots, x_n, y_n)^t,$$

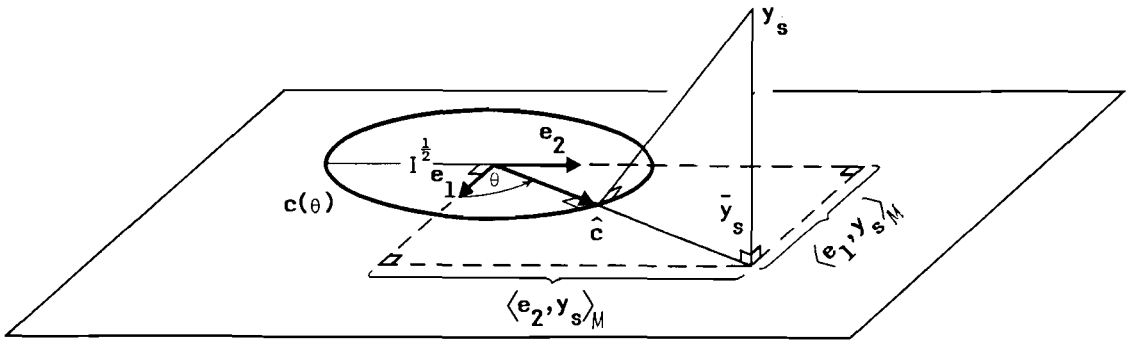
$$I = \sum_{i=1}^{n} (\bar{x}_i^2 + \bar{y}_i^2),$$

$$\mathbf{e}_1 = I^{-\frac{1}{2}} (\bar{x}_1, \bar{y}_1, \ldots, \bar{x}_n, \bar{y}_n)^t, \text{ and}$$

$$e_2 = I^{-\frac{1}{2}}(\bar{y}_1, -\bar{x}_1, \ldots, \bar{y}_n, -\bar{x}_n)^t,$$

with $\langle e_1, e_2 \rangle_M = 0, \quad \langle e_1, e_1 \rangle_M = \langle e_2, e_2 \rangle_M = 1$.

Hence our non-linear model (3.30) describes a circle which lies in the two-dimensional plane spanned by the orthonormal vectors e_1 and e_2 (see figure 26).



"Helmert transformation only admitting a rotation"

figure 26

The radius of this circle is given by the square root of I.
Thus we have immediately that

$$k_1 = I^{-\frac{1}{2}} = \left(\sum_{i=1}^{n} (\bar{x}_i^2 + \bar{y}_i^2) \right)^{-\frac{1}{2}} . \tag{3.33}$$

We also see at once that the arclength parameter s of $c(\theta)$ is given by

$$s = I^{\frac{1}{2}} \theta \quad ,$$

from which follows that the induced metric is constant along $c(\theta)$.
Hence, if by any chance the least-squares residual vector $y_s - \hat{c}$ is identical to zero, Gauss' iteration method as applied to (3.30) will have a third order convergence behaviour.
To compute the local linear convergence factor

$$cf. = |\langle k_1 N_1, \; y_s - c \rangle_{\hat{c}, M}| ,$$

of Gauss' iteration method as applied to (3.30), we need the length of the residual vector $y_s - \hat{c}$ projected onto N_1, the first normal of $c(\theta)$. Thus we need the length of the pseudo residual vector $\bar{y}_s - \hat{c}$, where $\bar{y}_s$ is the vector obtained by projecting y_s orthogonally onto the plane spanned by e_1 and e_2 (see figure 26). Hence,

$$\bar{y}_s = \langle e_1,\ y_s \rangle_M e_1 + \langle e_2,\ y_s \rangle_M e_2 \tag{3.34}$$

with $\quad \langle e_1, y_s \rangle_M = I^{-\frac{1}{2}} \sum_{i=1}^{p} (\bar{x}_i x_i + \bar{y}_i y_i), \quad \langle e_2, y_s \rangle_M = I^{-\frac{1}{2}} \sum_{i=1}^{p} (\bar{y}_i x_i - \bar{x}_i y_i)$.

Therefore

$$| \langle N_1, y_s - c \rangle_{\hat{c},M} | = || \bar{y}_s - \hat{c} ||_M = | \ || \bar{y}_s ||_M - || \hat{c} ||_M \ | =$$

$$= | \sqrt{\langle e_1, y_s \rangle_M^2 + \langle e_2, y_s \rangle_M^2} - I^{\frac{1}{2}} | .$$

With (3.33) follows then that

$$cf. = | 1 - \lambda |, \tag{3.35}$$

with $\quad \lambda = \sqrt{\dfrac{\langle e_1, y_s \rangle_M^2 + \langle e_2, y_s \rangle_M^2}{I}}$,

or

$$\lambda = \frac{\sqrt{\left(\sum_{i=1}^{p} (\bar{y}_i x_i - \bar{x}_i y_i) \right)^2 + \left(\sum_{i=1}^{p} (\bar{x}_i x_i + \bar{y}_i y_i) \right)^2}}{\sum_{i=1}^{p} (\bar{x}_i^2 + \bar{y}_i^2)} . \tag{3.35'}$$

Note that (3.35') is precisely the estimate of the scale parameter which one obtains when solving for the two dimensional Helmert transformation

$$\begin{aligned} \tilde{x}_i &= \bar{x}_i \lambda \cos\theta + \bar{y}_i \lambda \sin\theta \\ \tilde{y}_i &= -\bar{x}_i \lambda \sin\theta + \bar{y}_i \lambda \cos\theta, \end{aligned}$$

admitting a rotation **and** scale (see also (5.12)).

Of course the above discussion is only meant as illustration. In practice one will not solve model (3.30) by using an iteration method, since an exact non-linear solution is readily available. From figure 26 follows namely that

$$\tan \hat{\theta} = \frac{\langle e_2, y_s \rangle_M}{\langle e_1, y_s \rangle_M} .$$

Hence,

$$\hat{\theta} = \tan^{-1} \frac{\sum_{i=1}^{p} (\bar{y}_i x_i - \bar{x}_i y_i)}{\sum_{i=1}^{p} (\bar{x}_i x_i + \bar{y}_i y_i)} . \tag{3.36}$$

It will also be clear from the figure that solution (3.36) is a **global** minimum of the minimization problem

$$\min_{\mathbf{y} \in \mathbf{c}(\theta)} \langle \mathbf{y}_s - \mathbf{y}, \mathbf{y}_s - \mathbf{y} \rangle_M \ . \tag{3.37}$$

Except for the case $||\bar{\mathbf{y}}_s||_M = 0$. Then namely the solution is indefinite.

We will now give an alternative interpretation of the non-linear model (3.30). For the moment this alternative interpretation is only of theoretical interest. Observe that we can write model (3.30) in the form of

$$\begin{pmatrix} \tilde{x}_1 & \tilde{y}_1 \\ \vdots & \vdots \\ \tilde{x}_n & \tilde{y}_n \end{pmatrix} = \begin{pmatrix} \bar{x}_1 & \bar{y}_1 \\ \vdots & \vdots \\ \bar{x}_n & \bar{y}_n \end{pmatrix} \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}, \tag{3.38}$$

which we abbreviate as

$$\tilde{\mathbf{y}} = \mathbf{A}\,\mathbf{x}, \qquad \mathbf{x}\,\mathbf{x}^t = \mathbf{1}. \tag{3.39}$$

Thus $\tilde{\mathbf{y}}$ stands for the nx2 matrix on the left hand side of (3.38), **A** for the nx2 matrix on the right hand side and **x** for the 2x2 rotation matrix.

We will denote the linear vector space of nx2 real matrices by M(nx2), and the space of 2x2 orthogonal matrices by O(2):

It can be shown that O(n) is an $\frac{n(n-1)}{2}$ - dimensional manifold. Thus, with the usual abbreviations $M = M(nx2)$, $N = O(2)$, and $\bar{N} = \mathbf{A}O(2) \subset M$, we have that

$$\dim. M = 2n, \quad \dim. N = \dim. \bar{N} = 1, \tag{3.40}$$

and that **A** O(2) describes a curve in M.

To make our new formulation (3.39) compatible with (3.30) and the metric (3.31), we take for the metric tensor of $M = M(nx2)$ the following definition:

$$\langle . \ , \ . \rangle_M \overset{def.}{=} \text{trace}\,[(.)^t(.)] \ . \tag{3.41}$$

It is easily verified that $\langle . \ , \ . \rangle_M$ as given by (3.41) fulfils the necessary conditions of symmetry, bi-linearity and non-degeneracy.

With (3.39) and (3.41) we are now in the position of rephrasing our original least-squares problem (3.37) as

$$\min_{\mathbf{x} \in N = O(2)} \langle \mathbf{y}_s - \mathbf{A}\,\mathbf{x}, \mathbf{y}_s - \mathbf{A}\,\mathbf{x} \rangle_M = \min_{\mathbf{x} \in N = O(2)} \text{trace}[(\mathbf{y}_s - \mathbf{A}\,\mathbf{x})^t (\mathbf{y}_s - \mathbf{A}\,\mathbf{x})] .$$

And this is the formulation which we will use in our discussion of the three dimensional Helmert transformation (see subsection 5.5).

In the remaining four examples of this section we give some numerical results of some simple models to demonstrate the various effects mentioned of Gauss' iteration method. In all these examples we take the metric of M to be the standard metric.

Example 2: Orthogonal projection onto a unit circle

Our model reads as: $c^{i=1}(t) = \cos(t)$, $c^{i=2}(t) = \sin(t)$,
The observation point given is: $y_s^{i=1} = 0.5$, $y_s^{i=2} = 0.0$, and
our initial guess reads: $t_o = \frac{1}{4}\pi$ (rad.)
The numerical results are:

iteration step q	$c^{i=1}(t_q)$	$c^{i=2}(t_q)$	t_q
1	0.90822	0.41849	0.43178
2	0.97534	0.22070	0.22254
3	0.99371	0.11195	0.11218
4	0.99842	0.05618	0.05621
5	0.99960	0.02812	0.02812
6	0.99990	0.01406	0.01406
7	0.99997	0.00703	0.00703
8	0.99999	0.00352	0.00352
9	0.99999	0.00176	0.00176
10	0.99999	0.00088	0.00088
11	0.99999	0.00044	0.00044
12	1.00000	0.00022	0.00022
13	1.00000	0.00011	0.00011
14	1.00000	0.00005	0.00005
15	1.00000	0.00003	0.00003

table 1

Since the unitcircle has curvature $k_1 = 1$, we have with the observation point $y_s^{i=1} = 0.5$, $y_s^{i=2} = 0.0$ that $\langle k_1 \mathbf{N}_1, \mathbf{y}_s - \mathbf{c} \rangle_{\hat{\mathbf{c}},M} = 0.5$. And this local convergence factor is indeed clearly recognizable from the above given numerical results.

Example 3: Orthogonal projection onto a unitcircle

Again our model reads: $c^{i=1}(t) = \cos(t)$, $c^{i=2}(t) = \sin(t)$,
but this time we have as observation point: $y_s^{i=1} = 1.5$, $y_s^{i=2} = 0.0$,
our initial guess reads: $t_o = \frac{1}{4}\pi$ (rad.).

The numerical results are:

iteration step q	$c^{i=1}(t_q)$	$c^{i=2}(t_q)$	t_q
1	0.96235	-0.27180	-0.27526
2	0.99124	0.13205	0.13244
3	0.99785	-0.06560	-0.06564
4	0.99946	0.03274	0.03275
5	0.99987	-0.01637	-0.01637
6	0.99997	0.00818	0.00818

table 2

Again we have here a curvature $k_1=1$. In contrast with example 2, however, we have that $\langle k_1 \mathbf{N}_1, \mathbf{y}_s - \mathbf{c} \rangle_{\hat{\mathbf{c}},M} = -0.5$, which follows from the fact that the residual vector $\mathbf{y}_s - \hat{\mathbf{c}}$ has a direction opposite to that of $\mathbf{N}_1$. Thus, when compared to example 2, this third example reveals another feature, namely that when the observation point $\mathbf{y}_s$ and the centre of curvature are on opposite sides of the curve, the convergence factor will be negative. As a consequence the steplength of each iterationstep will be too long, resulting in an overshoot. Hence, the oscillatory behaviour of the above iteration.

In the previous example the obervation point $\mathbf{y}_s$ and centre of curvature were on the same side of the curve. And in that case the steplength will be too short (see figure 27). This effect is indeed clearly recognized from table 1 where the points in the sequence $t_1, t_2 \ldots$ approach $\hat{t}$ from one side.

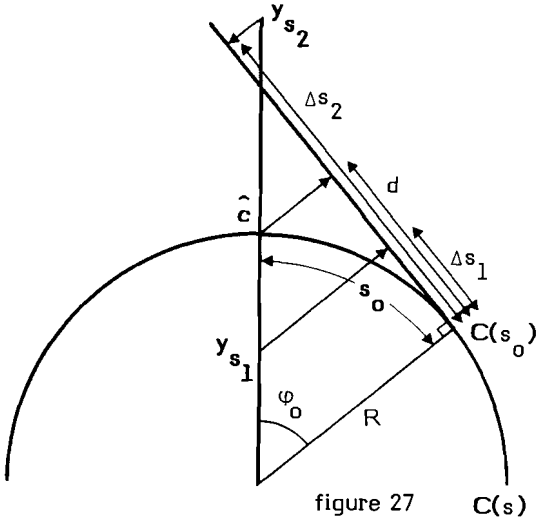


figure 27

$$\Delta s_1 < d = R \sin \frac{s_o}{R} < \Delta s_2$$

Example 4: Orthogonal projection onto a straight line

Our model reads as: $c^{i=1}(t) = e^{10t}$, $c^{i=2}(t) = e^{10t}$,
the observation point is given by: $y_s^{i=1} = 0$, $y_s^{i=2} = 2e$, and
the initial guess reads: $t_0 = 0$.

The numerical results are:

iteration step q	$c^{i=1}(t_q)$	$c^{i=2}(t_q)$	t_q
1	5.57494	5.57494	0.17183
2	3.33967	3.33967	0.12059
3	2.77267	2.77267	0.10198
4	2.71881	2.71881	0.10002
5	2.71828	2.71828	0.10000

table 3

Since the curve onto which the observation point is projected has no curvature, the local convergence behaviour of Gauss' iteration method as applied to the above model must be quadratic. In fact, with $\frac{1}{2}\,(s'(\hat{t}))^{-1}s''(\hat{t}) = 5$ for the above model, the local convergence rule of (3.29) is easily verified from table 3.

When viewing the last column of table 3 we also notice another interesting feature. We see that all iterates t_q except the initial guess t_o stay on the same side of the solution $\hat{t}$. The explanation is that the induced metric function, which for the above model reads $g(t) = 200\ e^{20t}$, is monotonic and increasing. With a monotonic and increasing metric function one will namely have an overshoot. In the above iteration this has the following effect. Since $t_o < \hat{t}$, we see that with the graph of g(t) we are going uphill. Hence, in the first iteration step we will have an overshoot. Thus $t_1 > \hat{t}$. But for the next step this means that with the graph of g(t) we are going downhill. Hence, for the second and succeeding steps we will have an undershoot, which explains why t_1, t_2... all approach $\hat{t}$ from the same side.

Example 5: Orthogonal projection onto a unitcircle with zero residualvector

Our model reads: $c^{i=1}(t) = \cos(t)$, $c^{i=2}(t) = \sin(t)$,
the observation point is given by: $y_s^{i=1} = 1.0$, $y_s^{i=1} = 0.0$, and
the initial guess reads: $t_o = \frac{1}{4}\pi$ (rad.).
The numerical results are:

iteration step q	$c^{i=1}(t_q)$	$c^{i=1}(t_q)$	t_q
1	0.99694	0.07821	0.07821
2	1.00000	0.00008	0.00008
3	1.00000	0.00000	0.00000

Although the unitcircle has a curvature of $k_1 = 1$, the observation point lies on the circle. Hence, we expect a local quadratic convergence behaviour governed by rule (3.29). However, a closer look at the above results reveals a third order behaviour instead of second order. The explanation is given by the fact that t equals the natural arclength parameter s of the unit circle. Thus $\frac{1}{2}(s'(t))^{-1}s''(t) = 0$.

3.7. Conclusions

In this section we have considered the univariate minimization problem of orthogonally projecting a given observation point $\mathbf{y}_s$ onto a smooth curve $\mathbf{c}$ in M. As a natural generalization of the linear least-squares problem we obtained Gauss' iteration method (3.7) which consists of successively solving a linear least-distance adjustment problem until the necessary condition of orthogonality,

$$\langle \mathbf{c}_*\left(\frac{d}{dt}\right),\ \mathbf{y}_s-\mathbf{c} \rangle_{\mathbf{c}(\hat{t}),M} = 0,$$

is fulfilled. At each iteration step q+1 the observation point $\mathbf{y}_s$ is orthogonally projected onto a new tangent space $T_{\mathbf{c}(t_{q+1})}(\mathbf{c}(N))$, which will be close to the previous one, $T_{\mathbf{c}(t_q)}(\mathbf{c}(N))$. Hence, the rate in which the tangential part of $\mathbf{y}_s-\mathbf{c}(t_q)$ decreases will depend on the rate of change of tangent spaces. And since curvature is defined as the measure of the rate of change of tangents, one can expect the local behaviour of Gauss' iteration method to depend on the curvature of curve $\mathbf{c}$. Through geometric reasoning we found that the local behaviour of Gauss' method is properly described by

$$t_{q+1} - \hat{t} = \langle k_1\mathbf{N}_1,\ \mathbf{y}_s-\mathbf{c} \rangle_{\hat{\mathbf{c}},M}(t_q-\hat{t}) + o((t_q-\hat{t})).$$

Hence, a necessary condition for convergence is

$$|\ \langle \mathbf{N}_1,\ \mathbf{y}_s-\mathbf{c} \rangle_{\hat{\mathbf{c}},M}| < k_1(\hat{t})^{-1},$$

and the rate of convergence is linear.
Moreover, it will be clear from the pictorial presentations given earlier that $\hat{\mathbf{c}}$ is a strict local minimum if

$$\langle \mathbf{N}_1,\ \mathbf{y}_s-\mathbf{c} \rangle_{\hat{\mathbf{c}},M} < k_1(\hat{t})^{-1}.$$

We also found that the local convergence behaviour of Gauss' method is invariant to any non-linear admissible parameter transformation.
The decisive factors which determine the local convergence rate are given by k_1 and $\mathbf{y}_s-\hat{\mathbf{c}}$. If either of them or both are equal to zero, then Gauss' method will have a local quadratic convergence behaviour:

$$t_{q+1} - \hat{t} = \tfrac{1}{2}\left[(s'(\hat{t}))^{-1}s''(\hat{t})\right](t_q-\hat{t})^2 + o((t_q-\hat{t})^2)\ .$$

Instead of solving the actual adjustment problem, one is then solving for the inverse mapping problem: given $\hat{\mathbf{c}}$ find the pre-image $\hat{t}$ under map $\mathbf{c}$: $t \in N = \mathbb{R} \to M$.
Consequently, the local quadratic convergence behaviour will not be invariant to a reparametrization.

In the next section we extend our results to the multivariate case. Can we expect the generalizations

to be simple and straightforward? In most cases yes, although there are two points which are worth mentioning. Firstly, when we consider manifolds other than curves, we must in some way take care of the increase in dimensions. And secondly, we must recognize that a surface in a three dimensional space is the simplest object having its own internal or intrinsic geometry. In our investigation of the space curve $\mathbf{c}(t)$ we were lead to the invariants of curvature. But these are invariants rather of the way the curve is situated in space, than internal to the curve. That is, they are **extrinsic** invariants. A curve has no intrinsic invariants, since essentially the only candidate for this status is the natural parameter of arclength s. But s is by itself inadequate for distinguishing the curve from, for instance, a straight line, i.e. we can coordinatize a straight line with the same parameter s in such a way that distances along both curve and straight line are measured in the same way. For surfaces and manifolds in general the situation is different. It is impossible, for instance, to coordinatize the sphere so that the formula for distance on the sphere in terms of these coordinates, is the same as the usual distance formula in the ambient space. A consequence is that where in the univariate case the possible local quadratic convergence behaviour of Gauss' method could be reduced to a third order behaviour by taking the arclength s as parameter, this will not be possible in the multivariate case.

4. Orthogonal projection onto a parametrized submanifold

4.1. Gauss' method

In this section we will consider Gauss' method for the multivariate case of non-linear least-squares adjustment. Thus we assume dim. $N = n \geq 1$. Furthermore we assume that the imbedding of the n-dimensional manifold N into the m-dimensional space M is established by the injective nonlinear map $\mathbf{y}$, i.e. $\mathbf{y}: N \to M$.

When we speak of the metric of N we mean as before the induced metric, i.e. the metric obtained by pulling the metric of M back to N:

$$\langle X,Y \rangle_N = \langle \mathbf{y}_*(X), \mathbf{y}_*(Y) \rangle_M \quad \text{for any vector fields } X,Y \text{ on } N.$$

Now, consider again the least-squares minimization problem

$$\min_{\mathbf{y} \in \bar{N} = \mathbf{y}(N)} \langle \mathbf{y}_s - \mathbf{y}, \mathbf{y}_s - \mathbf{y} \rangle_M = \langle \mathbf{y}_s - \hat{\mathbf{y}}, \mathbf{y}_s - \hat{\mathbf{y}} \rangle_M , \tag{4.1}$$

For $\hat{\mathbf{y}}$ to be a solution to (4.1) we have as necessary condition that the residual vector $\mathbf{y}_s - \hat{\mathbf{y}}$ must be orthogonal to the tangent space $T_{\hat{\mathbf{y}}}\bar{N}$ of $\bar{N}$ at $\hat{\mathbf{y}}$, i.e. we have that

$$\langle \mathbf{y}_*(\partial_\alpha), \mathbf{y}_s - \mathbf{y} \rangle_M = 0, \tag{4.2}$$

must hold at $\hat{\mathbf{y}} \in \bar{N}$.

Due, however, to the assumed nonlinearity of map $\mathbf{y}$, the tangent space $T_{\hat{\mathbf{y}}}\bar{N}$ is generally unknown a

priori. Hence, our adjustment problem cannot be solved directly in general. But as in the previous section, (4.2) suggests that we take as a first approximation the orthogonal projection of y_s onto a chosen nearby tangent space $T_{y_q}\bar{N}$ of $\bar{N}$ at $y_q = y(x_q)$. Of course then

$$\langle y_*(\partial_\alpha), y_s - y\rangle_{y_q, M} \neq 0. \tag{4.3}$$

But by pulling the non-orthogonality as measured by (4.3) back to the Riemannian manifold N, we get

$$\langle \partial_\alpha, \Delta x_q\rangle_{x_q N} = \langle y_*(\partial_\alpha), y_s - y\rangle_{y_q M} \quad , \quad \text{with } \Delta x_q \in T_{x_q} N . \tag{4.4}$$

And in local coordinates this expression suggests **Gauss' iteration method:**

$$\begin{aligned} \Delta x^\beta_q &= g^{\beta\alpha}(x_q)\partial_\alpha y^i(x_q) g_{ij}(y^j_s - y^j(x_q)) \\ x^\beta_{q+1} &= x^\beta_q + \Delta x^\beta_q, \qquad i,j=1,\dots,m;\ \alpha,\beta=1,\dots,n \end{aligned} \tag{4.5}$$

This scheme is thus the multivariate generalization of (3.7), and it consists of successively solving a linear least-distance adjustment problem until condition (4.2) is met.

In order to understand the local behaviour of Gauss' method we shall now proceed in a way similar to that of the previous section. One of the problems, however, we have to deal with is the increase in dimensions. Nevertheless, the **linearity** of the local rate of convergence of Gauss' method (4.5) is easily shown. From Taylorizing

$$\langle y_*(g^{\beta\alpha}\partial_\alpha), y_s - y\rangle_{y_q M}$$

about the least-squares solution follows namely

$$\begin{aligned} x^\beta_{q+1} - x^\beta_q &= \langle y_*(g^{\beta\alpha}\partial_\alpha), y_s - y\rangle_{\hat{y}M} + D_{y_*(\partial_\gamma)}\langle y_*(g^{\beta\alpha}\partial_\alpha), y_s - y\rangle_{\hat{y}M}(x^\gamma_q - \hat{x}^\gamma) + O(\|x_q - \hat{x}\|^2) \\ &= \langle y_*(g^{\beta\alpha}\partial_\alpha), y_s - y\rangle_{\hat{y}M} + \langle D_{y_*(\partial_\gamma)} y_*(g^{\beta\alpha}\partial_\alpha), y_s - y\rangle_{\hat{y}M}(x^\gamma_q - \hat{x}^\gamma) + \\ &\quad + \langle y_*(g^{\beta\alpha}\partial_\alpha), D_{y_*(\partial_\gamma)}(y_s - y)\rangle_{\hat{y}M}(x^\gamma_q - \hat{x}^\gamma) + O(\|x_q - \hat{x}\|^2). \end{aligned}$$

And since

$$\langle y_*(g^{\beta\alpha}\partial_\alpha), y_s - y\rangle_{\hat{y}M} = 0 \quad \text{and} \quad D_{y_*(\partial_\gamma)}(y_s - y) = -y_*(\partial_\gamma),$$

we get

$$x^\beta_{q+1} - \hat{x}^\beta = \langle D_{y_*(\partial_\gamma)} y_*(g^{\beta\alpha}\partial_\alpha), y_s - y\rangle_{\hat{y}M}(x^\gamma_q - \hat{x}^\gamma) + O(\|x_q - \hat{x}\|^2), \tag{4.6}$$

which proves our statement. Thus, for points close enough to the solution the coordinate-differences of the current point $\mathbf{x}_{q+1}$ and the solution $\hat{\mathbf{x}}$ depend linearly on the coordinate differences of the previous point $\mathbf{x}_q$ and $\hat{\mathbf{x}}$.
Upon comparing (4.6) with our univariate result (3.26) we see that we still lack a proper geometric interpretation of the convergence factor of Gauss' method (4.5) although we can expect that in some way the curvature behaviour of the submanifold $\bar{N}$ at $\hat{\mathbf{y}}$ will be involved. To make this statement precise it seems appropriate that we look for the multivariate analogon of

4.2. The Gauss' equation

as given in (3.23).
To do so, we first recall that the connection $\mathbf{D}$ of M satisfies

$$\mathbf{D}_{f\mathbf{V}}g\mathbf{W} = f\mathbf{V}(g)\mathbf{W} + fg\mathbf{D}_{\mathbf{V}}\mathbf{W}, \tag{4.7}$$

for all smooth functions $f, g: M \to \mathbb{R}$ and vector fields $\mathbf{V},\mathbf{W}$ on M; that it is torsionfree, i.e.

$$\mathbf{D}_{\mathbf{V}}\mathbf{W} - \mathbf{D}_{\mathbf{W}}\mathbf{V} = \mathbf{V}\,\mathbf{W} - \mathbf{W}\,\mathbf{V}, \tag{4.8}$$

for all vector fields $\mathbf{V},\mathbf{W}$ on M; and that it is metric, i.e.

$$\mathbf{D}_{\mathbf{U}}\langle \mathbf{V},\mathbf{W}\rangle_M = \langle \mathbf{D}_{\mathbf{U}}\mathbf{V},\mathbf{W}\rangle_M + \langle \mathbf{V},\mathbf{D}_{\mathbf{U}}\mathbf{W}\rangle_M, \tag{4.9}$$

for all vector fields $\mathbf{U},\mathbf{V},\mathbf{W}$ on M.
We say that a vector field $\mathbf{U}$ on M is an **extension** of a vector field $\mathbf{Z}$ on N, if $\mathbf{U}$ restricted to $\bar{N}$ equals the pushforward of $\mathbf{Z}$ on N, i.e.

$$\mathbf{U}_{|\bar{N}} = \mathbf{y}_*(\mathbf{Z}), \tag{4.10}$$

or in components

$$U^i_{|\bar{N}} = \partial_\alpha y^i Z^\alpha \quad .$$

Now, let $\mathbf{X},\mathbf{Y}$ and $\mathbf{Z}$ be three vector fields on N and let $\mathbf{V}$, $\mathbf{W}$ and $\mathbf{U}$ be their extensions. As in (3.23), we then decompose $\mathbf{D}_{\mathbf{V}}\mathbf{W}$ restricted to $\bar{N}$, into a tangential and normal part:

$$\begin{aligned} \mathbf{D}_{\mathbf{V}}\mathbf{W}_{|\bar{N}} &= \mathbf{Tang.}(\mathbf{D}_{\mathbf{V}}\mathbf{W}_{|\bar{N}}) + \mathbf{Norm.}\ (\mathbf{D}_{\mathbf{V}}\mathbf{W}_{|\bar{N}}) \\ &\overset{say}{=} \mathbf{y}_*(\nabla_X Y) + \mathbf{B}(X,Y) \quad . \end{aligned} \tag{4.11}$$

With the connection properties (4.7), (4.8) and (4.9) of $\mathbf{D}$ we can then derive the following properties for ∇ and the normalfield $\mathbf{B}$ (see e.g. Spivak, 1975):

(i) Let f and g be smooth functions on M and denote their pullbacks by $\bar{f}$ and $\bar{g}$ respectively, i.e. $\bar{f} = f \circ y$, $\bar{g} = g \circ y$. From (4.7), (4.10) and (4.11) follows then that

$$\begin{aligned} D_{fV}gW|_{\bar{N}} &\overset{(4.7)}{=} \{ fV(g)W + fgD_V W \}|_{\bar{N}} \\ &\overset{(4.10)}{=} \bar{f}X(\bar{g})y_*(Y) + \bar{f}\bar{g}\, D_V W|_{\bar{N}} \\ &\overset{(4.11)}{=} y_*(\bar{f}X(\bar{g})Y + \bar{f}\bar{g}\, \nabla_X Y) + \bar{f}\bar{g}\, B(X,Y) \end{aligned}$$

or

$$y_*(\nabla_{\bar{f}X}\bar{g}Y) + B(\bar{f}X,\bar{g}Y) = y_*(\bar{f}X(\bar{g})Y + \bar{f}\bar{g}\, \nabla_X Y) + \bar{f}\bar{g}\, B(X,Y) \ .$$

Hence,

$$\nabla_{\bar{f}X}\bar{g}Y = \bar{f}X(\bar{g})Y + \bar{f}\bar{g}\, \nabla_X Y \quad \text{and} \quad B(\bar{f}X,\bar{g}Y) = \bar{f}\bar{g}\, B(X,Y). \tag{4.12}$$

Since additivity is trivial to prove, these two equations show that ∇ defines an affine connection on N and that B is bilinear in its arguments.

(ii) From (4.8) follows that

$$(D_V W - D_W V)|_{\bar{N}} = (VW-WV)|_{\bar{N}} = (V^i \partial_i W^j - W^i \partial_i V^j)\, \partial_j|_{\bar{N}} \ .$$

And with (4.10) this gives

$$\begin{aligned} (D_V W - D_W V)|_{\bar{N}} &= (X^\beta \partial_\beta (\partial_\alpha y^j Y^\alpha) - Y^\beta \partial_\beta (\partial_\alpha y^j X^\alpha))\, \partial_j \\ &= \partial_\alpha y^j (X^\beta \partial_\beta Y^\alpha - Y^\beta \partial_\beta X^\alpha)\, \partial_j \\ &= y_*(XY-YX) \ . \end{aligned}$$

Hence, with (4.11) we have

$$y_*(\nabla_X Y) + B(X,Y) - y_*(\nabla_Y X) - B(Y,X) = y_*(XY-YX) \ ,$$

or

$$\nabla_X Y - \nabla_Y X = XY - YX \quad \text{and} \quad B(X,Y) = B(Y,X) \ . \tag{4.13}$$

But this shows that the torsionfreeness of D implies that ∇ is torsionfree and that B is symmetric in its arguments.

(iii) From (4.9), (4.10) and (4.11) follows that

$$\begin{aligned} D_U\langle V,W\rangle|_{\bar{N}} &= (\langle D_U V, W\rangle + \langle V, D_U W\rangle)|_{\bar{N}} \\ &= \langle y_*(\nabla_Z X), y_*(Y)\rangle_M + \langle y_*(X), y_*(\nabla_Z Y)\rangle_M \\ &= \langle \nabla_Z X, Y\rangle_N + \langle X, \nabla_Z Y\rangle_N \end{aligned}$$

And since

$$D_U\langle V,W\rangle_{|\bar{N}} = U\langle V,W\rangle_{|\bar{N}} = Z\langle X,Y\rangle_N = \nabla_Z\langle X,Y\rangle_N \ ,$$

it follows that also ∇ is metric, i.e.

$$\nabla_Z\langle X,Y\rangle_N = \langle \nabla_Z X,Y\rangle_N + \langle X,\nabla_Z Y\rangle_N . \tag{4.14}$$

Concluding, (4.12), (4.13) and (4.14) taken together show that ∇ is an affine, torsionfree and metric connection of N and that the normalfield **B** is bilinear and symmetric in its arguments. Hence, ∇ is the unique Riemannian connection (also known as the induced or Levi-Civita connection) of N which is completely described by the induced metric $\langle .,. \rangle_N$.
Those familiar with Gaussian surface theory will probably recognize the connection ∇ more easily if we show how the connection coefficients $\Gamma^{\gamma}_{\alpha\beta}$, defined by

$$\nabla_{\partial_\alpha}\partial_\beta = \Gamma^{\gamma}_{\alpha\beta}\,\partial_\gamma, \quad \alpha,\beta,\gamma = 1,\ldots,n, \tag{4.15}$$

can be computed from the coefficients of the induced metric tensor $\langle .,. \rangle_N$. Since we assume that partial derivatives commute, it follows from (4.13) for $X = \partial_\alpha$, $Y = \partial_\beta$, that

$$\nabla_{\partial_\alpha}\partial_\beta = \nabla_{\partial_\beta}\partial_\alpha \quad \text{or} \quad \Gamma^{\gamma}_{\alpha\beta} = \Gamma^{\gamma}_{\beta\alpha} . \tag{4.16}$$

With $X = \partial_\alpha$, $Y = \partial_\beta$, $Z = \partial_\gamma$ and (4.15), we can write (4.14) as

$$\partial_\gamma g_{\alpha\beta} = \Gamma^{\delta}_{\gamma\alpha} g_{\delta\beta} + \Gamma^{\delta}_{\gamma\beta} g_{\delta\alpha} .$$

Cyclically permuting the indices gives then three equations which, with (4.16) show that

$$\Gamma^{\gamma}_{\alpha\beta} = \frac{1}{2}\, g^{\gamma\delta}\left\{\partial_\alpha g_{\beta\delta} + \partial_\beta g_{\delta\alpha} - \partial_\delta g_{\alpha\beta}\right\} . \tag{4.17}$$

This is Christoffel's second identity well known from surface theory.

The decomposition formula

$$\boxed{D_V W_{|\bar{N}} = y_*(\nabla_X Y) + B(X,Y)} \ , \tag{4.18}$$

which brought the above derived properties of ∇ and **B** about is known as **Gauss' equation.**
Its complementary counterpart, i.e. **Weingarten's equation,** is obtained from applying D_V to a normalfield, **N** say, on $\bar{N}$, followed by an orthogonal decomposition:

$$D_V N_{|\bar{N}} = - y_*(K_N(X)) + D^{\perp}_V N . \tag{4.19}$$

And with a similar derivation as used above one can show that $K_N(X)$ is bilinear in **N** and **X**, and that

$D^{\perp}$ is a metric connection for the normalbundle $T^{\perp}\bar{N}$ of $\bar{N}$ in M (see e.g. Spivak, 1975).

We shall now show how equation (4.18) for $V = y_*(\partial_\alpha)$, $W = y_*(\partial_\beta)$, i.e.

$$D_{y_*(\partial_\alpha)} y_*(\partial_\beta) = y_*(\nabla_{\partial_\alpha} \partial_\beta) + B(\partial_\alpha, \partial_\beta), \tag{4.20}$$

specializes to the first equation of (3.23) if we assume n=1, and replace y by c and ∂_α by $\frac{d}{dt}$. With these assumptions, (4.20) becomes

$$D_{c_*(\frac{d}{dt})} c_*(\frac{d}{dt}) = c_*(\nabla_{\frac{d}{dt}} \frac{d}{dt}) + B_c(\frac{d}{dt}, \frac{d}{dt}) \quad . \tag{4.21}$$

(We have given B_c the subindex "c" to emphasize that the normalfield B_c belongs to the spacecurve c viewed as a one dimensional manifold).

With (4.16) and (4.17) it follows that the first term of (4.21) can be written as

$$c_*(\nabla_{\frac{d}{dt}} \frac{d}{dt}) = c_*(\frac{1}{2}\, g(t)^{-1} \frac{dg}{dt}(t)\, \frac{d}{dt}) \quad . \tag{4.22}$$

For the second term of (4.21) it follows from

$$0 = D_V \langle V, N \rangle_M = \langle D_V V, N \rangle_M + \langle V, D_V N \rangle_M \ ,$$

and (4.19), (4.21) and $V = c_*(\frac{d}{dt})$ that

$$\langle B_c(\frac{d}{dt}, \frac{d}{dt}), N \rangle_M = \langle c_*(\frac{d}{dt}), c_*(K_N(\frac{d}{dt})) \rangle_M \ .$$

Hence, if we put $B_c(\frac{d}{dt}, \frac{d}{dt}) = B_c^1(t) N_1$, $N = N_1$ where N_1 is a unitnormal, and $K_{N_1}(\frac{d}{dt}) = k_1(t)\, \frac{d}{dt}$ we get for the second term of (4.21) that

$$B_c(\frac{d}{dt}, \frac{d}{dt}) = g(t)\, k_1(t) N_1 \ . \tag{4.23}$$

Thus, with (4.22) and (4.23), (4.21) can be written as

$$D_{c_*(\frac{d}{dt})} c_*(\frac{d}{dt}) = c_*(\frac{1}{2}\, g^{-1}(t) \frac{dg}{dt}\,(t)\, \frac{d}{dt}) + g(t) k_1(t) N_1$$

or as

$$D_V V = (s'(t))^{-1} s''(t) V + (s'(t))^2 k_1(t) N_1, \tag{4.24}$$

since $g(t) = (s'(t))^2$ and $V = c_*(\frac{d}{dt})$ .And (4.24) is indeed the equation which we already derived in (3.23).

Note from comparing (4.20) and (4.24) that

$(s'(t))^{-1}s''(t)\,\frac{d}{dt}$ generalizes to $\Gamma^{\gamma}_{\alpha\beta}\partial_{\gamma}$,

and

$(s'(t))^{2}k_1(t)N_1$ generalizes to $B(\partial_{\alpha},\partial_{\beta})$.

Hence, we can expect that the curvature behaviour of submanifold $\tilde{N}$ is contained in the normalfield **B**. Let us therefore study

4.3. The normalfield B

in more detail.

According to (4.23), the first curvature $k_1(t)$ of a curve $c:\mathbb{R} \to M$ can be obtained from the normalfield B_c through

$$\left\langle B_c\left(\frac{d}{dt}, \frac{d}{dt}\right), N_1\right\rangle_M = k_1(t)\left\langle \frac{d}{dt}, \frac{d}{dt}\right\rangle . \tag{4.25}$$

Now in order to find the proper multivariate generalization of this expression, one of the problems we have to deal with is the increase in dimensions. We can, however, get round this difficulty if we consider two curves, one in N, which we denote by $c_1: \mathbb{R} \to N$, and one in $\tilde{N} \subset M$, which we denote by $c_2: \mathbb{R} \to \tilde{N} \subset M$. And furthermore we assume that $c_2 = y \circ c_1$. Thus we have the following situation

$$t \in \mathbb{R} \xrightarrow{c_1} N \xrightarrow{y} \tilde{N} \subset M, \qquad t \in \mathbb{R} \xrightarrow{c_2} \tilde{N} \subset M$$

With the connections **D** and ∇ of M and N respectively, we can then apply the univariate Gauss' decomposition formula twice. Namely to curve c_1 and to curve c_2. With

$$V = c_{2*}\left(\frac{d}{dt}\right) \text{ and } X = c_{1*}\left(\frac{d}{dt}\right)$$

this gives

$$D_V V = c_{2*}\left((s'(t))^{-1}s''(t)\,\frac{d}{dt}\right) + (s'(t))^{2}k_{2,1}(t)N_{2,1}, \tag{4.26.a}$$

and

$$\nabla_X X = c_{1*}\left((s'(t))^{-1}s''(t)\,\frac{d}{dt}\right) + (s'(t))^{2}k_{1,1}(t)N_{1,1}, \tag{4.26.b}$$

where $k_{2,1}$ and $N_{2,1}$ are the first curvature and first normal of curve c_2 in M, and $k_{1,1}$ and $N_{1,1}$ are the first curvature and first normal of curve c_1 in N.

Note that the arclength parameter s is equal for both curves since $c_2 = y \circ c_1$. Hence $V = y_*(X)$. Application of the multivariate Gauss' decomposition formula gives then

$$D_V V = y_*(\nabla_X X) + B(X,X) .$$

And substitution of (4.26.b) gives

$$D_V V = y_*(c_{1*}((s'(t))^{-1}s''(t)\frac{d}{dt})) + y_*(s'(t))^2 k_{1,1}(t)N_{1,1}) + B(X,X)$$

or

$$D_V V = c_{2*}((s'(t))^{-1}s''(t)\frac{d}{dt}) + (s'(t))^2 k_{1,1}(t)y_*(N_{1,1}) + B(X,X) \ . \qquad (4.27)$$

From comparing (4.27) with (4.26.a) follows then that

$$(s'(t))^2 k_{2,1}(t)N_{2,1} = (s'(t))^2 k_{1,1}(t)y_*(N_{1,1}) + B(X,X) \ .$$

Hence, for curve c_2 which lies entirely in $\bar{N} \subset M$, the normalfield **B** equals the orthogonal component of $(s'(t))^2 k_{2,1}(t)N_{2,1}$. Thus for an arbitrary unit normal $N \in T_y^{\perp}\bar{N}$ we have

$$\langle B(X,X),N\rangle_M = \langle k_{2,1}(t)N_{2,1},N\rangle_M \langle X,X\rangle_N \ , \qquad (4.28)$$

since $(s'(t))^2 = \langle X,X\rangle_N$.

We call $\langle k_{2,1}(t)N_{2,1},N\rangle_M$ the **extrinsic curvature** of curve c_2 with respect to the unitnormal **N** and denote it by $k_N(t)$ (the first curvature $k_{1,1}(t)$ of curve c_1 in N is sometimes called the intrinsic or geodesic curvature). We can now write (4.28) as

$$\boxed{\langle B(X,X),N\rangle_M = k_N\langle X,X\rangle_N \ , \text{ with } N \in T_y^{\perp}\bar{N} \ , \ X \in T_x N} \qquad (4.29)$$

and this expression can considered to be the proper generalization of (4.25). As a consequence of the increase in dimensions we thus see that to every combination of a tangent vector **X** and a normalvector **N**, there belongs an extrinsic curvature k_N.

Tangent directions for which the extrinsic curvature k_N attains extreme values are called **principal directions** with respect to the unitnormal **N**. And the corresponding extrinsic curvatures are called **principal curvatures** with respect to **N**. Thus in order to find the principal directions- and curvatures for a chosen unit normal **N**, we need the extreme values of the ratio

$$\frac{\langle B(X,X),N\rangle_M}{\langle X,X\rangle_N} \ , \ \forall \ X = X^\alpha \partial_\alpha \in T_x N \ .$$

Recall from linear algebra that this problem reduces to the eigenvalue problem

$$\langle B(X,Y),N\rangle_M = \lambda_N\langle X,Y\rangle_N, \ \forall \ Y \in T_x N \ . \qquad (4.30)$$

The eigenvectors **X** determine then the mutually orthogonal principal directions and the corresponding eigenvalues the principal curvatures.

We will denote the n principal curvatures for the normal directions **N** by k_N^r, r=1,...,n, and assume that

$$k_N^1 \geq k_N^2 \geq \cdots \geq k_N^n . \tag{4.31}$$

The corresponding mutually orthogonal principal directions are denoted by X_r, r=1,...,n.
For later reference we define the **mean curvature** $\bar{k}_N$ of submanifold $\bar{N} = \mathbf{y}(N)$ for the normal direction **N** as the average trace of **B:**

$$\bar{k}_N = \frac{1}{n} \langle g^{\alpha\beta} B(\partial_\alpha, \partial_\beta), N \rangle_M = \frac{1}{n} \sum_{r=1}^{n} k_N^r , \tag{4.32}$$

and the unique **mean curvature normal** $\bar{\mathbf{N}}$ of $\bar{N}$ as

$$\bar{\mathbf{N}} = \frac{1}{n} \langle g^{\alpha\beta} B(\partial_\alpha, \partial_\beta), N_p \rangle_M N_p = \bar{k}_{N_p} N_p , \tag{4.33}$$

where N_p, p=1,...,(m-n) is an orthonormal basis of $T_{\hat{y}}^{\perp} \bar{N}$.

Now that we have found the geometric interpretation associated with the normalfield **B**, let us return to our nonlinear least-squares adjustment problem and apply our results to obtain a geometric interpretation of

4.4. The local rate of convergence

of Gauss' iteration method.
Recall from (4.6) that

$$x_{q+1}^{\beta} - \hat{x}^{\beta} = \langle D_{y_*(\partial_\gamma)} y_*(g^{\beta\alpha}\partial_\alpha),\ y_s - y \rangle_{\hat{y},M} (x_q^{\gamma} - \hat{x}^{\gamma}) + O(||x_q - \hat{x}||^2).$$

But Gauss' decomposition formula states that

$$D_{y_*(\partial_\gamma)} y_*(g^{\beta\alpha}\partial_\alpha) = y_*(\nabla_{\partial_\gamma} g^{\beta\alpha}\partial_\alpha) + B(g^{\beta\alpha}\partial_\alpha, \partial_\gamma) .$$

Hence,

$$x_{q+1}^{\beta} - \hat{x}^{\beta} = \langle g^{\beta\alpha} B(\partial_\alpha, \partial_\gamma), y_s - y \rangle_{\hat{y},M} (x_q^{\gamma} - \hat{x}^{\gamma}) + O(||x_q - \hat{x}||^2) , \tag{4.34}$$

since $y_s - \hat{y} \in T_{\hat{y}}^{\perp} \bar{N}$.
Thus we see that indeed the extrinsic curvatures of submanifold $\bar{N}$ at $\hat{y}$ with respect to the normal direction $y_s - \hat{y}$ govern the local convergence factor of Gauss' method. We can rewrite (4.34) in a form which better resembles our univariate result (3.25) if we make use of the eigenvalue problem (4.30). Assume therefore that X_r, r = 1,...,n, forms an orthonormal basis of principal directions in $T_x N$. Then

$$\langle B(\partial_\alpha, \partial_\beta), N \rangle_{\hat{y},M} X_r^{\beta} = k_N^r g_{\alpha\beta} X_r^{\beta} \quad \text{(no summation over r).}$$

With

$$x^{\beta}_{q+1} - \xi^{\beta} = u^{r}_{q+1} X^{\beta}_{r}, \qquad x^{\beta}_{q} - \xi^{\beta} = u^{r}_{q} X^{\beta}_{r}$$

and

$$\mathbf{N} = (\mathbf{y}_s - \hat{\mathbf{y}}) / || \mathbf{y}_s - \hat{\mathbf{y}} ||_M ,$$

expression (4.34) can then be written as

$$u^{r}_{q+1} = k^{r}_{\mathbf{N}} || \mathbf{y}_s - \hat{\mathbf{y}} ||_M \; u^{r}_{q} + O(u^{t}_{q} \delta_{ts} u^{s}_{q}).$$

Hence, we have

$$\boxed{u^{r}_{q+1} = \langle k^{r}_{\mathbf{N}} \mathbf{N}, \mathbf{y}_s - \mathbf{y} \rangle_{\hat{\mathbf{y}} M} \; u^{r}_{q} + O(u^{t}_{q} \delta_{ts} u^{s}_{q}) \quad r=1,\ldots,n \quad \text{(no summation over r).}} \tag{4.35}$$

Compare this with our univariate result (3.26).
With (4.35) we are now able to generalize some of our conclusions of section three:

(i) If

$$| \langle k^{r}_{\mathbf{N}} \mathbf{N}, \; \mathbf{y}_s - \mathbf{y} \rangle_{\hat{\mathbf{y}} M} | < 1, \qquad r=1,\ldots,n, \tag{4.36}$$

and

$\mathbf{x}_0$ is sufficiently close to $\boldsymbol{\xi}$,

then the sequence $\{\mathbf{x}_q\}$ generated by Gauss' method converges to $\boldsymbol{\xi}$.

(ii) The local convergence behaviour of Gauss' method is determined by the combined effect of the curvature behaviour of submanifold $\bar{N}$ at $\hat{\mathbf{y}}$, and the residual vector $\mathbf{y}_s - \hat{\mathbf{y}}$.

(iii) Since the extrinsic curvatures are a property of the submanifold $\bar{N}$ itself, the local convergence behaviour of Gauss' method is invariant to any admissible parameter transformation. Hence, we cannot expect to speed up convergence in general by choosing a particular parametrization.

(iv) Gauss' method has a local linear rate of convergence. From (4.35) follows that

$$\lim_{q \to \infty} \frac{(x^{\alpha}_{q+1} - \xi^{\alpha}) g_{\alpha\beta}(\xi) (x^{\beta}_{q+1} - \xi^{\beta})}{(x^{\alpha}_{q} - \xi^{\alpha}) g_{\alpha\beta}(\xi) (x^{\beta}_{q} - \xi^{\beta})} \leq \left(\max. \{ \, | k^{1}_{\mathbf{N}} | \; || \mathbf{y}_s - \hat{\mathbf{y}} ||_M , \, | k^{n}_{\mathbf{N}} | \; || \mathbf{y}_s - \hat{\mathbf{y}} ||_M \} \right)^2 .$$

(4.37.a)

Hence, the local convergence factor (lcf.) of Gauss' method reads

$$\text{lcf.} = \max. \{ | k^{1}_{\mathbf{N}} | \; || \mathbf{y}_s - \hat{\mathbf{y}} ||_M \; , \; | k^{n}_{\mathbf{N}} | \; || \mathbf{y}_s - \hat{\mathbf{y}} ||_M \} \; . \tag{4.37.b}$$

Note that since $\langle B(\partial_\alpha,\partial_\beta),N\rangle_M$ need not be positive definite, the extrinsic curvatures can either be positive, zero or negative. But they are always real, since B is symmetric in its arguments.

(v) From the geometry of our non-linear least-squares problem follows that the solution $\hat{\mathbf{y}}$ is a strict local minimum if $1 - k_N^1 \, ||\mathbf{y}_s - \hat{\mathbf{y}}||_M > 0$. The fact that $\hat{\mathbf{y}}$ is a strict local minimum does however not ensure local convergence of Gauss' method. See (4.36).

(vi) If $k_N^r < 0$, then the observation point $\mathbf{y}_s$ and the with k_N^r corresponding centre of curvature lie on opposite sides of the submanifold $\bar{N}$. Consequently, one will overshoot the target $\hat{\mathbf{x}}$ along the principal direction $\mathbf{X}_r$ in each iteration step if $k_N^r < 0$. Hence, the iteration will then show an oscillatory behaviour along the direction $\mathbf{X}_r$.
Similarly, one will have an undershoot along the direction $\mathbf{X}_r$ if $k_N^r > 0$ (see also example 3 of the previous section).

An interesting point of the above conclusion (vi) is that it indicates the possibility of adjusting the steplength in each iteration step with the aid of the curvature behaviour of $\bar{N}$, so as to improve the convergence behaviour (4.35) of Gauss' method. Let us therefore pursue this argument a bit further. Instead of (4.5) we take

$$x_{q+1}^\alpha = x_q^\alpha + t_q^\alpha \, \Delta x_q^\alpha \,, \tag{4.38}$$

where $\Delta\mathbf{x}_q$ is provided by Gauss' method and t_q is a positive scalar, chosen so as to adjust the steplength. Instead of (4.35) one would then get

$$u_{q+1}^r = \left[(1-t_q) + \langle k_N^r \mathbf{N}, \mathbf{y}_s - \mathbf{y}\rangle_{\hat{\mathbf{y}}M} \, t_q\right] u_q^r + O(u_q^t \delta_{ts} u_q^s), \tag{4.39}$$

$r = 1,\ldots,n$; no summation over r.

As could be expected, it follows from (4.39) that the scalar t_q should be chosen less than one if all extrinsic curvatures are negative, and greater than one if all extrinsic curvatures are positive. Now let us investigate what the optimal choice of t_q would be. Since the in absolute value largest coefficient of u_q^r, $r=1,\ldots,n$, in (4.39) is given by

$$\max.\{|(1-t_q) + \langle k_N^1 \mathbf{N}, \mathbf{y}_s - \mathbf{y}\rangle_{\hat{\mathbf{y}}M} \, t_q \,|, |(1-t_q) + \langle k_N^n \mathbf{N}, \mathbf{y}_s - \mathbf{y}\rangle_{\hat{\mathbf{y}}M} \, t_q \,|\},$$

it follows that the optimal choice of t_q is given by the solution of

$$\min_{t_q>0}.\left(\max.\{|(1-t_q) + \langle k_N^1 \mathbf{N}, \mathbf{y}_s - \mathbf{y}\rangle_{\hat{\mathbf{y}}M} \, t_q \,|, |(1-t_q) + \langle k_N^n \mathbf{N}, \mathbf{y}_s - \mathbf{y}\rangle_{\hat{\mathbf{y}}M} \, t_q \,|\}\right).$$

From figure 28 follows then, that if $\hat{\mathbf{y}}$ is a strict local minimum, the optimal choice for t_q is:

$$t_q = \frac{2}{2 - \langle (k_N^1 + k_N^n)\mathbf{N}, \mathbf{y}_s - \mathbf{y}\rangle_{\hat{\mathbf{y}}M}} \,. \tag{4.40}$$

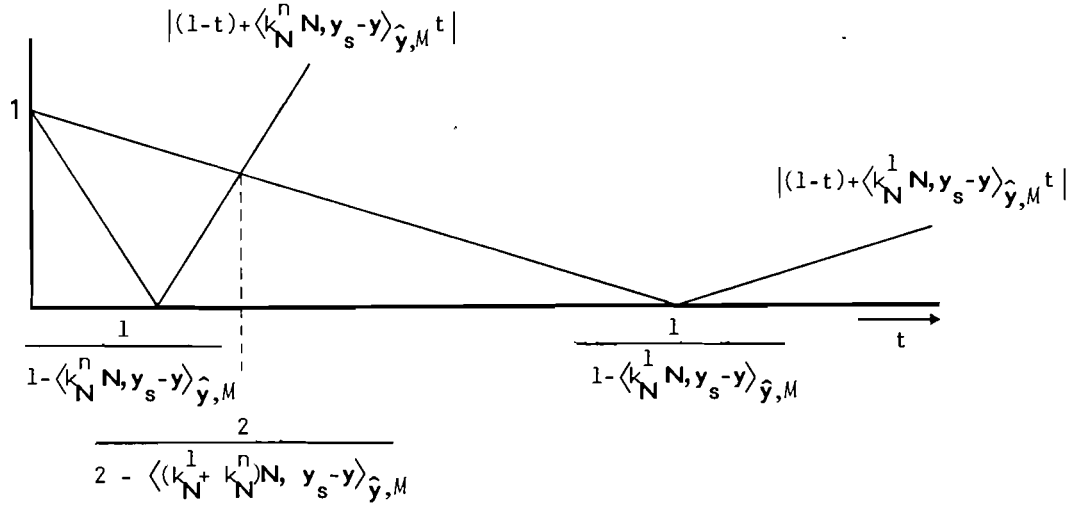


figure 28

Substitution of (4.40) into (4.39) gives then

$$u_{q+1}^r = \left(\frac{\langle (k_N^1-k_N^n)N, y_s-y\rangle_{\hat{y}M}}{2-\langle (k_N^1+k_N^n)N, y_s-y\rangle_{\hat{y}M}} - 2\,\frac{\langle (k_N^1-k_N^r)N, y_s-y\rangle_{\hat{y}M}}{2-\langle (k_N^1+k_N^n)N, y_s-y\rangle_{\hat{y}M}} \right) u_q^r + O(u_q^t \delta_{ts} u_q^s). \qquad (4.41)$$

And from this follows that the smallest attainable linear convergence factor (lcf.) for Gauss' method with a line search strategy is given by:

$$\text{lcf.} = \frac{\langle (k_N^1-k_N^n)N, y_s-y\rangle_{\hat{y}M}}{2-\langle (k_N^1+k_N^n)N, y_s-y\rangle_{\hat{y}M}} . \qquad (4.42)$$

Note that although now local convergence is guaranteed if $\hat{y}$ is a strict local minimum, convergence can still be very slow; namely when $k_N^1 - k_N^n >> 0$ for instance.

The above discussed Gauss' method with the optimal choice (4.40) is of course not practical executable as such, since we generally lack the curvature information needed. Nevertheless, the above results are of some importance since with (4.42) we have obtained a lower bound on the linear convergence factor attainable for Gauss' method with a line search strategy. This means that when one decides to use a line search strategy in practice, one should choose a strategy which gives a rate of convergence close to (4.42).

Apart from the minimization rule which will be used in the next section to establish global convergence, we shall not discuss in the sequel any of the existing line search strategies. For details the reader is therefore referred to the relevant literature (see e.g. Ortega & Rheinboldt, 1970). Our decision of not including a discussion on various line search strategies is mainly based on the

following important conclusion:

(vii) If $\langle (k_N^1+k_N^n)\mathbf{N}, \mathbf{y}_s-\mathbf{y}\rangle_{\hat{\mathbf{y}}M}$ is small, then $t_q=1$ is a good choice for a line search strategy (see (4.40)). Hence, for small residual adjustment problems and moderately curved submanifolds $\bar{N}$, Gauss' method without a line search strategy has a close to optimal rate of convergence. In fact, if either $\mathbf{B}\equiv\mathbf{0}$ or $\mathbf{y}_s = \hat{\mathbf{y}}$, one must choose $t_q = 1$ in order to assure a local quadratic convergence behaviour.

(viii) From (4.35) follows that Gauss' method has a local **quadratic** convergence behaviour if either the normalfield $\mathbf{B}$ vanishes identically on $\bar{N}$, i.e. $\mathbf{B}\equiv\mathbf{0}$, or $\mathbf{y}_s \in \bar{N}$, i.e. $\mathbf{y}_s = \hat{\mathbf{y}}$. Submanifolds for which $\mathbf{B}\equiv\mathbf{0}$ are called **totally geodesic.** This as a generalization of the concept of a geodesic ("straight line") for which the first curvature vanishes identically.
The local quadratic convergence behaviour is described by

$$x^\gamma_{q+1} - \hat{x}^\gamma = \frac{1}{2}\,\Gamma^\gamma_{\alpha\beta}(\hat{x})\,(x^\alpha_q-\hat{x}^\alpha)(x^\beta_q-\hat{x}^\beta) + O(\|x_q-\hat{x}\|^3)\,. \tag{4.43}$$

Of course, we still have to prove (4.43). But it is reasonable to expect that (4.43) holds, since we know from the previous section that for geodesics Gauss' method has a local quadratic convergence behaviour with convergence factor $\frac{1}{2}\,(s'(\hat{t}))^{-1}s''(\hat{t})$. And we also know that $(s'(t))^{-1}s''(t)$ generalizes to the Christoffel symbols of the second kind $\Gamma^\gamma_{\alpha\beta}$.
If $\mathbf{B}\equiv\mathbf{0}$, then $T_{\mathbf{y}}\bar{N} = \bar{N}$ which means that our actual adjustment problem is linear. Hence, if $\mathbf{B}\equiv\mathbf{0}$ then

$$\hat{\mathbf{y}} = \mathbf{y}_1 + P_{T\bar{N},T\bar{N}^\perp}(\mathbf{y}_s-\mathbf{y}_1) \text{ for some } \mathbf{y}_1\in\bar{N}\,, \tag{4.44}$$

from which follows that

$$\langle \mathbf{y}_*(\partial_\alpha),\mathbf{y}_s-\mathbf{y}\rangle_{\mathbf{y}_qM} = \langle \mathbf{y}_*(\partial_\alpha),P_{T\bar{N},T\bar{N}^\perp}(\mathbf{y}_s-\mathbf{y})\rangle_{\mathbf{y}_qM} = \langle \mathbf{y}_*(\partial_\alpha),\hat{\mathbf{y}}-\mathbf{y}\rangle_{\mathbf{y}_qM}\,. \tag{4.45}$$

This already shows that indeed the convergence behaviour will be the same if either $\mathbf{B}\equiv\mathbf{0}$ or $\mathbf{y}_s = \hat{\mathbf{y}}$ holds. Remember that in both cases we are actually solving the inverse mapping problem: given $\hat{\mathbf{y}} = \mathbf{y}_1 + P_{T\bar{N},T\bar{N}^\perp}(\mathbf{y}_s-\mathbf{y}_1)$ for some $\mathbf{y}_1\in\bar{N}$, find the pre-image $\hat{\mathbf{x}}$ under map $\mathbf{y}$. To prove the quadratic convergence behaviour (4.43), we Taylorize the right-hand side of

$$\Delta x^\beta_q = \langle \mathbf{y}_*(g^{\beta\alpha}\partial_\alpha),\mathbf{y}_s-\mathbf{y}\rangle_{\mathbf{y}_qM}\,,$$

about the least-squares solution $\hat{\mathbf{y}}$. With (4.45) and $x^\beta_{q+1} = x^\beta_q + \Delta x^\beta_q$ this gives

$$x^\beta_{q+1}-\hat{x}^\beta = -\frac{1}{2}\langle D_{\mathbf{y}_*(\partial_\gamma)}\mathbf{y}_*(g^{\beta\alpha}\partial_\alpha),\mathbf{y}_*(\partial_\delta)\rangle_{\hat{\mathbf{y}}M}\,(x^\gamma_q-\hat{x}^\gamma)(x^\delta_q-\hat{x}^\delta) + O(\|x_q-\hat{x}\|^3). \tag{4.46}$$

But according to Gauss' decomposition formula we have

$$\langle D_{y_*(\partial_\gamma)} y_*(g^{\beta\alpha}\partial_\alpha), y_*(\partial_\delta)\rangle_M = \langle y_*(\nabla_{\partial_\gamma} g^{\beta\alpha}\partial_\alpha), y_*(\partial_\delta)\rangle_M = \langle \nabla_{\partial_\gamma} g^{\beta\alpha}\partial_\alpha, \partial_\delta\rangle_N .$$

And since

$$\nabla_{\partial_\gamma} g^{\beta\alpha}\partial_\alpha = -\Gamma^{\beta}_{\gamma\tau} g^{\tau\rho}\partial_\rho ,$$

this gives

$$\langle D_{y_*(\partial_\gamma)} y_*(g^{\beta\alpha}\partial_\alpha), y_*(\partial_\delta)\rangle_M = -\Gamma^{\beta}_{\gamma\delta} .$$

Hence, with (4.46) the quadratic convergence rule (4.43) follows.

(ix) As another generalization of the univariate case we have that the local quadratic convergence rule (4.43) is **not** invariant to nonlinear reparametrization. This follows from the fact that the Christoffel symbols are not the components of a tensor.

With the reparametrization $\bar{x}^{\bar\alpha}(x^\alpha)$, their transformation law reads namely as

$$\bar\Gamma^{\bar\gamma}_{\bar\alpha\bar\beta} = \left\{ \Gamma^{\gamma}_{\alpha\beta} \frac{\partial x^\alpha}{\partial \bar x^{\bar\alpha}} \frac{\partial x^\beta}{\partial \bar x^{\bar\beta}} + \frac{\partial^2 x^\gamma}{\partial \bar x^{\bar\alpha} \partial \bar x^{\bar\beta}} \right\} \frac{\partial \bar x^{\bar\gamma}}{\partial x^\gamma} . \tag{4.47}$$

Note that this is the generalization of the easily verifiable tranformation rule

$$(s'(\bar t))^{-1} s''(\bar t) = \left\{ (s'(t))^{-1} s''(t) \frac{dt}{d\bar t}\frac{dt}{d\bar t} + \frac{d^2 t}{d\bar t^2} \right\} \frac{d\bar t}{dt} .$$

With respect to the univariate case there is however one big difference. In the univariate case we could always find a parametrization for which $(s'(t))^{-1}s''(t)$ would vanish identically. In the multivariate case however this is only possible if $\mathbf{B} \equiv \mathbf{0}$. The explanation is that in the univariate case $T_t N$ and N are identifiable irrespective the curvature of the space curve $\mathbf{c}$, whereas in the multivariate case $T_x N$ and N are only identifiable if $\mathbf{B} \equiv \mathbf{0}$. Namely, only if $\mathbf{B} \equiv \mathbf{0}$ can one find a parametrization for which $\langle .,. \rangle_N$ reduces to the standard metric globally.
Nevertheless there do exist parametrizations for which the Christoffel symbols $\Gamma^{\gamma}_{\alpha\beta}$ vanish locally. Coordinates for which the Christoffel symbols vanish at a point, $\mathbf{x}_0$ say, are geodesic polar coordinates.
The procedure of finding geodesic polar coordinates is the following:
According to the theory of ordinary differential equations a geodesic $\mathbf{c}(s)$ through a point $\mathbf{x}_0$ is locally **uniquely** characterized by the coordinates of $\mathbf{x}_0 = \mathbf{c}(0)$ and the tangent vector $\mathbf{c}_*(\frac{d}{ds})$ at $\mathbf{x}_0$. Hence a point $\mathbf{x} = \mathbf{c}(s) \in N$ on this geodesic can be identified by $\mathbf{c}_*(\frac{d}{ds})$ at $\mathbf{x}_0$ and s. Or in coordinates: the point $\mathbf{x} \in N$ with coordinates $x^\alpha = c^\alpha(s)$ can be identified locally with the point $\mathbf{X} \in T_{x_0} N$ having coordinates

$$X^\alpha = x_0^\alpha + s \frac{dc^\alpha}{ds}(0) . \tag{4.48}$$

Thus, since the geodesic $\mathbf{c}(s)$ is locally uniquely characterized by $\mathbf{x}_o = \mathbf{c}(o)$ and $\mathbf{c}_*(\frac{d}{ds})$ at $\mathbf{x}_o$, there exists locally a diffeomorphism from N into $T_{\mathbf{x}_o} N$. Let us denote this map in coordinates by

$$X^\alpha = X^\alpha(x^\alpha). \tag{4.49}$$

From the Taylor expansion of $\mathbf{c}(s)$,

$$x^\alpha = c^\alpha(s) = x_o^\alpha + \frac{dc^\alpha}{ds}(0)s + \frac{1}{2}\frac{d^2c^\alpha}{ds^2}(0)\,s^2 + \dots ,$$

follows then with

$$\frac{d^2c^\alpha}{ds^2} + \Gamma^\alpha_{\beta\gamma}\frac{dc^\beta}{ds}\frac{dc^\gamma}{ds} = 0,$$

that

$$x^\alpha = c^\alpha(s) = x_o^\alpha + \frac{dc^\alpha}{ds}(0)\,s - \frac{1}{2}\Gamma^\alpha_{\beta\gamma}(0)\frac{dc^\beta}{ds}(0)\frac{dc^\gamma}{ds}(0)\,s^2 + \dots .$$

Or with (4.48),

$$x^\alpha = X^\alpha - \frac{1}{2}\Gamma^\alpha_{\beta\gamma}(X^\beta - x_o^\beta)(X^\gamma - x_o^\gamma) + \dots$$

The inverse of this relation gives then

$$X^\alpha = x^\alpha + \frac{1}{2}\Gamma^\alpha_{\beta\gamma}(x^\beta - x_o^\beta)(x^\gamma - x_o^\gamma) + \dots \tag{4.50}$$

as the desired expression for (4.49).

We can now view (4.50) as a nonlinear parametertransformation. It is admissible since the Jacobian determinant equals 1 at $\mathbf{x}_o$. The new coordinates X^α are known as **geodesic polar coordinates.**

In these new coordinates the geodesic $\mathbf{c}(s)$ is found as the solution of

$$\frac{d^2X^\alpha}{ds^2} + \bar{\Gamma}^\alpha_{\beta\gamma}\frac{dX^\beta}{ds}\frac{dX^\gamma}{ds} = 0 \quad ,$$

where the new Christoffel symbols $\bar{\Gamma}^\alpha_{\beta\gamma}$ follow from (4.47) using (4.50). But as is easily verified the coefficients $\bar{\Gamma}^\alpha_{\beta\gamma}$ vanish at $\mathbf{x}_o$. Hence in a neighbourhood of $\mathbf{x}_o$ the geodesic $\mathbf{c}(s)$ is given in geodesic polar coordinates as

$$X^\alpha(s) = x_o^\alpha + s\frac{dc^\alpha}{ds}(0) \; . \tag{4.51}$$

From the above discussion follows that if the coordinates x^α in (4.46) are geodesic polar coordinates at $\hat{\mathbf{x}}$ by chance, and $\mathbf{B} \neq \mathbf{0}$ but $\mathbf{y}_s = \hat{\mathbf{y}}$, then Gauss' method has a local **third order** convergence behaviour. Note by the way that since the geodesic polar coordinates X^α are linear in s we are indeed dealing here with the proper multivariate generalization of the case considered in the previous

section where the univariate parameter t was chosen as linear function of s so as to eliminate the necessity of iteration for solving the inverse mapping problem.

4.5. Global convergence

In the above local analysis of Gauss' method we have seen that both the initial guess $\mathbf{x}_0$ had to be sufficiently close to the solution $\hat{\mathbf{x}}$ and $| \langle k_N^r \mathbf{N}, \mathbf{y}_s - \mathbf{y} \rangle_{\hat{\mathbf{y}} M} | < 1$ had to hold for all r = 1,..,n in order to assure convergence. For most practical problems we indeed believe that these conditions are satisfied. Nevertheless, it would be dissatisfactory not to have an iteration method which guarantees convergence almost independently of the chosen initial guess and curvature behaviour of the submanifold $\bar{N}$. In the following we will discuss therefore the necessary conditions which assure global convergence. Note that the adjective "global" does not refer to $\hat{\mathbf{x}}$, but to the almost independency of the initial guess $\mathbf{x}_0$, i.e. usually one will have global convergence to a local minimum. The method we will discuss is essentially the above discussed Gauss' method, but now with the so-called minimization rule as line search strategy. In formulating the method we have chosen to start from some general principles so as to get a better understanding of how the various assumptions contribute to the overall proof of global convergence.

As a start we assume

$$\text{that we are given a sequence } \{\mathbf{x}_q\} \text{ for which } E(\mathbf{x}_{q+1}) \leq E(\mathbf{x}_q), \quad \text{for all } q = 0,1,\ldots \tag{4.52}$$

This seems a natural conditon to start with since we are looking for an iteration method which can locate a local minimum of E. From (4.52) follows that the sequence $\{E(\mathbf{x}_q)\}$ converges to a limit, since the sum of squares function E is bounded from below $[0 \leq E(\mathbf{x}), \forall \mathbf{x}]$ and the sequence $\{E(\mathbf{x}_q)\}$ is non-increasing.

Now, in order to find an appropriate iteration method which generates a sequence $\{\mathbf{x}_q\}$ satisfying the conditions of (4.52), we first need to know, given an initial guess, in which direction to proceed. In ordinary vector analysis the gradient of a scalar field E is defined as the vector field $\partial_\alpha E$, $\alpha=1,\ldots,n$. And it is well known that $-\partial_\alpha E$ points in the direction in which the function E decreases most rapidly locally. In view of (4.52) it seems therefore appropriate to proceed in the direction of $-\partial_\alpha E$. However, this ordinary definition of gradient is not invariant under a change of coordinates. With our geometric exposition of the preceding sections in mind we can therefore expect that the simplicity of the ordinary vector analytic definition of the gradient almost inevitably forces difficulties and awkwardness when problems involving change of coordinates are encountered. A way out of this dilemma is offered if we bring the requirements of invariance under change of coordinates to the foreground. Therefore, given a function $E: N \to \mathbb{R}$ we define the gradient field, denoted by **grad** E, invariantly by

$$\langle \mathbf{grad}\ E, X \rangle_N = X(E) \quad \text{for all vector fields } X \text{ on } N. \tag{4.53}$$

In local coordinates this expression reads as

$$(\text{grad } E)^\alpha g_{\alpha\beta} X^\beta = X^\beta \partial_\beta E.$$

And this gives

$$(\text{grad } E)^\alpha = g^{\alpha\beta} \partial_\beta E. \tag{4.54}$$

Since the direction for which

$$\frac{\langle \mathbf{grad}\ E, \Delta\mathbf{x} \rangle_N}{\langle \Delta\mathbf{x}, \Delta\mathbf{x} \rangle_N^{\frac{1}{2}}}$$

is minimized as function of $\Delta\mathbf{x} \neq \mathbf{0}$, is given by

$$\mathbf{x} = \Delta\mathbf{x}(\mathbf{x}) = -\mathbf{grad}\ E(\mathbf{x}) \in T_{\mathbf{x}} N \quad , \tag{4.55}$$

it follows that $\Delta\mathbf{x}(\mathbf{x})$ points in the direction of maximal local decrease of E. Note that since

$$\partial_\alpha E(x) = -\partial_\alpha y^i(x) g_{ij}(y_s^j - y^j(x)),$$

the vector

$$\Delta x_q^\alpha = \Delta x^\alpha(x_q) = -\, g^{\alpha\beta}(x_q) \partial_\beta E(x_q)$$

equals the incremental step as produced by Gauss' method (4.5). Hence, both the geometry of our non-linear least-squares problem as well as the fact that -grad E points in the direction of maximal local decrease of E, suggest that the vector $\Delta\mathbf{x}(\mathbf{x})$ as given by (4.55) is an appropriate choice for the direction of search. However, although $\Delta\mathbf{x}(\mathbf{x})$ points in the direction of maximal local decrease of E, this does not necessarily imply that the function value of $E(\mathbf{x})$ decreases by taking $\Delta\mathbf{x}(\mathbf{x})$ as incremental step. In fact we already saw in the previous section that the descent property only holds if $\bar{N}$ is moderately curved and $\mathbf{x}$ sufficiently close to $\hat{\mathbf{x}}$. So, we still need a rule according to which we can compute an appropriate $\mathbf{x}_{q+1}$ from $\mathbf{x}_q$. Nevertheless, the above discussion is not without meaning since by agreeing upon taking $\Delta\mathbf{x}(\mathbf{x}_q)$ as the direction of search we have reduced the dimensions of our problem essentially from n to 1. That is, by choosing a curve $\mathbf{c}_q$: $t \in \mathbb{R} \to N$, with

$$\mathbf{c}_q(t=0) = \mathbf{x}_q \quad \text{and} \quad \mathbf{c}_{q*}\left(\frac{\mathbf{d}}{\mathbf{dt}}\right)_{\mathbf{x}_q} = \Delta\mathbf{x}_q = \Delta\mathbf{x}(\mathbf{x}_q) = -\mathbf{grad}\ E(\mathbf{x}_q), \tag{4.56}$$

we can define $\mathbf{x}_{q+1}$ by $\mathbf{x}_{q+1} = \mathbf{c}_q(t_q)$, where t_q is an appropriate scalar so that

$$E(\mathbf{x}_{q+1}) = E(\mathbf{c}_q(t_q)) \leq E(\mathbf{c}_q(0)) = E(\mathbf{x}_q)$$

holds. That such a scalar exists is seen as follows. Since

$$\lim_{t\to 0} \frac{E(\mathbf{c}_q(t))-E(\mathbf{c}_q(o))}{t} = \partial_\alpha E(x_q)\,\frac{dc_q^\alpha}{dt}(0) = \left\langle \mathbf{grad}\, E,\ \mathbf{c}_{q*}\left(\frac{d}{dt}\right)\right\rangle_{\mathbf{x}_q N},$$

it follows with $\mathbf{c}_{q*}\left(\frac{d}{dt}\right)_{\mathbf{x}_q} = \Delta\mathbf{x}(\mathbf{x}_q) = -\,\mathbf{grad}\, E\,(\mathbf{x}_q)$, that if $\Delta\mathbf{x}(\mathbf{x}_q) \neq \mathbf{0}$,

$$\lim_{t\to 0} \frac{E(\mathbf{c}_q(t)) - E(\mathbf{c}_q(o))}{t} = -\left\langle \mathbf{grad}\, E,\ \mathbf{grad}\, E\right\rangle_{\mathbf{x}_q N} < 0.$$

Hence, if $\Delta\mathbf{x}(\mathbf{x}_q) \neq \mathbf{0}$, there exists a $\delta > 0$ so that $E(\mathbf{c}_q(t)) < E(\mathbf{c}_q(o))$ for all $t \in (0,\delta)$. Thus if $\mathbf{x}_q$ is not a critical point of E it is always possible to choose a positive scalar t_q so that

$$E(\mathbf{x}_{q+1}) = E(\mathbf{c}_q(t_q)) < E(\mathbf{c}_q(o)) = E(\mathbf{x}_q)). \tag{4.57}$$

It seems appropriate to choose t_q so that the maximal possible decrease in E is obtained. This is the case when t_q is chosen so as to minimize E along the curve $\mathbf{c}_q(t)$. That is, when t_q is computed as the scalar satisfying the minimization rule

$$E(\mathbf{c}_q(t_q)) = \min_{t>0} E(\mathbf{c}_q(t)). \tag{4.58}$$

So far we did not specify the type of curve $\mathbf{c}_q(t)$ chosen. The simplest way computationwise would be to choose the curve $\mathbf{c}_q(t)$ so that its coordinate functions are given by

$$c_q^\alpha(t) = x_q^\alpha + t\ \Delta x_q^\alpha.$$

But other choices are also possible. And since the particular type of curve chosen is not important for our convergence analysis, we just assume that a rule is given which smoothly assigns a unique curve $\mathbf{c}_q\colon t\in\mathbb{R}\to N$ to every point $\mathbf{x}_q$ so that the initial conditions (4.56) hold. That is, we assume that the coordinate functions c_q^α, $\alpha=1,\ldots,n$, of the curve $\mathbf{c}_q$ are smooth functions of not only the parameter t but also of the initial conditions. Instead of $\mathbf{c}_q(t)$ we may therefore write $\mathbf{c}(t,\mathbf{x}_q,\Delta\mathbf{x}(\mathbf{x}_q))$ and by Taylor's formula we have

$$\begin{aligned} c^\alpha(t,x_q,\Delta x_q) &= c_q^\alpha(0) + \frac{dc_q^\alpha}{dt}(0)t + \frac{1}{2}\,\phi^\alpha(t,x_q,\Delta x_q)t^2 \\ &= x_q^\alpha + \Delta x_q^\alpha t + \frac{1}{2}\,\phi^\alpha(t,x_q,\Delta x_q)t^2. \end{aligned} \tag{4.59}$$

where the smooth functions ϕ^α depend on the rule given.

Summarizing, we have come up with the following iteration method:

(i) Choose an initial guess $\mathbf{x}_0$ and set q = 0. Choose a rule which smoothly assigns a unique curve $\mathbf{c}_q : t \in \mathbb{R} \to N$ to every point $\mathbf{x}_q$ with the prescribed initial conditions

$$\mathbf{c}_q(o) = \mathbf{x}_q \quad \text{and} \quad \mathbf{c}_{q*}\left(\frac{d}{dt}\right)_{x_q} = \Delta\mathbf{x}(\mathbf{x}_q) \neq \mathbf{0}.$$

(ii) Compute $\Delta\mathbf{x}(\mathbf{x}_q) = -\,\mathbf{grad}\, E(\mathbf{x}_q)$. If $\Delta\mathbf{x}(\mathbf{x}_q) = \mathbf{0}$ then stop. (4.60)

(iii) Compute the scalar t_q satisfying $E(\mathbf{c}_q(t_q)) = \min_{t > 0} E(\mathbf{c}_q(t))$.

(iv) Compute $\mathbf{x}_{q+1} = \mathbf{c}_q(t_q)$ and set q = q+1. Return to (ii).

The sequence $\{\mathbf{x}_q\}$ generated by (4.60) is either finite or infinite. If it is finite then clearly its limit is a critical (or stationary) point of E by virtue of the stop command in (4.60). But if it is infinite then the only thing we know for sure is that the sequence $\{E(\mathbf{x}_q)\}$ has a limit. It is important to realize, however, that this by itself implies nothing about the validity of the final convergence statement which we set out to prove, namely that $\lim_{q\to\infty} \mathbf{x}_q = \hat{\mathbf{x}}$, with $\hat{\mathbf{x}}$ being a critical point of E. This is best seen by means of an example: Take $E(\mathbf{x}) = m.e^{2\mathbf{x}}$, where m is a real-valued constant, and $\mathbf{x}_q = 2^{-q}$. Then $\lim_{q\to\infty} E(\mathbf{x}_q) = m$ and $\lim_{q\to\infty} \mathbf{x}_q = \mathbf{0}$, but $\mathbf{x} = \mathbf{0}$ is clearly not a critical point of E.
In fact, the convergence of the sequence $\{E(\mathbf{x}_q)\}$ does in general not even imply the convergence of the sequence $\{\mathbf{x}_q\}$. Therefore, in order to assure that the sequence $\{\mathbf{x}_q\}$ as generated by (4.60) converges to a critical point of E, we assume in addition to (4.52),

that the initial guess $\mathbf{x}_0$ is chosen such that the level set
$L(\mathbf{x}_0) = \{\mathbf{x} \mid E(\mathbf{x}) \leq E(\mathbf{x}_0)\}$ is bounded, and that the function values of E at critical points in $L(\mathbf{x}_0)$ are distinct. (4.61)

With (4.61) we are now in the position to prove that the sequence $\{\mathbf{x}_q\}$ converges to a critical point of E. We will assume that the sequence $\{\mathbf{x}_q\}$ is infinite.
According to (4.52) we have $E(\mathbf{x}_{q+1}) \leq E(\mathbf{x}_q)$ for all q=0,1,2... . Hence, $\mathbf{x}_q \in L(\mathbf{x}_0)$ for all q=0,1,2... . And since the level set $L(\mathbf{x}_0)$ is bounded by assumption, it follows that $\{\mathbf{x}_q\}$ has at least one convergent subsequence, say $\{\mathbf{x}_{q_i}\}$, where $q_{i+1} > q_i$, and with limit $\lim_{i\to\infty} \mathbf{x}_{q_i} = \hat{\mathbf{x}}$.
We shall now proof by contradiction that $\hat{\mathbf{x}}$ is a critical point of E. Assume therefore that $\hat{\mathbf{x}}$ is not a critical point of E.
We denote the unique curve assigned to $\hat{\mathbf{x}}$ by $\mathbf{c}(t,\hat{\mathbf{x}},\Delta\mathbf{x}(\hat{\mathbf{x}}))$, and the positive scalar $\hat{t}$ satisfying $E(\mathbf{c}(\hat{t},\hat{\mathbf{x}},\Delta\mathbf{x}(\hat{\mathbf{x}}))) = \min_{t > 0} E(\mathbf{c}(t,\hat{\mathbf{x}},\Delta\mathbf{x}(\hat{\mathbf{x}})))$ by $\hat{t} = t(\hat{\mathbf{x}})$. Similarly, we denote the unique curve assigned to an arbitrary point $\mathbf{x}$ by $\mathbf{c}(t,\mathbf{x},\Delta\mathbf{x}(\mathbf{x}))$; and the scalar t' satisfying $E(\mathbf{c}(t',\mathbf{x},\Delta\mathbf{x}(\mathbf{x}))) = \min_{t>0} E(\mathbf{c}(t,\mathbf{x},\Delta\mathbf{x}(\mathbf{x})))$ by t'=t(x).
Now we define a function $F(\mathbf{x})$ as

$$F(\mathbf{x}) = E(\mathbf{c}(t(\hat{\mathbf{x}}),\mathbf{x},\mathbf{x}(\mathbf{x}))) - E(\mathbf{x}) \ . \tag{4.62}$$

Since $F(\mathbf{x})$ is continuous by inspection and $\lim_{i\to\infty} \mathbf{x}_{q_i} = \hat{\mathbf{x}}$, it follows that

$$\lim_{i\to\infty} F(\mathbf{x}_{q_i}) = F(\hat{\mathbf{x}}).$$

From the definition of the limit of a convergent sequence (see e.g. W. Flemming, 1977) follows then that for every $\varepsilon > 0$ there exists a positive integer r such that

$$|F(\mathbf{x}_{q_i}) - F(\hat{\mathbf{x}})| \leq \varepsilon \text{ for every } i \geq r. \tag{4.63}$$

Since we assumed $\hat{\mathbf{x}}$ to be a non-critical point, we have

$$F(\hat{\mathbf{x}}) = E(\mathbf{c}(t(\hat{\mathbf{x}}),\hat{\mathbf{x}}, \mathbf{x}(\hat{\mathbf{x}}))) - E(\hat{\mathbf{x}}) < 0.$$

Hence, we can take $\varepsilon > 0$ in (4.63) to be $\varepsilon = \frac{1}{2}|F(\hat{\mathbf{x}})|$. This gives us then

$$F(\mathbf{x}_{q_i}) \leq \tfrac{1}{2} F(\hat{\mathbf{x}}) < 0 \quad \text{for every } i \geq r \tag{4.64}$$

From

$$E(\mathbf{c}(t(\mathbf{x}),\mathbf{x}, \mathbf{x}(\mathbf{x}))) \leq E(\mathbf{c}(t,\mathbf{x}, \mathbf{x}(\mathbf{x})))$$

or

$$E(\mathbf{c}(t(\mathbf{x}),\mathbf{x}, \mathbf{x}(\mathbf{x}))) - E(\mathbf{x}) \leq F(\mathbf{x}),$$

follows then that

$$E(\mathbf{c}(t(\mathbf{x}_{q_i}),\mathbf{x}_{q_i}, \mathbf{x}(\mathbf{x}_{q_i}))) - E(\mathbf{x}_{q_i}) \leq F(\mathbf{x}_{q_i}) \leq \tfrac{1}{2} F(\hat{\mathbf{x}}) < 0 \text{ for every } i \geq r,$$

or

$$E(\mathbf{x}_{q_i+1}) \leq E(\mathbf{x}_{q_i}) + \tfrac{1}{2} F(\hat{\mathbf{x}}), \quad \text{with } F(\hat{\mathbf{x}}) \leq 0, \text{ for every } i \geq r.$$

With $E(\mathbf{x}_{q_{i+1}}) \leq E(\mathbf{x}_{q_i+1})$ follows that

$$E(\mathbf{x}_{q_{i+1}}) \leq E(\mathbf{x}_{q_i}) - \tfrac{1}{2}|F(\hat{\mathbf{x}})| \text{ with } |F(\hat{\mathbf{x}})| \neq 0, \text{ for every } i \geq r.$$

Hence,

$$\lim_{i\to\infty} E(\mathbf{x}_{q_i}) = -\infty . \tag{4.65}$$

Thus if $\hat{\mathbf{x}}$ is not a critical point then (4.65) must hold. But this contradicts the fact that $\{E(\mathbf{x}_{q_i})\}$ converges to $E(\hat{\mathbf{x}})$. Hence, $\hat{\mathbf{x}}$ must be a critical point of E.

To prove that the sequence $\{\mathbf{x}_q\}$ itself converges to a critical point of E, suppose that $\hat{\mathbf{x}}$ and $\hat{\hat{\mathbf{x}}}$ are distinct limits of two convergent subsequences of $\{\mathbf{x}_q\}$. We know then that $\hat{\mathbf{x}}$ and $\hat{\hat{\mathbf{x}}}$ must be critical points of E. And since $\{E(\mathbf{x}_q)\}$ converges, we must have $E(\hat{\mathbf{x}}) = E(\hat{\hat{\mathbf{x}}})$. But this contradicts with our assumption that the critical values of E are distinct. Hence we must have that $\hat{\mathbf{x}} = \hat{\hat{\mathbf{x}}}$ which means that the sequence $\{\mathbf{x}_q\}$ itself converges to a critical point.

This concludes the proof of the following **global convergence theorem** (Ortega & Rheinhold, 1970):

Let an initial guess $\mathbf{x}_0$ be chosen such that the level set $L(\mathbf{x}_0)=\{\mathbf{x} \mid E(\mathbf{x}) \leq E(\mathbf{x}_0)\}$ is bounded, and let the function values of E be distinct at critical points in $L(\mathbf{x}_0)$. Then the sequence $\{\mathbf{x}_q\}$ defined by (4.60) is either finite and terminates at a critical point of E, or it is infinite and converges to a critical point, i.e. $\lim_{q\to\infty} \mathbf{x}_q = \hat{\mathbf{x}}$ with $\mathbf{grad}\, E(\hat{\mathbf{x}}) = \mathbf{0}$. (4.66)

To conclude this section we will prove the following result on the rate of convergence of the globally convergent iteration method (4.60):

If

$$k_N^1 \, ||\mathbf{y}_s-\hat{\mathbf{y}}||_M < 1 \, ,$$

then

$$\lim_{q\to\infty} \frac{||\mathbf{y}_s-\mathbf{y}(\mathbf{x}_{q+1})||_M^2 - ||\mathbf{y}_s-\hat{\mathbf{y}}||_M^2}{||\mathbf{y}_s-\mathbf{y}(\mathbf{x}_q)||_M^2 - ||\mathbf{y}_s-\hat{\mathbf{y}}||_M^2} \leq \frac{(k_N^1-k_N^n)^2 ||\mathbf{y}_s-\hat{\mathbf{y}}||_M^2}{(2-(k_N^1+k_N^n)||\mathbf{y}_s-\hat{\mathbf{y}}||_M)^2}. \tag{4.67}$$

Recall from (4.60) that in order to generate the sequence $\{\mathbf{x}_q\}$ one should first decide upon a descent curve $\mathbf{c}(t,\mathbf{x}_q,\Delta\mathbf{x}(\mathbf{x}_q))$. Fortunately all methods for selecting such a curve are asymptotically equivalent in the sense that the curves are all tangent at the starting point $\mathbf{x}_q$. That is, as the stepsize goes to zero the methods all move approximately along the same curve, which implies that the asymptotic properties of the sequence $\{\mathbf{x}_q\}$ are independent of the type of curve chosen provided that the initial conditions (4.56) hold. Hence, for the determination of the local rate of convergence we are free in choosing the type of curve $\mathbf{c}_q(t)$. For convenience we will assume therefore that the descent curve $\mathbf{c}_q(t)$ is a geodesic.

Now, before we prove (4.67) we will first prove that the linear map $\mathbf{H}: T_x N \to T_x N$ defined by

$$\mathbf{H}\,\mathbf{X} = \nabla_{\mathbf{X}} \mathbf{grad}\, E \qquad \text{for all} \quad \mathbf{X} \in T_x N, \tag{4.68}$$

satisfies

$$\langle \mathbf{H}\,\mathbf{X},\mathbf{Y}\rangle_N = \langle \mathbf{X},\mathbf{Y}\rangle_N - \langle \mathbf{B}(\mathbf{X},\mathbf{Y}),\mathbf{N}\rangle_M \quad \text{for all} \quad \mathbf{X},\mathbf{Y} \in T_x N, \tag{4.69}$$

where

$$\mathbf{N} = P_{T\bar{N}^{\perp},T\bar{N}}(\mathbf{y}_s-\mathbf{y}) \, . \tag{4.70}$$

From (4.70) and the definition of the pushforward of **grad** E ,

$$\mathbf{y}_*(\mathbf{grad}\, E) = -P_{T\bar{N},T\bar{N}^{\perp}}(\mathbf{y}_s-\mathbf{y}),$$

follows that

$$\mathbf{y}_s-\mathbf{y} = P_{T\bar{N},T\bar{N}^{\perp}}(\mathbf{y}_s-\mathbf{y}) + P_{T\bar{N}^{\perp},T\bar{N}}(\mathbf{y}_s-\mathbf{y}) = -\mathbf{y}_*(\mathbf{grad}\, E) + \mathbf{N}.$$

And with $D_{\mathbf{y}_*(\mathbf{X})}(\mathbf{y}_s-\mathbf{y}) = -\mathbf{y}_*(\mathbf{X})$, this gives

$$-y_*(X) = -D_{y_*(X)}y_*(\text{grad } E) + D_{y_*(X)}N.$$

Hence,

$$\langle X,Y\rangle_N = \langle y_*(X),y_*(Y)\rangle_M = \langle D_{y_*(X)}y_*(\text{grad } E),\ y_*(Y)\rangle_M - \langle D_{y_*(X)}N,y_*(Y)\rangle_M. \tag{4.71}$$

Since,

$$0 = D_{y_*(X)}\langle N,y_*(Y)\rangle_M = \langle D_{y_*(X)}N,y_*(Y)\rangle_M + \langle N,D_{y_*(X)}y_*(Y)\rangle_M\ ,$$

we can write (4.71) also as

$$\langle X,Y\rangle_N = \langle D_{y_*(X)}y_*(\text{grad } E),y_*(Y)\rangle_M + \langle D_{y_*(X)}y_*(Y),N\rangle_M\ .$$

Two times application of Gauss' decomposition formula (4.18) gives then

$$\langle X,Y\rangle_N = \langle y_*(\nabla_X\text{grad } E),y_*(Y)\rangle_M + \langle B(X,Y),N\rangle_M\ .$$

or

$$\langle \nabla_X\text{grad } E,Y\rangle_N = \langle H\,X,Y\rangle_N = \langle X,Y\rangle_N - \langle B(X,Y),N\rangle_M\ ,$$

which proves (4.69).

With (4.29), it follows from (4.69) that

$$\frac{\langle H\,X,X\rangle_N}{\langle X,X\rangle_N} = 1 - k_N\,||P_{T\bar{N}^{\perp},T\bar{N}}(y_s-y)||_M\ . \tag{4.72}$$

But for $X = \text{grad } E(x_q)$, this is precisely to a first order approximation the inverse of the scalar t_q satisfying the minimization rule $E(c_q(t_q)) = \min_{t>0} E(c_q(t))$. To see this, take a plane section of the submanifold $\bar{N}$ through the points $y(x_q)$, y_s and $y(x_q) - y_*(\text{grad } E(x_q))$, and approximate the resulting plane curve by its circle of curvature (see figure 29).

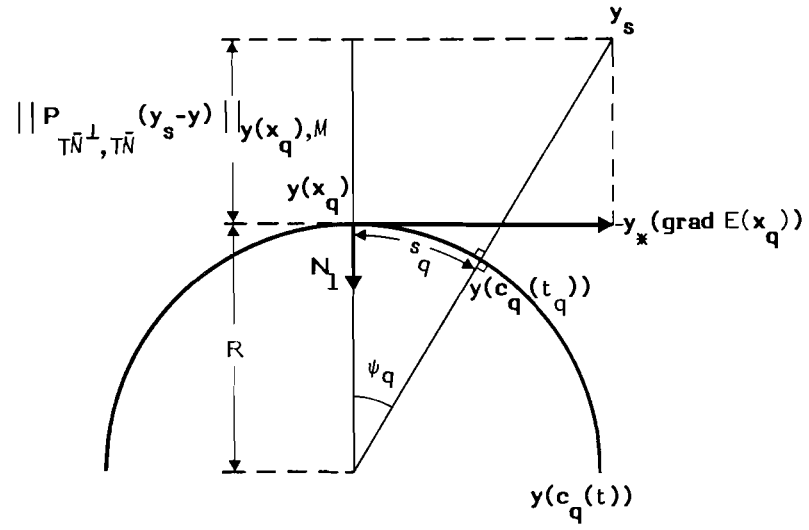


figure 29

In a neighbourhood of $y(x_q)$ this circle of curvature can then be considered as a sufficient approximation of the curve $y(c_q(t))$. Since the curve $c_q(t)$ is a geodesic by assumption it follows that

$$g_{\alpha\beta}(c_q(t)) \frac{dc_q^{\alpha}}{dt}(t) \frac{dc_q^{\beta}}{dt}(t)$$

is constant along the curve. Hence, s, the parameter of arclength, is proportional to t. Since

$$c_{q*}(\frac{d}{dt})_{x_q} = - \text{grad } E(x_q),$$

it follows therefore that

$$s = || \text{ grad } E(x_q) ||_N t = ||y_*(\text{grad } E(x_q))||_M t \quad . \tag{4.73}$$

Furthermore we know that the scalar t_q satisfies the minimization rule. Therefore

$$\langle y_*(c_{q*}(\frac{d}{dt})), y_s - y \rangle_{y(c_q(t_q))M} = 0$$

must hold. From figure 29 follows then that

$$\frac{s_q}{R} = \psi_q \simeq \tan \psi_q = \frac{||y_*(\text{grad } E)||_{y(x_q)M}}{R - \langle N_1, P_{T\bar{N}^{\perp}, T\bar{N}}(y_s - y)\rangle_{y(x_q)M}}, \tag{4.74}$$

where N_1 is the first normal of $y(c_q(t))$.
With (4.73) follows then

$$t_q \simeq \frac{R}{R - \langle N_1, P_{T\bar{N}^{\perp}, T\bar{N}}(y_s - y)\rangle_{y(x_q)M}} = \frac{1}{1 - k_{N_1} ||P_{T\bar{N}^{\perp}, T\bar{N}}(y_s - y)||_{y(x_q)M}} . \tag{4.75}$$

Compare with (4.72).
To make relation (4.75) precise we recall that geodesics are characterized by

$$\nabla_V V = 0, \quad \text{with} \quad V = c_{q*}(\frac{d}{dt}).$$

From the fact that t_q satisfies the minimization rule follows then

$$0 = \langle \text{grad } E, V \rangle_{c_q(t_q)} = \langle \text{grad } E, V \rangle_{c_q(o)} + \langle \nabla_V \text{grad } E, V \rangle_{c_q(o)} t_q + O(t_q^2).$$

And with (4.68) and $V_{c_q(o)} = c_{q*}(\frac{d}{dt})_{x_q} = - \text{grad } E(x_q)$ this gives

$$t_q = \frac{\langle \mathbf{grad}\, E,\ \mathbf{grad}\, E\rangle_{x_q N}}{\langle H\ \mathbf{grad}\, E,\ \mathbf{grad}\, E\rangle_{x_q N}} + O(\ ||\mathbf{grad}\, E||_{x_q N}\ ). \tag{4.76}$$

Compare with (4.75).

Now, to continue our proof of (4.67), we substitute (4.76) into

$$\begin{aligned} E(c_q(t_q)) - E(c_q(o)) &= \langle \mathbf{grad}\, E,\ V\rangle_{c_q(o)}\ t_q + \tfrac{1}{2}\langle \nabla_V \mathbf{grad}\, E,\ V\rangle_{c_q(o)}\ t_q^2 + O(t_q^3) \\ &= -\langle \mathbf{grad}\, E,\ \mathbf{grad}\, E\rangle_{c_q(o)}\ t_q + \\ &\quad + \tfrac{1}{2}\langle H\ \mathbf{grad}\, E,\ \mathbf{grad}\, E\rangle_{c_q(o)}\ t_q^2 + O(t_q^3)\ , \end{aligned}$$

and find

$$E(x_{q+1}) - E(x_q) = -\frac{1}{2}\frac{\langle \mathbf{grad}\, E,\ \mathbf{grad}\, E\rangle^2_{x_q N}}{\langle H\ \mathbf{grad}\, E,\ \mathbf{grad}\, E\rangle_{x_q N}} + O(||\mathbf{grad}\, E||^3_{x_q N}\ ). \tag{4.77}$$

By assuming that x_q and $\hat{x}$ are connected by a geodesic $c(s)$ with $c(o) = x_q$ and $c(\hat{s}) = \hat{x}$, we can write

$$E(c(\hat{s})) - E(c(o)) = \langle \mathbf{grad}\, E,\ W\rangle_{x_q}\ \hat{s} + \tfrac{1}{2}\langle H\,W,\ W\rangle_{x_q}\ \hat{s}^2 + O\ (\hat{s}^3) \tag{4.78}$$

where $\nabla_W W = 0$ and $W = c_*(\frac{d}{ds})$.
Since $\mathbf{grad}\, E(\hat{x}) = 0$, we have for an arbitrary parallel field U (i.e. $\nabla_W U = 0$) along $c(s)$,

$$0 = \langle \mathbf{grad}\, E,\ U\rangle_{\hat{x}} = \langle \mathbf{grad}\, E,\ U\rangle_{x_q} + \langle H\,W,\ U\rangle_{x_q}\ \hat{s} + O(\hat{s}^2).$$

Hence,

$$\hat{s} W_{x_q} = -H^{-1}\mathbf{grad}\, E(x_q) + O(||\mathbf{grad}\, E||^2_{x_q N}\ ).$$

Substitution into (4.78) gives then

$$E(\hat{x}) - E(x_q) = -\tfrac{1}{2}\langle H^{-1}\mathbf{grad}\, E,\ \mathbf{grad}\, E\rangle_{x_q N} + O(||\mathbf{grad}\, E||^3_{x_q N}\ ). \tag{4.79}$$

And subtracting this from (4.77) gives

$$E(x_{q+1}) - E(\hat{x}) =$$

$$\left(1 - \frac{\langle \mathbf{grad}\, E,\ \mathbf{grad}\, E\rangle^2_{x_q}}{\langle H\ \mathbf{grad}\, E,\ \mathbf{grad}\, E\rangle_{x_q}\ \langle H^{-1}\mathbf{grad}\, E,\ \mathbf{grad}\, E\rangle_{x_q}}\right)\left(E(x_q) - E(\hat{x}) + O(||\mathbf{grad}\, E||^3_{x_q N}\ )\right),$$

or

$$\frac{E(x_{q+1})-E(\hat{x})}{E(x_q)-E(\hat{x})} = \left[1 - \frac{\langle \Delta x,\Delta x\rangle^2_{x_q}}{\langle H\Delta x,\Delta x\rangle_{x_q}\,\langle H^{-1}\Delta x,\Delta x\rangle_{x_q}}\right]\left[1 + O(||\Delta x||_{x_q N})\right], \tag{4.80}$$

with $\Delta x(x_q) = -\,\mathbf{grad}\, E(x_q)$.
By assuming that $\hat{x}$ is a strict local minimum of

$$E(x) = \frac{1}{2}||y_s - y(x)||^2_M,$$

we can now apply Kantorovich' inequality to (4.80). Kantorovich' inequality (see Rao, 1973, p. 74) states namely that if a linear map

$$A: T_x N \to T_x N$$

is positive definite and selfadjoint with eigenvalues

$$\lambda^n \geq \lambda^{n-1} \geq \ldots \geq \lambda^1 > 0,$$

then

$$1 \leq \langle A\Delta x,\Delta x\rangle_N \langle A^{-1}\Delta x,\Delta x\rangle_N \leq \frac{1}{4}\left[\left(\frac{\lambda^1}{\lambda^n}\right)^{\frac{1}{2}} + \left(\frac{\lambda^n}{\lambda^1}\right)^{\frac{1}{2}}\right]^2, \tag{4.81}$$

for all normalized $\Delta x \in T_x N$, i.e. $\langle \Delta x,\Delta x\rangle_N = 1$.
Since the eigenvalues of the linear map H read

$$\lambda^r = 1 - k^r_N ||y_s - y||_M, \qquad r=1,\ldots,n,$$

application of (4.81) to (4.80) finally gives the desired result (4.67).

5. Supplements and examples

In this section we will consider some examples to illustrate the theory developed in the previous sections. Apart from the examples, we also present new results on the Helmert transformation and give some suggestions as to how to estimate the extrinsic curvatures.

5.1. The two dimensional Helmert transformation

In subsection 3.6 we have seen that the solution of the Helmert transformation only admitting a

rotation could be found by orthogonally projecting the observation point onto a circle with radius equalling the square root of the moment of inertia of the network involved. We will now generalize this result and consider the full Helmert transformation. That is, we will assume the scale- and translation parameters to be included as well.

Of couse, the solution to the two dimensional Helmert transformation is well known (see e.g. Köchle, 1982). It is therefore not so much our purpose to present the solution, but to emphasize the geometry involved. And the method chosen for deriving the solution prepares us for the case considered in our next example.

The model of the Helmert transformation reads

$$\begin{aligned} x_i &= u_i\lambda \cos\theta + v_i\lambda \sin\theta + t_x + e_{x_i} \\ y_i &= -u_i\lambda \sin\theta + v_i\lambda \cos\theta + t_y + e_{y_i} \end{aligned} , \tag{5.1}$$

where:

- $i = 1,\ldots,n$ = number of points,
- x_i, y_i are the cartesian coordinates of the network points in the first coordinate system, and
- u_i, v_i are the coordinates in the second coordinate system,
- λ, θ, t_x and t_y are respectively the scale, orientation and translation parameters, which need to be estimated, and
- e_{x_i}, e_{y_i} are the errors to be minimized in the 2-norm.

If we write model (5.1) as

$$\mathbf{y_s} = \lambda \cos\theta\, \mathbf{x_1} + \lambda \sin\theta\, \mathbf{x_2} + t_x\mathbf{x_3} + t_y\mathbf{x_4} + \mathbf{e}, \tag{5.2}$$

where:

$$\begin{aligned} \mathbf{y_s} &= (x_1,y_1,\ldots,x_n,y_n)^t, \quad \mathbf{e} = (e_{x_1},e_{y_1},\ldots,e_{x_n},e_{y_n})^t, \\ \mathbf{x_1} &= (u_1,v_1,\ldots,u_n,v_n)^t, \quad \mathbf{x_2} = (v_1,-u_1,\ldots,v_n,-u_n)^t, \\ \mathbf{x_3} &= (1,\ 0,\ldots,\ 1,\ 0)^t, \quad \mathbf{x_4} = (0,\ 1,\ldots,\ 0,\ 1)^t, \end{aligned} \tag{5.2'}$$

our least-squares problem becomes

$$\min_{\lambda,\theta,t_x,t_y} E(\lambda,\theta,t_x,t_y) = \min_{\lambda,\theta,t_x,t_y} \|\mathbf{y_s} - \lambda\cos\theta\, \mathbf{x_1} - \lambda\sin\theta\, \mathbf{x_2} - t_x\mathbf{x_3} - t_y\mathbf{x_4}\|_M^2. \tag{5.3}$$

We shall solve (5.3) by proceeding in two steps. First we assume λ and θ fixed and solve the subproblem

$$\min_{t_x,t_y} E(\lambda,\theta,t_x,t_y). \tag{5.4}$$

Let $t_x(\lambda,\theta)$, $t_y(\lambda,\theta)$ denote the solution to (5.4) and formulate the second problem as

$$\min_{\lambda,\theta} E(\lambda,\theta,t_x(\lambda,\theta),t_y(\lambda,\theta)). \tag{5.5}$$

Let $\hat{\lambda},\hat{\theta}$ denote the solution of (5.5). The overall solution of our original least-squares problem (5.3) is then

$$\hat{\lambda},\hat{\theta},t_x(\hat{\lambda},\hat{\theta}),\ t_y(\hat{\lambda},\hat{\theta}). \tag{5.6}$$

By taking this two-step procedure we have separated our original four-dimensional least-squares problem (5.3) into two two-dimensional least-squares problems (5.4) and (5.5).
With the abbreviation

$$\mathbf{y}_s(\lambda,\theta) = \mathbf{y}_s - \lambda \cos\theta\ \mathbf{x}_1 - \lambda \sin\theta\ \mathbf{x}_2, \tag{5.7}$$

the first subproblem (5.4) becomes

$$\min_{t_x,t_y} E(\lambda,\theta,t_x,t_y) = \min_{t_x,t_y} ||\mathbf{y}_s(\lambda,\theta) - t_x\mathbf{x}_3 - t_y\mathbf{x}_4||^2_M. \tag{5.8}$$

And geometrically this problem can of course be seen as the problem of finding the point in the plane spanned by the orthogonal vectors $\mathbf{x}_3$ and $\mathbf{x}_4$ (as before we assume that the observation space is endowed with the standard metric) which is nearest to $\mathbf{y}_s(\lambda,\theta)$. (see figure 30).

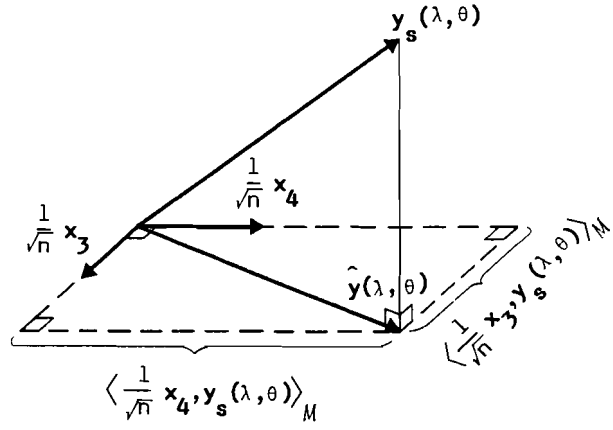


figure 30

Since the two vectors $\frac{1}{\sqrt{n}}\mathbf{x}_3$ and $\frac{1}{\sqrt{n}}\mathbf{x}_4$ are orthonormal, it follows that the point in the plane spanned by $\mathbf{x}_3$ and $\mathbf{x}_4$ closest to $\mathbf{y}_s(\lambda,\theta)$ is

$$\hat{\mathbf{y}}(\lambda,\theta) = \langle \tfrac{1}{\sqrt{n}}\mathbf{x}_3, \mathbf{y}_s(\lambda,\theta)\rangle_M \tfrac{1}{\sqrt{n}}\mathbf{x}_3 + \langle \tfrac{1}{\sqrt{n}}\mathbf{x}_4, \mathbf{y}_s(\lambda,\theta)\rangle_M \tfrac{1}{\sqrt{n}}\mathbf{x}_4.$$

Hence,

$$t_x(\lambda,\theta) = \frac{1}{n}\langle x_3, y_s(\lambda,\theta)\rangle_M \quad , \quad t_y(\lambda,\theta) = \frac{1}{n}\langle x_4, y_s(\lambda,\theta)\rangle_M \ ,$$

or with (5.7)

$$\begin{aligned} t_x(\lambda,\theta) &= \frac{1}{n}\langle x_3, y_s - \lambda\cos\theta\, x_1 - \lambda\sin\theta\, x_2\rangle_M \ , \\ t_y(\lambda,\theta) &= \frac{1}{n}\langle x_4, y_s - \lambda\cos\theta\, x_1 - \lambda\sin\theta\, x_2\rangle_M \ . \end{aligned} \tag{5.9}$$

This concludes the first step.

To solve (5.5), we substitute (5.9) into (5.3) and find

$$\min_{\lambda,\theta} E(\lambda,\theta,t_x(\lambda,\theta),t_y(\lambda,\theta)) = \min_{\lambda,\theta} \| y_s^c - \lambda\cos\theta\, x_1^c - \lambda\sin\theta\, x_2^c \|_M^2 , \tag{5.10}$$

where:

$$\left.\begin{aligned} y_s^c &= y_s - \frac{1}{n}\langle x_3, y_s\rangle_M\, x_3 - \frac{1}{n}\langle x_4, y_s\rangle_M\, x_4 \\ x_1^c &= x_1 - \frac{1}{n}\langle x_3, x_1\rangle_M\, x_3 - \frac{1}{n}\langle x_4, x_1\rangle_M\, x_4 \\ x_2^c &= x_2 - \frac{1}{n}\langle x_3, x_2\rangle_M\, x_3 - \frac{1}{n}\langle x_4, x_2\rangle_M\, x_4 \end{aligned}\right\} \tag{5.10'}$$

The geometry of problem (5.10) is illustrated in figure 31.

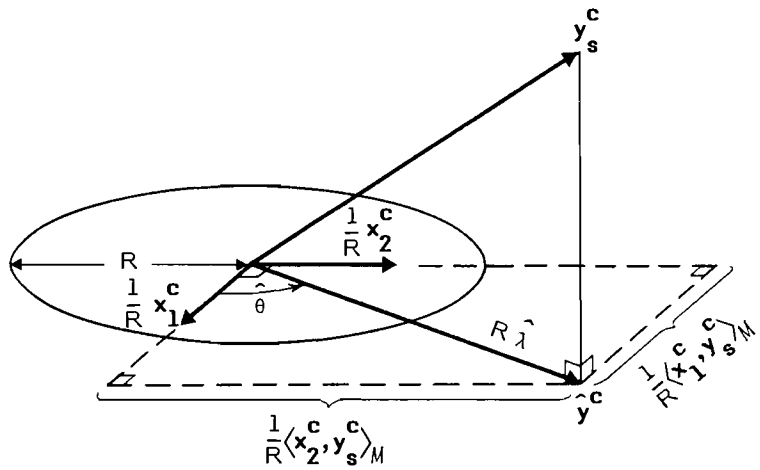


figure 31

Since the two vectors $\frac{1}{R}x_1^c$ and $\frac{1}{R}x_2^c$, with $R = \|x_1^c\| = \|x_2^c\|$, are orthonormal, it follows that the point in the plane spanned by x_1^c and x_2^c closest to y_s^c is

$$\hat{y}^c = \langle \tfrac{1}{R}x_1^c, y_s^c\rangle_M \tfrac{1}{R}x_1^c + \langle \tfrac{1}{R}x_2^c, y_s^c\rangle_M \tfrac{1}{R}x_2^c .$$

Hence,

$$\hat{\lambda}\cos\hat{\theta} = \frac{1}{R^2}\langle x_1^c, y_s^c\rangle_M = \frac{\langle x_1^c, y_s^c\rangle_M}{\langle x_1^c, x_1^c\rangle_M}, \quad \hat{\lambda}\sin\hat{\theta} = \frac{1}{R^2}\langle x_2^c, y_s^c\rangle_M = \frac{\langle x_2^c, y_s^c\rangle_M}{\langle x_2^c, x_2^c\rangle_M} \tag{5.11}$$

or

$$\hat{\theta} = \tan^{-1}\frac{\langle x_2^c, y_s^c\rangle_M}{\langle x_1^c, y_s^c\rangle_M}, \qquad \hat{\lambda} = \sqrt{\frac{\langle x_1^c, y_s^c\rangle_M^2}{\langle x_1^c, x_1^c\rangle_M^2} + \frac{\langle x_2^c, y_s^c\rangle_M^2}{\langle x_2^c, x_2^c\rangle_M^2}}.$$

With (5.10') and (5.2') this can be written as

$$\hat{\theta} = \tan^{-1}\frac{\sum_{i=1}^{n} v_i^c x_i^c - u_i^c y_i^c}{\sum_{i=1}^{n} u_i^c x_i^c + v_i^c y_i^c}, \quad \hat{\lambda} = \frac{\sqrt{\left(\sum_{i=1}^{n} u_i^c x_i^c + v_i^c y_i^c\right)^2 + \left(\sum_{i=1}^{n} v_i^c x_i^c - u_i^c y_i^c\right)^2}}{\sum_{i=1}^{n} (u_i^c)^2 + (v_i^c)^2},$$

$$\text{where: } x_i^c = x_i - \frac{\sum_{j=1}^{n} x_j}{n}, \; y_i^c = y_i - \frac{\sum_{j=1}^{n} y_j}{n}, \; u_i^c = u_i - \frac{\sum_{j=1}^{n} u_j}{n}, \; v_i^c = v_i - \frac{\sum_{j=1}^{n} v_j}{n}. \tag{5.12}$$

To find the least-squares solution for the translation parameters we substitute (5.11) into (5.9) and find

$$\hat{t}_x = t_x(\hat{\lambda},\hat{\theta}) = \frac{1}{n}\left\langle x_3, y_s - \frac{\langle x_1^c, y_s^c\rangle_M}{\langle x_1^c, x_1^c\rangle_M} x_1 - \frac{\langle x_2^c, y_s^c\rangle_M}{\langle x_2^c, x_2^c\rangle_M} x_2 \right\rangle_M,$$

$$\hat{t}_y = t_y(\hat{\lambda},\hat{\theta}) = \frac{1}{n}\left\langle x_4, y_s - \frac{\langle x_1^c, y_s^c\rangle_M}{\langle x_1^c, x_1^c\rangle_M} x_1 - \frac{\langle x_2^c, y_s^c\rangle_M}{\langle x_2^c, x_2^c\rangle_M} x_2 \right\rangle_M.$$

With (5.10') and (5.2') this gives

$$\hat{t}_x = x^c - \frac{\sum_{i=1}^{n} u_i^c x_i^c + v_i^c y_i^c}{\sum_{i=1}^{n} (u_i^c)^2 + (v_i^c)^2} u^c - \frac{\sum_{i=1}^{n} v_i^c x_i^c - u_i^c y_i^c}{\sum_{i=1}^{n} (u_i^c)^2 + (v_i^c)^2} v^c,$$

$$\hat{t}_y = y^c - \frac{\sum_{i=1}^{n} u_i^c x_i^c + v_i^c y_i^c}{\sum_{i=1}^{n} (u_i^c)^2 + (v_i^c)^2} v^c + \frac{\sum_{i=1}^{n} v_i^c x_i^c - u_i^c y_i^c}{\sum_{i=1}^{n} (u_i^c)^2 + (v_i^c)^2} u^c, \tag{5.13}$$

$$\text{where: } x^c = \frac{\sum_{i=1}^{n} x_i}{n}, \; y^c = \frac{\sum_{i=1}^{n} y_i}{n}, \; u^c = \frac{\sum_{i=1}^{n} u_i}{n}, \; v^c = \frac{\sum_{i=1}^{n} v_i}{n}.$$

(5.12) together with (5.13) constitute the well known solution of the two dimensional Helmert transformation (see e.g. Köchle, 1982).

Note that although the functions occurring in model (5.1) are non-linear, the actual adjustment problem is linear. That is, the submanifold $\bar{N}$ as described by model (5.1) is a typical example of a totally geodesic manifold. The non-linearity present in (5.1) effects therefore only the inverse mapping problem. This follows from (5.11) if one solves for the parameters λ and θ.

Also note that we are by no means restricted to the particular two-step procedure chosen in (5.4) and (5.5). Instead of taking the above two-step procedure, we could for instance have decided to only fix θ first. In the first step we would then have to solve for $\lambda(\theta)$, $t_x(\theta)$ and $t_y(\theta)$. And this is still a linear adjustment problem. But when solving for θ in the second step, we would get a non-linear adjustment problem namely that of orthogonally projecting onto a circle. Hence we see that where we started with an essentially linear adjustment problem we end up with two subproblems of which the second is non-linear. What has happened is of course that by fixing θ we have chosen to project onto a non-linear curve lying in an otherwise flat manifold. Thus generally speaking such a step procedure would not be very recommendable since it produces a non-linear problem out of a linear one. An interesting point is, however, that if we reverse the argument one should be able in some cases to get a linear subproblem out of an essentially non-linear problem by applying the appropriate step procedure. Think for instance of parametrized curved submanifolds which have linear, i.e. straight, coordinate lines. In the following we will consider a typical class of such manifolds.

5.2. Orthogonal projection onto a ruled surface

A **ruled surface** is a surface which has the property that through every point of the surface there passes a straight line which lies entirely in the surface. Thus the surface is covered by straight lines, called **rulings** which form a family depending on one parameter.

In order to find a parametrization of a ruled surface choose on the surface a curve transversal to the rulings. Let this curve be given by $\mathbf{c}(t_1)$. At any point of this curve take a vector $\mathbf{T}$ of the ruling which passes through this point. This vector obviously depends on t_1. Thus we have $\mathbf{T}(t_1)$.

Now we can write the equation of the surface as

$$\mathbf{y}(t_1, t_2) = \mathbf{c}(t_1) + t_2\mathbf{T}(t_1) \ . \tag{5.14}$$

The parameter t_1 indicates the ruling on the surface, and the parameter t_2 shows the position on the ruling.

If in an adjustment context the submanifold $\bar{N}$ turns out to be a ruled surface, one can expect to take advantage of the special properties of $\bar{N}$. $\bar{N}$ will namely be flat in the directions of the rulings, whilst curved in the directions transversal to it. Hence, it might turn out to be advantageous to perform the adjustment in two steps. In the first step one would then solve for a **linear** least-squares adjustment problem, and in the second step for a non-linear adjustment problem of a **reduced dimension.** For the ruled surface (5.14) for instance, we would choose a point on the curve

$\mathbf{c}(t_1)$, $\mathbf{c}(t_1^o)$ say. To this point there corresponds a ruling with direction $\mathbf{T}(t_1^o)$. The linear least-squares adjustment step consists then of orthogonally projecting the observation point $\mathbf{y_s}$ onto the ruling given by

$$\mathbf{y}(t_2) = \mathbf{c}(t_1^o) + t_2\mathbf{T}(t_1^o) \quad . \tag{5.15}$$

As solution we get an adjusted point on the surface which depends on the choice of ruling, i.e. on the choice t_1^o:

$$\begin{aligned} \hat{\mathbf{y}}(t_1^o) &= \mathbf{c}(t_1^o) + \hat{t}_2(t_1^o)\mathbf{T}(t_1^o) = \\ &= \mathbf{c}(t_1^o) + \langle \mathbf{T}(t_1^o), \mathbf{T}(t_1^o) \rangle_M^{-1} \langle \mathbf{T}(t_1^o), \mathbf{y_s} - \mathbf{c}(t_1^o) \rangle_M \mathbf{T}(t_1^o) \quad . \end{aligned} \tag{5.16}$$

The second step consists then of orthogonally projecting $\mathbf{y_s}$ onto the curve given by (5.16). This problem is of course in general still non-linear, but it has the advantage of being of a smaller dimension than the original adjustment problem.
As an example one could think of a cylinder (this is in fact a very special ruled surface, since it is developable). Then we have (see figure 32):

$$c^{i=1}(t_1) = R \cos(t_1), \qquad c^{i=2}(t_1) = R \sin(t_1), \qquad c_1^{i=3}(t_1) = 0,$$
$$T^{i=1}(t_1) = 0, \qquad T^{i=2}(t_1) = 0, \qquad T^{i=3}(t_1) = 1.$$

In the first step we would choose t_1^o. This would give us then

$$\hat{\mathbf{y}}(t_1^o) = \mathbf{c}(t_1^o) + y_s^{i=3}\mathbf{T} \quad .$$

For the second step we would then need to minimize

$$\min_{t_1^o} . \langle \mathbf{y_s} - y_s^{i=3}\mathbf{T} - \mathbf{c}(t_1^o), \mathbf{y_s} - y_s^{i=3}\mathbf{T} - \mathbf{c}(t_1^o) \rangle_M \quad .$$

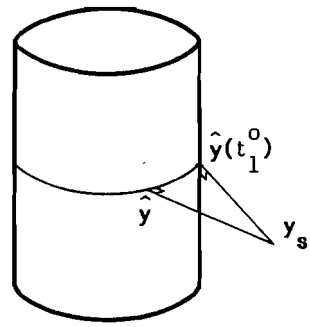


figure 32

It will be clear that the above described procedure also holds for ruled-type of manifolds.

5.3. The two dimensional Symmetric Helmert transformation

As a nice application of the idea described in the previous example we have what we shall call the two dimensional **Symmetric** Helmert transformation.

Recall the model of the two dimensional Helmert transformation (see (5.1)) and note that the model in its classical formulation favours one point field above the other. This can also be seen from the rather asymmetric solution of the scale parameter (see (5.12)).

It has bothered the present author for some time that one was satisfied with the classical formulation (5.1). A better formulation would namely be:

$$\left.\begin{aligned} \tilde{x}_i &= u_i\lambda \cos\theta + v_i\lambda \sin\theta + t_x \\ \tilde{y}_i &= -u_i\lambda \sin\theta + v_i\lambda \cos\theta + t_y \\ \tilde{\bar{x}}_i &= u_i \\ \tilde{\bar{y}}_i &= v_i \end{aligned}\right\} \qquad (5.17)$$

where:

- $i = 1,\ldots,n$ = number of points,
- the tilde "$\sim$" sign stands for the mathematical expectation,
- x_i, y_i and $\bar{x}_i$, $\bar{y}_i$ are the "observed" cartesian coordinates of the network points in the two coordinate systems,
- λ, θ, t_x and t_y are the transformation parameters which need to be estimated, and
- u_i, v_i are the cartesian coordinates which need to be estimated.

Of course, the submanifold as described by the classical Helmert transformation is totally geodesic. Hence, one could fear in the first instance that (5.17) can only be solved iteratively, i.e. through the process of linearization. However, in this example we will show that if one views model (5.17) as a **ruled-type of manifold,** one can in fact find its least-squares solution also analytically.

Note that if we fix u_i, v_i, $i=1,\ldots,n$, in (5.17) we are back at the classical Helmert transformation, which was linear. Hence, manifold $\bar{N}$ as described by (5.17) is flat in directions transversal to the u_i,v_i-coordinate lines. But if we fix λ and θ, we see that it is also flat in the directions transversal to the λ,θ- coordinate lines. Thus in the first adjustment step we can either fix the u_i,v_i, $i=1,\ldots,n$, or λ and θ. It turns out that the choice of fixing λ and θ is the most advantageous one.

Skipping the tedious but trivial adjustment derivation we find for fixed λ and θ the solution of the first adjustment step as:

$$\left.\begin{aligned} \hat{u}_i(\lambda,\theta) &= \bar{x}_c + (1+\lambda^2)^{-1}(\bar{x}_i^c + x_i^c\lambda \cos\theta - y_i^c \lambda \sin\theta), \\ \hat{v}_i(\lambda,\theta) &= \bar{y}_c + (1+\lambda^2)^{-1}(\bar{y}_i^c + x_i^c\lambda \sin\theta + y_i^c \lambda \cos\theta), \\ \hat{t}_x(\lambda,\theta) &= \bar{x}_c - \bar{x}_c \lambda \cos\theta - \bar{y}_c \lambda \sin\theta, \\ \hat{t}_y(\lambda,\theta) &= \bar{y}_c + \bar{x}_c \lambda \sin\theta - \bar{y}_c \lambda \cos\theta, \end{aligned}\right\} \qquad (5.18)$$

where:

$$\left.\begin{aligned} \bar{x}_c &= n^{-1}\sum_{j=1}^{n}\bar{x}_j, \quad \bar{y}_c = n^{-1}\sum_{j=1}^{n}\bar{y}_j, \quad x_c = n^{-1}\sum_{j=1}^{n}x_j, \quad y_c = n^{-1}\sum_{j=1}^{n}y_j, \\ \bar{x}_i^c &= \bar{x}_i - \bar{x}_c, \quad \bar{y}_i^c = \bar{y}_i - \bar{y}_c, \quad x_i^c = x_i - x_c, \quad y_i^c = y_i - y_c. \end{aligned}\right\} \quad (5.18')$$

Hence, for the second adjustment step we get

$$\left.\begin{aligned} x_i &= x_c + (1+\lambda^2)^{-1}(\lambda^2 x_i^c + \bar{x}_i^c \lambda \cos\theta + \bar{y}_i^c \lambda \sin\theta) + e_{x_i}, \\ y_i &= y_c + (1+\lambda^2)^{-1}(\lambda^2 y_i^c - \bar{x}_i^c \lambda \sin\theta + \bar{y}_i^c \lambda \cos\theta) + e_{y_i}, \\ \bar{x}_i &= \bar{x}_c + (1+\lambda^2)^{-1}(\bar{x}_i^c + x_i^c \lambda \cos\theta - y_i^c \lambda \sin\theta) + e_{\bar{x}_i}, \\ \bar{y}_i &= \bar{y}_c + (1+\lambda^2)^{-1}(\bar{y}_i^c + x_i^c \lambda \sin\theta + y_i^c \lambda \cos\theta) + e_{\bar{y}_i}, \end{aligned}\right\} \quad (5.19)$$

or

$$\begin{pmatrix} e_{x_i} \\ e_{y_i} \end{pmatrix} = (1+\lambda^2)^{-1}\left[\begin{pmatrix} x_i^c \\ y_i^c \end{pmatrix} - \lambda \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix}\begin{pmatrix} \bar{x}_i^c \\ \bar{y}_i^c \end{pmatrix}\right], \quad (5.19')$$

$$\begin{pmatrix} e_{\bar{x}_i} \\ e_{\bar{y}_i} \end{pmatrix} = -\lambda(1+\lambda^2)^{-1}\begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}\left[\begin{pmatrix} x_i^c \\ y_i^c \end{pmatrix} - \lambda \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix}\begin{pmatrix} \bar{x}_i^c \\ \bar{y}_i^c \end{pmatrix}\right],$$

where e are the residuals.

The sum of the squared residuals reads then:

$$\sum_{i=1}^{n}(e_{x_i}^2 + e_{y_i}^2 + e_{\bar{x}_i}^2 + e_{\bar{y}_i}^2) = \quad (5.20)$$

$$= (1+\lambda^2)^{-1}\sum_{i=1}^{n}\left[\begin{pmatrix} x_i^c \\ y_i^c \end{pmatrix} - \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix}\begin{pmatrix} \lambda\bar{x}_i^c \\ \lambda\bar{y}_i^c \end{pmatrix}\right]^t \left[\begin{pmatrix} x_i^c \\ y_i^c \end{pmatrix} - \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix}\begin{pmatrix} \lambda\bar{x}_i^c \\ \lambda\bar{y}_i^c \end{pmatrix}\right].$$

And this function needs to be minimized in order to find the least-squares estimates $\hat{\lambda}$ and $\hat{\theta}$. The function is of course still non-linear (and non-quadratic) in λ and θ. However, observe that if we fix $\lambda = 1$, the model underlying the function of (5.20) equals, apart from the fact that we are dealing here with coordinates referring to the centres of gravity, the Helmert transformation (3.30) admitting only a rotation. Hence, for an arbitrarily fixed value of λ the minimum $\hat{\theta}(\lambda)$ of (5.20) follows readily from (3.36) as

$$\hat{\theta}(\lambda) = \tan^{-1} \frac{\sum_{i=1}^{n} (\lambda \bar{y}_i^c x_i^c - \lambda \bar{x}_i^c y_i^c)}{\sum_{i=1}^{n} (\lambda \bar{x}_i^c x_i^c + \lambda \bar{y}_i^c y_i^c)} = \tan^{-1} \frac{\sum_{i=1}^{n} (\bar{y}_i^c x_i^c - \bar{x}_i^c y_i^c)}{\sum_{i=1}^{n} (\bar{x}_i^c x_i^c + \bar{y}_i^c y_i^c)} \quad . \tag{5.21}$$

Note that not too surprisingly the estimated rotation angle is invariant to scale changes.
From substituting (5.21) into (5.20) we find

$$f(\lambda) = (1 + \lambda^2)^{-1} \sum_{i=1}^{n} \left(\begin{bmatrix} x_i^c \\ y_i^c \end{bmatrix} - \lambda \begin{bmatrix} \cos\hat{\theta} & \sin\hat{\theta} \\ -\sin\hat{\theta} & \cos\hat{\theta} \end{bmatrix} \begin{bmatrix} \bar{x}_i^c \\ \bar{y}_i^c \end{bmatrix} \right)^t \left(\begin{bmatrix} x_i^c \\ y_i^c \end{bmatrix} - \lambda \begin{bmatrix} \cos\hat{\theta} & \sin\hat{\theta} \\ -\sin\hat{\theta} & \cos\hat{\theta} \end{bmatrix} \begin{bmatrix} \bar{x}_i^c \\ \bar{y}_i^c \end{bmatrix} \right) \tag{5.22}$$

which needs to be minimized in order to find $\hat{\lambda}$.
With the reparametrization

$$\lambda = \tan\phi, \quad 0 < \phi < \frac{1}{2}\pi \quad , \tag{5.23}$$

we can write (5.22) also as

$$f(\phi) = \langle \cos\phi\ \mathbf{e}_1 + \sin\phi\ \mathbf{e}_2, \cos\phi\ \mathbf{e}_1 + \sin\phi\ \mathbf{e}_2 \rangle_M \quad , \tag{5.24}$$

where:

$$\mathbf{e}_1 = (x_1^c, y_1^c, \ldots, x_n^c, y_n^c)^t \quad ,$$

$$\mathbf{e}_2 = (-\cos\hat{\theta}\ \bar{x}_1^c - \sin\hat{\theta}\ \bar{y}_1^c, \sin\hat{\theta}\ \bar{x}_1^c - \cos\hat{\theta}\ \bar{y}_1^c, \ldots, -\cos\hat{\theta}\ \bar{x}_n^c - \sin\hat{\theta}\ \bar{y}_n^c, \sin\hat{\theta}\ \bar{x}_n^c - \cos\hat{\theta}\ \bar{y}_n^c)^t \quad . \tag{5.24'}$$

Observe that the function $f(\phi)^{\frac{1}{2}}$ describes the distance from the origin to an ellipse lying in the plane spanned by the vectors $\mathbf{e}_1$ and $\mathbf{e}_2$. Hence, to minimize $f(\phi)$ we need to find that point on the ellipse

$$\mathbf{y}(\phi) = \cos\phi\ \mathbf{e}_1 + \sin\phi\ \mathbf{e}_2 \quad ,$$

which is closest to the origin. This minimization problem results then in the following eigenvalue problem

$$\begin{bmatrix} \langle \mathbf{e}_1, \mathbf{e}_1 \rangle_M & \langle \mathbf{e}_1, \mathbf{e}_2 \rangle_M \\ \langle \mathbf{e}_2, \mathbf{e}_1 \rangle_M & \langle \mathbf{e}_2, \mathbf{e}_2 \rangle_M \end{bmatrix} \begin{bmatrix} \cos\phi \\ \sin\phi \end{bmatrix} = \mu \begin{bmatrix} \cos\phi \\ \sin\phi \end{bmatrix}, \tag{5.25}$$

And the minimum of $f(\phi)$ equals the smallest eigenvalue $\mu_{min.}$ of (5.25). The eigenvalues of (5.25) follow from

$$\begin{vmatrix} \langle \mathbf{e}_1, \mathbf{e}_1 \rangle_M - \mu & \langle \mathbf{e}_1, \mathbf{e}_2 \rangle_M \\ \langle \mathbf{e}_2, \mathbf{e}_1 \rangle_M & \langle \mathbf{e}_2, \mathbf{e}_2 \rangle_M - \mu \end{vmatrix} = 0 \quad ,$$

as

$$\mu = \frac{\langle e_1,e_1\rangle + \langle e_2,e_2\rangle \pm \sqrt{(\langle e_1,e_1\rangle + \langle e_2,e_2\rangle)^2 - 4(\langle e_1,e_1\rangle\langle e_2,e_2\rangle - (\langle e_1,e_2\rangle)^2)}}{2}.$$

Hence,

$$\mu_{min.} = \frac{\langle e_1,e_1\rangle+\langle e_2,e_2\rangle - \sqrt{(\langle e_1,e_1\rangle + \langle e_2,e_2\rangle)^2 - 4(\langle e_1,e_1\rangle\langle e_2,e_2\rangle - (\langle e_1,e_2\rangle)^2)}}{2}$$

Substitution of $\mu = \mu_{min}$ into (5.25) gives

$$\tan\phi = \frac{(\langle e_2,e_2\rangle-\langle e_1,e_1\rangle) - \sqrt{(\langle e_1,e_1\rangle+\langle e_2,e_2\rangle)^2 - 4(\langle e_1,e_1\rangle\langle e_2,e_2\rangle-(\langle e_1,e_2\rangle)^2)}}{2\langle e_1,e_2\rangle},$$

or with (5.23) and $\langle e_1,e_2\rangle = \operatorname{sgn}[\langle e_1,e_2\rangle]\,|\langle e_1,e_2\rangle|$

$$\hat{\lambda} = \frac{(\langle e_2,e_2\rangle - \langle e_1,e_1\rangle)}{2\langle e_1,e_2\rangle} + - \operatorname{sgn}[\langle e_1,e_2\rangle]\sqrt{\left(\frac{\langle e_2,e_2\rangle - \langle e_1,e_1\rangle}{2\langle e_1,e_2\rangle}\right)^2 + 1}\,. \tag{5.26}$$

With (5.24'), (5.21) and (5.18) the least-squares solution of the two dimensional Symmetric Helmert transformation (5.17) finally becomes:

$$\hat{\lambda} = \frac{\sum_{i=1}^{n}((x_i^c)^2 + (y_i^c)^2) - \sum_{i=1}^{n}((\bar{x}_i^c)^2 + (\bar{y}_i^c)^2)}{2\sqrt{(\sum_{i=1}^{n}(\bar{x}_i^c x_i^c + \bar{y}_i^c y_i^c))^2 + (\sum_{i=1}^{n}(\bar{y}_i^c x_i^c - \bar{x}_i^c y_i^c))^2}} +$$

$$+\sqrt{1 + \left(\frac{\sum_{i=1}^{n}((x_i^c)^2 + (y_i^c)^2) - \sum_{i=1}^{n}((\bar{x}_i^c)^2 + (\bar{y}_i^c)^2)}{2\sqrt{(\sum_{i=1}^{n}(\bar{x}_i^c x_i^c + \bar{y}_i^c y_i^c))^2 + (\sum_{i=1}^{n}(\bar{y}_i^c x_i^c - \bar{x}_i^c y_i^c))^2}}\right)^2},$$

$$\hat{\theta} = \tan^{-1}\frac{\sum_{i=1}^{n}(\bar{y}_i^c x_i^c - \bar{x}_i^c y_i^c)}{\sum_{i=1}^{n}(\bar{x}_i^c x_i^c + \bar{y}_i^c y_i^c)}, \tag{5.27}$$

$$\hat{t}_x = x_c - \bar{x}_c\,\hat{\lambda}\cos\hat{\theta} - \bar{y}_c\,\hat{\lambda}\sin\hat{\theta},$$

$$\hat{t}_y = y_c + \bar{x}_c\,\hat{\lambda}\sin\hat{\theta} - \bar{y}_c\,\hat{\lambda}\cos\hat{\theta},$$

$$\hat{u}_i = \bar{x}_c + (1 + \hat{\lambda}^2)^{-1}(\bar{x}_i^c + x_i^c\,\hat{\lambda}\cos\hat{\theta} - y_i^c\,\hat{\lambda}\sin\hat{\theta}),$$

$$\hat{v}_i = \bar{y}_c + (1 + \hat{\lambda}^2)^{-1}(\bar{y}_i^c + x_i^c\,\hat{\lambda}\sin\hat{\theta} + y_i^c\,\hat{\lambda}\cos\hat{\theta}).$$

Note that the reciprocal scale parameter reads as

$$\hat{\lambda}^{-1} = \frac{\sum_{i=1}^{p} ((\bar{x}_i^c)^2 + (\bar{y}_i^c)^2) - \sum_{i=1}^{p}((x_i^c)^2 + (y_i^c)^2)}{2\sqrt{(\sum_{i=1}^{p}(x_i^{c-c}x_i + y_i^{c-c}y_i))^2 + (\sum_{i=1}^{p}(y_i^{c-c}x_i - x_i^{c-c}y_i))^2}}$$

$$+\sqrt{1 + \left[\frac{\sum_{i=1}^{p}((\bar{x}_i^c)^2 + (\bar{y}_i^c)^2) - \sum_{i=1}^{p}((x_i^c)^2 + (y_i^c)^2)}{2\sqrt{(\sum_{i=1}^{p}(x_i^{c-c}x_i + y_i^{c-c}y_i))^2 + (\sum_{i=1}^{p}(y_i^{c-c}x_i - x_i^{c-c}y_i))^2}}\right]^2}\ ,$$

which demonstrates the symmetry in our least-squares solution of the scale parameter. This in contrast to solution (5.12) of the classical Helmert transformation.

5.4. The two dimensional Symmetric Helmert transformation with a rotational invariant covariance structure

Up til now we assumed the simplest structure possible for the covariance matrices of the observed cartesian coordinates. In many practical applications this assumption will do, but it will not be sufficient for all applications. Unfortunately one can not expect to find a solution like (5.27) if the observed coordinates are allowed to have an arbitrary covariance matrix. One of the reasons that the derivation of (5.27) went so smoothly is namely that the covariance matrices used for the two sets of coordinates are scaled versions of each other and are invariant to rotations. This indicates, however, that if we assume the covariance matrices Q and $\bar{Q}$ of the two coordinate sets $(\ldots x_i, y_i, \ldots)^t$ and $(\ldots \bar{x}_i, \bar{y}_i, \ldots)^t$ to be of such a structure that

$$k^2 Q = \bar{Q} \quad \text{for some } k \in \mathbb{R}^+, \tag{5.28}$$

and

$$R^t Q R = Q, \tag{5.29}$$

where R is a $2n \times 2n$ block diagonal matrix with equal 2×2 blocks

$$\begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix},$$

one should be able to generalize (5.27) accordingly.
Note, that it follows from (5.28) and (5.29) that $\bar{Q}^{-1}$ consists of 2×2 diagonal blocks of the type

$$\begin{pmatrix} d^{ij} & 0 \\ 0 & d^{ij} \end{pmatrix} . \tag{5.30}$$

Hence, the Baarda-Alberda criterium matrix (see e.g. Baarda, 1973 or Teunissen, 1984b) is a proper

candidate for $\bar{Q}$.

To solve for the Symmetric Helmert transformation with the new covariance structure (5.28), (5.29) we apply the same two-step procedure as used before.

For fixed λ and θ we get then as solution of the first step:

$$\left.\begin{aligned} u_i(\lambda,\theta) &= \bar{x}_c + (1 + (k\lambda)^2)^{-1}(\bar{x}_i^c + k^2\lambda \cos\theta\, x_i^c - k^2\lambda \sin\theta\, y_i^c), \\ v_i(\lambda,\theta) &= \bar{y}_c + (1 + (k\lambda)^2)^{-1}(\bar{y}_i^c + k^2\lambda \sin\theta\, x_i^c + k^2\lambda \cos\theta\, y_i^c), \\ t_x(\lambda,\theta) &= x_c - \lambda \cos\theta\, \bar{x}_c - \lambda \sin\theta\, \bar{y}_c, \\ t_y(\lambda,\theta) &= y_c + \lambda \sin\theta\, \bar{x}_c - \lambda \cos\theta\, \bar{y}_c, \end{aligned}\right\} \quad (5.31)$$

where:

$$\left.\begin{aligned} \bar{x}_c &= \Big(\sum_j \big(\sum_i d^{ij}\big)\Big)^{-1} \sum_i (d^{ij}\bar{x}_j); \quad \bar{y}_c = \Big(\sum_j \big(\sum_i d^{ij}\big)\Big)^{-1} \sum_i (d^{ij}\bar{y}_j); \\ x_c &= \Big(\sum_j \big(\sum_i d^{ij}\big)\Big)^{-1} \sum_i (d^{ij}x_j); \quad y_c = \Big(\sum_j \big(\sum_i d^{ij}\big)\Big)^{-1} \sum_i (d^{ij}y_j); \\ \bar{x}_i^c &= \bar{x}_i - \bar{x}_c, \quad \bar{y}_i^c = \bar{y}_i - \bar{y}_c, \quad x_i^c = x_i - x_c, \quad y_i^c = y_i - y_c. \end{aligned}\right\} \quad (5.31')$$

From this follows that we get for the second adjustment step:

$$\begin{pmatrix} e_{x_i} \\ e_{y_i} \end{pmatrix} = (1 + (k\lambda)^2)^{-1} \left[\begin{pmatrix} x_i^c \\ y_i^c \end{pmatrix} - \lambda \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix} \begin{pmatrix} \bar{x}_i^c \\ \bar{y}_i^c \end{pmatrix} \right], \quad (5.32)$$

$$\begin{pmatrix} \bar{e}_{x_i} \\ \bar{e}_{y_i} \end{pmatrix} = -k^2\lambda\,(1 + (k\lambda)^2)^{-1} \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix} \left[\begin{pmatrix} x_i^c \\ y_i^c \end{pmatrix} - \lambda \begin{pmatrix} \cos\theta & \cos\theta \\ -\sin\theta & \cos\theta \end{pmatrix} \begin{pmatrix} \bar{x}_i^c \\ \bar{y}_i^c \end{pmatrix} \right]$$

where e are the residuals.

Hence, the weighted sum of the squared residuals reads then

$$k^2(1+(k\lambda)^2)^{-1} \left[\begin{pmatrix} \vdots \\ x_i^c \\ y_i^c \\ \vdots \end{pmatrix} - \begin{pmatrix} \vdots & \vdots \\ -\bar{x}_i^c & -\bar{y}_i^c \\ -\bar{y}_i^c & -\bar{x}_i^c \\ \vdots & \vdots \end{pmatrix} \begin{pmatrix} \lambda\cos\theta \\ \lambda\sin\theta \end{pmatrix} \right]^t \begin{pmatrix} \ddots & & \vdots & \vdots & \\ & d^{i,i-1} & d^{i-1,i} & 0_{i-1,i} & \ddots \\ 0 & d^{ii} & 0 & d^{i,i+1} & 0 \\ & d^{i+1,i} & 0_{i+1,i} & d^{ii} & 0 & d^{i,i+1} \\ & 0 & \vdots & \vdots & \ddots \end{pmatrix} \left[\begin{pmatrix} \vdots \\ x_i^c \\ y_i^c \\ \vdots \end{pmatrix} - \begin{pmatrix} \vdots & \vdots \\ -\bar{x}_i^c & -\bar{y}_i^c \\ -\bar{y}_i^c & -\bar{x}_i^c \\ \vdots & \vdots \end{pmatrix} \begin{pmatrix} \lambda\cos\theta \\ \lambda\sin\theta \end{pmatrix} \right] \quad (5.33)$$

and this function needs to be minimized in order to find $\hat{\lambda}$ and $\hat{\theta}$.
With the reparametrization

$$k\lambda = \tan\phi, \qquad 0 < \phi < \frac{1}{2}\pi \tag{5.34}$$

we can rewrite (5.33) as

$$\min_{\theta,\phi} \begin{pmatrix} \sin\phi\cos\theta \\ \sin\phi\sin\theta \\ \cos\phi \end{pmatrix}^t \begin{pmatrix} (\bar{x}_i^c d^{ij} \bar{x}_j^c + \bar{y}_i^c d^{ij} \bar{y}_j^c) & 0 & -k(\bar{x}_i^c d^{ij} x_j^c + \bar{y}_i^c d^{ij} y_j^c) \\ 0 & (\bar{x}_i^c d^{ij} \bar{x}_j^c + \bar{y}_i^c d^{ij} \bar{y}_j^c) & -k(\bar{y}_i^c d^{ij} x_j^c - \bar{x}_i^c d^{ij} y_j^c) \\ -k(\bar{x}_i^c d^{ij} x_j^c + \bar{y}_i^c d^{ij} y_j^c) & -k(\bar{y}_i^c d^{ij} x_j^c - \bar{x}_i^c d^{ij} y_j^c) & k^2(x_i^c d^{ij} x_j^c + y_i^c d^{ij} y_j^c) \end{pmatrix} \begin{pmatrix} \sin\phi\cos\theta \\ \sin\phi\sin\theta \\ \cos\phi \end{pmatrix} \tag{5.35}$$

where we have used Einstein's summation convention.
The least-squares problem (5.35) results in the following eigenvalue problem:

$$\begin{pmatrix} (\bar{x}_i^c d^{ij} \bar{x}_j^c + \bar{y}_i^c d^{ij} \bar{y}_j^c) & 0 & -k(\bar{x}_i^c d^{ij} x_j^c + \bar{y}_i^c d^{ij} y_j^c) \\ 0 & (\bar{x}_i^c d^{ij} \bar{x}_j^c + \bar{y}_i^c d^{ij} \bar{y}_j^c) & -k(\bar{y}_i^c d^{ij} x_j^c - \bar{x}_i^c d^{ij} y_j^c) \\ -k(\bar{x}_i^c d^{ij} x_j^c + \bar{y}_i^c d^{ij} y_j^c) & -k(\bar{y}_i^c d^{ij} x_j^c - \bar{x}_i^c d^{ij} y_j^c) & k^2(x_i^c d^{ij} x_j^c + y_i^c d^{ij} y_j^c) \end{pmatrix} \begin{pmatrix} \sin\phi\cos\theta \\ \sin\phi\sin\theta \\ \cos\phi \end{pmatrix} = \mu \begin{pmatrix} \sin\phi\cos\theta \\ \sin\phi\sin\theta \\ \cos\phi \end{pmatrix} \tag{5.36}$$

And the minimum of (5.35) equals the smallest eigenvalue $\mu_{min.}$ of (5.36).
The smallest eigenvalue reads:

$$\mu_{min.} = \frac{1}{2}\{(\bar{x}_i^c d^{ij} \bar{x}_j^c + \bar{y}_i^c d^{ij} \bar{y}_j^c) + k^2(x_i^c d^{ij} x_j^c + y_i^c d^{ij} y_j^c) - \tag{5.37}$$

$$\sqrt{((\bar{x}_i^c d^{ij} \bar{x}_j^c + \bar{y}_i^c d^{ij} \bar{y}_j^c) - k^2(x_i^c d^{ij} x_j^c + y_i^c d^{ij} y_j^c))^2 + 4k^2[(\bar{x}_i^c d^{ij} x_j^c + \bar{y}_i^c d^{ij} y_j^c)^2 + (\bar{y}_i^c d^{ij} x_j^c - \bar{x}_i^c d^{ij} y_j^c)^2]}\ \}$$

From the first two equations of (5.36) we find that

$$\hat{\theta} = \tan^{-1} \frac{\bar{y}_i^c d^{ij} x_j^c - \bar{x}_i^c d^{ij} y_j^c}{\bar{x}_i^c d^{ij} x_j^c + \bar{y}_i^c d^{ij} y_j^c}, \tag{5.38}$$

and from the third equation we get

$$k\hat{\lambda} = \tan\hat{\phi} = \frac{\mu_{min.} - k^2(x_i^c d^{ij} x_j^c + y_i^c d^{ij} y_j^c)}{-k((\bar{x}_i^c d^{ij} x_j^c + \bar{y}_i^c d^{ij} y_j^c)\cos\hat{\theta} + (\bar{y}_i^c d^{ij} x_j^c - \bar{x}_i^c d^{ij} y_j^c)\sin\hat{\theta})} .$$

Together with (5.37) and (5.38) this finally gives

$$k\hat{\lambda} = \frac{k^2(x_i^c d^{ij} x_j^c + y_i^c d^{ij} y_j^c) - (\bar{x}_i^c d^{ij} \bar{x}_j^c + \bar{y}_i^c d^{ij} \bar{y}_j^c)}{2k\sqrt{(\bar{x}_i^c d^{ij} x_j^c + \bar{y}_i^c d^{ij} y_j^c)^2 + (\bar{y}_i^c d^{ij} x_j^c - \bar{x}_i^c d^{ij} y_j^c)^2}} + \sqrt{1 + \left[\frac{k^2(x_i^c d^{ij} x_j^c + y_i^c d^{ij} y_j^c) - (\bar{x}_i^c d^{ij} \bar{x}_j^c + \bar{y}_i^c d^{ij} \bar{y}_j^c)}{2k\sqrt{(\bar{x}_i^c d^{ij} x_j^c + \bar{y}_i^c d^{ij} y_j^c)^2 + (\bar{y}_i^c d^{ij} x_j^c - \bar{x}_i^c d^{ij} y_j^c)^2}}\right]^2} . \qquad (5.39)$$

The adjusted coordinates and translation parameters can be found by substituting (5.38) and (5.39) into (5.31).

5.5. The three dimensional Helmert transformation and its symmetrical generalization

Now that we have found the solution to the two dimensional Helmert transformation and its non-linear generalization, it is natural to try to generalize these results to three dimensions.
We will first consider the classical three dimensional Helmert transformation. The model for the three dimensional Helmert transformation reads:

$$\begin{pmatrix} x_i \\ y_i \\ z_i \end{pmatrix} = \lambda R_3(\gamma)R_2(\beta)R_1(\alpha) \begin{pmatrix} u_i \\ v_i \\ w_i \end{pmatrix} + \begin{pmatrix} t_x \\ t_y \\ t_z \end{pmatrix} + \begin{pmatrix} e_{x_i} \\ e_{y_i} \\ e_{z_i} \end{pmatrix}, \qquad (5.40)$$

where:
- $i = 1,\ldots,n$ = number of network points,
- x_i, y_i, z_i are the "observed" coordinates of the network points in the first coordinate system,
- u_i, v_i, w_i are the fixed given coordinates in the second coordinate system,
- $\lambda, (\alpha, \beta, \gamma)$ and (t_x, t_y, t_z) are respectively the scale, orientation and translation parameters,
- $e_{x_i}, e_{y_i}, e_{z_i}$ are the errors, and

$$- \; R_1(\alpha)=\begin{pmatrix}1 & 0 & 0\\ 0 & \cos\alpha & \sin\alpha\\ 0 & -\sin\alpha & \cos\alpha\end{pmatrix} \;, R_2(\beta)=\begin{pmatrix}\cos\beta & 0 & -\sin\beta\\ 0 & 1 & 0\\ \sin\beta & 0 & \cos\beta\end{pmatrix}, R_3(\gamma)=\begin{pmatrix}\cos\gamma & \sin\gamma & 0\\ -\sin\gamma & \cos\gamma & 0\\ 0 & 0 & 1\end{pmatrix}.$$

In contrast to the two dimensional case, the submanifold of the three dimensional Helmert transformation is curved. This complicates matters considerably. However, a number of simplifications can be obtained if we again apply the appropriate two-step procedure. In the first step we therefore assume the orientation parameters α, β and γ to be fixed, and solve for the scale $\lambda(\alpha,\beta,\gamma)$ and translation parameters $t_x(\alpha,\beta,\gamma), t_y(\alpha,\beta,\gamma), t_z(\alpha,\beta,\gamma)$. Since the first step consists of a linear adjustment problem, it is relatively easy to solve. The second adjustment step, where we have to solve for the orientation parameters, is however still non-linear. We will solve this second adjustment step by making use of the alternative formulation as discussed in example 1 of section 3.6 .

To apply the alternative formulation which makes use of the trace operator, we take the abbreviations

$$\underset{n\times 3}{X} = \begin{pmatrix} x_1 & y_1 & z_1 \\ \vdots & \vdots & \vdots \\ x_n & y_n & z_n \end{pmatrix}, \quad \underset{n\times 3}{U} = \begin{pmatrix} u_1 & v_1 & w_1 \\ \vdots & \vdots & \vdots \\ u_n & v_n & w_n \end{pmatrix}, \quad \underset{n\times 3}{E} = \begin{pmatrix} e_{x_1} & e_{y_1} & e_{z_1} \\ \vdots & \vdots & \vdots \\ e_{x_n} & e_{y_n} & e_{z_n} \end{pmatrix},$$

$$\underset{n\times 1}{H} = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}, \quad t = \begin{pmatrix} t_x \\ t_y \\ t_z \end{pmatrix}, \quad R = R_3(\gamma)R_2(\beta)R_1(\alpha) \;, \tag{5.41}$$

and write (5.40) as

$$\underset{n\times 3}{X} = \lambda \underset{n\times 3}{U} \underset{}{R^t} + \underset{n\times 3}{H} \underset{}{t^t} + \underset{n\times 3}{E} \;. \tag{5.42}$$

The first step of our adjustment problem reads then

$$\min_{\lambda,t} f(\lambda,t) = \min_{\lambda,t} \operatorname{trace}\left[(X-\lambda U R^t - H t^t)^t (X-\lambda U R^t - H t^t)\right]. \tag{5.43}$$

To find the critical point of the function $f(\lambda,t)$ the following results on matrix differentiation will be used:

If $\partial/\partial L = \partial/\partial L^{ij}$, then

$$\left.\begin{aligned} &\text{(a)} \quad \frac{\partial \operatorname{trace}(K L M)}{\partial L} = K^t M^t \\ &\text{(b)} \quad \frac{\partial \operatorname{trace}(L^t K L M)}{\partial L} = 2 K L M, \text{ when K and M symmetric} \\ &\text{(c)} \quad \frac{\partial \operatorname{trace}(L M L^t K)}{\partial L} = 2 K L M, \text{ when K and M symmetric} \end{aligned}\right\} \tag{5.44}$$

The proofs of these relations are straightforward, and we illustrate the method by proving (5.44.a):
Let

$$\text{trace } (K\ L\ M) = K_{ij} L^{jk} M_{ki},$$

where Einstein's summation convention is understood.
Then

$$\frac{\partial \text{ trace } (K\ L\ M)}{\partial\ L^{mn}} = K_{im} M_{ni} = K^{t}_{mi} M^{t}_{in}$$

or

$$\frac{\partial \text{ trace } (K\ L\ M)}{\partial\ L} = K^{t} M^{t}.$$

With the aid of (5.44) it follows from (5.43) that

$$\left.\begin{aligned} &(a) \quad \frac{\partial f}{\partial t} = -2\ (X - \lambda\ U\ R^{t})^{t} H + 2\ nt = 0 \\ &(b) \quad \frac{\partial f}{\partial \lambda} = 2\lambda\ \text{trace } (U^{t}U) - 2\ \text{trace } [(X - Ht^{t})^{t} (U\ R^{t})] = 0 \end{aligned}\right\} \qquad (5.45)$$

From (5.45.a) follows that

$$\boxed{t = \frac{1}{n}\ (X - \lambda\ U\ R^{t})^{t} H} \quad . \qquad (5.46)$$

Substitution of (5.46) into (5.45.b) gives

$$\lambda\ \text{trace } [U^{t}(I - \frac{1}{n} H\ H^{t})U] - \text{trace } [X^{t}(I - \frac{1}{n} H\ H^{t})U\ R^{t}] = 0. \qquad (5.47)$$

Note that $I - \frac{1}{n} H\ H^{t}$ is a projector, i.e. $(I - \frac{1}{n} H\ H^{t})(I - \frac{1}{n} H\ H^{t}) = (I - \frac{1}{n} H\ H^{t})$. With the abbreviations

$$\underset{n\times 3}{U^{c}} = \underset{n\times n}{(I - \frac{1}{n} H\ H^{t})}\ \underset{n\times 3}{U} \quad \text{and} \quad \underset{n\times 3}{X^{c}} = \underset{n\times n}{(I - \frac{1}{n} H\ H^{t})}\ \underset{n\times 3}{X} \qquad (5.48)$$

it therefore follows from (5.47) that

$$\boxed{\lambda = \frac{\text{trace } [(X^{c})^{t} (U^{c}) R^{t}]}{\text{trace } [(U^{c})^{t} (U^{c})]}} \quad . \qquad (5.49)$$

Formula (5.46) together with (5.49) constitute the solution of the first step. To formulate the second adjustment step, we substitute (5.46) and (5.49) into

$$\text{trace } [(X - \lambda\ U\ R^{t} - Ht^{t})^{t} (X - \lambda\ U\ R^{t} - Ht^{t})] \ .$$

This gives for the second adjustment step:

$$\min_{\alpha,\beta,\gamma} \left\{ \text{trace}\left[(X^c)^t(X^c)\right] - \frac{\text{trace}^2\left[(X^c)^t(U^c)R^t\right]}{\text{trace}\left[(U^c)^t(U^c)\right]} \right\} , \qquad \text{subject to } R = R_3(\gamma)R_2(\beta)R_1(\alpha) . \tag{5.50}$$

Since we know that the scale parameter λ must be positive and that $\text{trace}\left[(U^c)^t(U^c)\right]$ is positive, it follows from (5.49) that $\text{trace}\left[(X^c)^t(U^c)R^t\right]$ must be positive. We can therefore rephrase our second adjustment step as

$$\max_{\alpha,\beta,\gamma} \text{trace}\left[(X^c)^t(U^c)R^t\right] \quad \text{subject to } R = R_3(\gamma)R_2(\beta)R_1(\alpha). \tag{5.51}$$

To find the solution $\hat{R}$ to (5.51) we apply the singular value decomposition theorem (see e.g. Teunissen, 1984a) to the matrix $(X^c)^t(U^c)$. This theorem says that the matrix $(X^c)^t(U^c)$ may be factorized in the form

$$\underset{3\times3}{(X^c)^t(U^c)} = \underset{3\times3}{V_1}\ \underset{3\times3}{D}\ \underset{3\times3}{V_2^t} , \tag{5.52}$$

where V_1 and V_2 are orthogonal matrices of order 3x3 respectively, and D is a diagonal matrix of the form

$$\underset{3\times3}{D} = \begin{bmatrix} d_1 & 0 & 0 \\ 0 & d_2 & 0 \\ 0 & 0 & d_3 \end{bmatrix} ,$$

where d_i, i=1,2,3, are the singular values of $(X^c)^t(U^c)$, which may be ordered so that $d_1 \geq d_2 \geq d_3 \geq 0$.

From (5.52) follows that $(U^c)^t(X^c)(X^c)^t(U^c) = V_2 D^2 V_2^t$. Hence, the columns of V_2 give an orthonormal set of eigenvectors of the symmetric matrix $(U^c)^t(X^c)(X^c)^t(U^c)$ and the d_i^2 are the corresponding eigenvalues.
Substitution of (5.52) into (5.51) gives

$$\max_{\alpha,\beta,\gamma} \text{trace}\left[V_1 D V_2^t R^t\right] \quad \text{subject to } R = R_3(\gamma)R_2(\beta)R_1(\alpha) . \tag{5.53}$$

Since for arbitrary matrices A and B, $\text{trace}\left[A\ B\right] = \text{trace}\left[B\ A\right]$, we can rewrite (5.53) as

$$\max_{\alpha,\beta,\gamma} \text{trace}\left[V_2^t R^t V_1 D\right] \quad \text{subject to } R = R_3(\gamma)R_2(\beta)R_1(\alpha) . \tag{5.54}$$

If we denote the diagonal elements of the matrix $V_2^t R^t V_1$ by a_i, i=1,2,3, it follows that

$$\text{trace}\,(V_2^t R^t V_1 D) = a_1 d_1 + a_2 d_2 + a_3 d_3. \tag{5.55}$$

Let us now first assume that all three singular values are non-zero. Then, since the singular values d_i are positive and the matrices in the triple product $V_2^t R^t V_1$ are orthogonal, it follows that (5.55) is maximal if $a_i = 1$, $i = 1,2,3$. This implies then that (5.55) is maximal if and only if

$$V_2^t R^t V_1 = I.$$

Hence, our solution becomes

$$\hat{R} = V_1 V_2^t,$$

or with (5.52)

$$\boxed{\hat{R} = (X^c)^t (U^c) V_2 D^{-1} V_2^t} \quad . \tag{5.56}$$

Thus in case of non-zero singular values d_i, $i=1,2,3$, the orthogonal matrix $\hat{R}$ can be found from the eigenvectors and corresponding eigenvalues of the symmetric matrix $(U^c)^t (X^c)(X^c)^t (U^c)$. Since this matrix is of order 3x3 its characteristic equation is a cubic, say

$$\mu^3 + a\mu^2 + b\mu + c = 0.$$

Substitution of

$$\mu = x - \frac{1}{3} a \tag{5.57}$$

gives

$$x^3 + px + q = 0, \tag{5.58}$$

where $p = b - \frac{1}{3} a^2$ and $q = c + \frac{2}{27} a^3 - \frac{1}{3} ab.$

According to the Cardanian formula (see e.g. Griffiths, 1947) the three roots of (5.58) are:

$$\left.\begin{aligned} x_1 &= \left(-\tfrac{1}{2}q+\sqrt{(\tfrac{1}{2}q)^2+(\tfrac{1}{3}p)^3}\,\right)^{1/3} + \left(-\tfrac{1}{2}q-\sqrt{(\tfrac{1}{2}q)^2+(\tfrac{1}{3}p)^3}\,\right)^{1/3}, \\ x_2 &= \omega\left(-\tfrac{1}{2}q+\sqrt{(\tfrac{1}{2}q)^2+(\tfrac{1}{3}p)^3}\,\right)^{1/3} + \omega^2\left(-\tfrac{1}{2}q-\sqrt{(\tfrac{1}{2}q)^2+(\tfrac{1}{3}p)^3}\,\right)^{1/3}, \\ x_3 &= \omega^2\left(-\tfrac{1}{2}q+\sqrt{(\tfrac{1}{2}q)^2+(\tfrac{1}{3}p)^3}\,\right)^{1/3} + \omega\left(-\tfrac{1}{2}q-\sqrt{(\tfrac{1}{2}q)^2+(\tfrac{1}{3}p)^3}\,\right)^{1/3}, \end{aligned}\right\} \tag{5.59}$$

where $\omega = \cos\frac{2}{3}\pi + i \sin\frac{2}{3}\pi$ and $i^2 = -1$.

Thus with (5.59) and (5.57) one can compute the eigenvalues of the symmetric matrix $(U^c)^t (X^c)(X^c)^t (U^c)$. Once the eigenvalues are known it becomes straightforward to compute the corresponding eigenvectors.

Although the case of zero singular values will not occur very often in practice, let us now assume that one of the singular values, say d_j, equals zero. It follows then again that (5.55) is maximal if and only if $\hat{R} = V_1 V_2^t$. With (5.52) we can therefore write

$$\hat{R} = (X^c)^t (U^c) V_2 D^+ V_2^t + V_{1j} V_{2j}^t, \tag{5.60}$$

where D^+ is the pseudo-inverse of D, and V_{1j} and V_{2j} are the j-th column vectors of V_1 and V_2 respectively.

Finally we consider the case of multiple zero singular values. The case $d_1 = d_2 = d_3 = 0$ is trivial, since then the orthogonal matrix $\hat{R}$ is indeterminate and may take any arbitrary form. In case only two of the singular values, say d_2 and d_3, equal zero, we find that (5.55) is maximized if $\hat{R}$ takes the form

$$R = V_1 \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\phi & \pm\sin\phi \\ 0 & -\sin\phi & \pm\cos\phi \end{bmatrix} V_2^t, \qquad \text{where } \phi \text{ is arbitrary.}$$

Thus in case the two singular values d_j and d_k equal zero we find with (5.52) that the orthogonal matrix $\hat{R}$ takes the form

$$\hat{R} = (X^c)^t (U^c) V_2 D^+ V_2^t + \cos\phi (V_{1j} V_{2j}^t \pm V_{1k} V_{2k}^t) + \sin\phi (\pm V_{1j} V_{2k}^t - V_{1k} V_{2j}^t), \tag{5.61}$$

where ϕ is arbitrary.

In the geodetic literature a number of authors have studied the three dimensional Helmert transformation. The two most recent papers on the subject are (Sansò, 1973) and (Köchle, 1982). References to earlier papers can be found in (Schwidefsky and Ackermann, 1976).
Using the factorization of Cayley, (Köchle, 1982) arrives at an iterative solution for the orthogonal matrix $\hat{R}$. (Sansò, 1973) on the other hand, formulates the solution for $\hat{R}$ through the use of quaternion algebra in terms of an eigenvalue problem of a symmetric 4x4 matrix. His result is therefore to some extent comparable with our solution. Note, however, that our derivation is more general than Sansò's, since it does not require any restrictions on the number of columns in the matrices X and U in (5.42).

Now that we have found the solution of the three dimensional Helmert transformation (5.42), we will consider the three dimensional generalization of the Symmetric Helmert transformation (5.17). Using our alternative formulation the model can be written as

$$\left.\begin{aligned} X &= \lambda U R^t + H t^t + E \\ \bar{X} &= U + \bar{E} \end{aligned}\right\} \tag{5.62}$$

As in section 5.4 we assume that the covariance matrices Q and $\bar{Q}$ of the two coordinate sets $(\ldots x_i, y_i, z_i, \ldots)^t$ and $(\ldots \bar{x}_i, \bar{y}_i, \bar{z}_i \ldots)^t$ are of such a structure that $\bar{Q}^{-1}$ consists of 3x3 diagonal blocks of the type

$$\begin{bmatrix} d^{ij} & 0_{ij} & 0 \\ 0 & d^{ij} & 0_{ij} \\ 0 & 0 & d^{ij} \end{bmatrix},$$

and

$$k^2 Q = \bar{Q} \text{ for some } k \in \mathbb{R}^+.$$

Our adjustment problem becomes then

$$\underset{u_i, v_i, w_i, \lambda, t_x, t_y, t_z, \alpha, \beta, \gamma}{\text{minimize}} \quad f(u_i, v_i, w_i, \lambda, t_x, t_y, t_z, \alpha, \beta, \gamma), \tag{5.63}$$

with

$$f = k^2 \text{trace}\{(X - \lambda U R^t - H t^t)^t G (X - \lambda U R^t - H t^t)\} + \text{trace}\{(\bar{X} - U)^t G (\bar{X} - U)\}, \tag{5.63'}$$

and where the element of the nxn symmetric matrix G on place ij is given by d^{ij}.
To solve (5.63) we will proceed in three steps. First we will fix the scale λ and orientation parameters α, β, γ:

$$\begin{aligned} &\underset{u_i, v_i, w_i, t_x, t_y, t_z}{\text{minimize}} \quad g(u_i, v_i, w_i, t_x, t_y, t_z) = \\ &\underset{u_i, v_i, w_i, t_x, t_y, t_z}{\text{minimize}} \quad \{k^2 \text{trace}\{(X - \lambda U R^t - H t^t)^t G (X - \lambda U R^t - H t^t)\} + \text{trace}\{(\bar{X} - U)^t G (\bar{X} - U)\}\}. \end{aligned} \tag{5.64}$$

With the aid of the matrix differentiation rules of (5.44) we find that the critical point of g should satisfy:

$$\left.\begin{aligned} &\text{(a)} \quad \frac{\partial g}{\partial U} = 2(k^2\lambda^2 + 1) G U - 2\lambda k^2 G (X - H t^t) R - 2 G \bar{X} = 0 \\ &\text{(b)} \quad \frac{\partial g}{\partial t} = -2 k^2 (X - \lambda U R^t)^t G H + 2 k^2 t H^t G H = 0 \end{aligned}\right\} \tag{5.65}$$

From (5.65.b) we find that

$$t = (H^t G H)^{-1} (X - \lambda U R^t)^t G H. \tag{5.66}$$

Substitution of (5.66) into (5.65.a) gives

$$((k\lambda)^2 + 1) U - k^2 \lambda (I - H(H^t G H)^{-1} H^t G) X R - (k\lambda)^2 H (H^t G H)^{-1} H^t G U - \bar{X} = 0. \tag{5.67}$$

Premultiplication with $H(H^tGH)^{-1}H^tG$ shows that

$$H(H^t G H)^{-1} H^t G U = H(H^t G H)^{-1} H^t G \bar{X}.$$

Hence, we can write (5.67) also as

$$(k^2\lambda^2+1)U-k^2\lambda(I-H(H^tG\,H)^{-1}H^tG)X\,R-(I-H(H^tG\,H)^{-1}H^tG)\bar{X}-(k^2\lambda^2+1)H(H^tG\,H)^{-1}H^tG\,\bar{X}=0$$

With the abbreviations

$$X^c = (I-H(H^tG\,H)^{-1}H^tG)X \quad \text{and} \quad \bar{X}^c = (I-H(H^tG\,H)^{-1}H^tG)\bar{X}\ ,$$

we thus find that

$$\boxed{U = H(H^tG\,H)^{-1}H^tG\,\bar{X} + ((k\lambda)^2+1)^{-1}(\bar{X}^c+k^2\lambda\,X^cR)} \quad . \qquad (5.68)$$

When we substitute (5.68) into (5.66) we find the translation vector as

$$\boxed{t^t = (H^tG\,H)^{-1}H^tG(X-\lambda\,\bar{X}\,R^t)} \quad . \qquad (5.69)$$

(5.68) and (5.69) constitute the solution of our first adjustment step. Compare (5.68) and (5.69) with (5.31).

To commence with our second adjustment step we substitute (5.68) and (5.69) into (5.63') and find

$$f = k^2((k\lambda)^2+1)^{-1}\,\mathrm{trace}[(X^c-\lambda\,\bar{X}^cR^t)^tG(X^c-\lambda\,\bar{X}^cR^t)]. \qquad (5.70)$$

In a similar way as (5.56) was derived, we find that for fixed scale the conditional minimum of (5.70) is obtained by

$$\boxed{\hat{R} = (X^c)^tG(\bar{X}^c)V_2D^{-1}V_2^t} \quad , \qquad (5.71)$$

where the diagonal matrix D contains the singular values of $(X^c)^tG(\bar{X}^c)$ and the column vectors of the orthogonal matrix V_2 are provided by the eigenvectors of the 3x3 matrix $(\bar{X}^c)^tG(X^c)(X^c)^tG(\bar{X}^c)$.

To find the least-squares estimate of λ, we substitute (5.71) into (5.70) and use the reparametrization

$$k\lambda = \tan\phi, \quad 0 < \phi < \frac{1}{2}\pi. \qquad (5.72)$$

This gives

$$k^2\,\mathrm{trace}[(\cos\phi\,X^c-\sin\phi\,k^{-1}\bar{X}^c\,\hat{R}^t)^tG(\cos\phi\,X^c-\sin\phi\,k^{-1}\bar{X}^c\,\hat{R}^t)]. \qquad (5.73)$$

The minimization of (5.73) leads then to the following eigenvalue problem

$$\begin{bmatrix} k^2\,\mathrm{trace}\left((X^c)^t G(X^c)\right) & -k\,\mathrm{trace}\left((X^c)^t G(\bar{X}^c)\hat{R}^t\right) \\ -k\,\mathrm{trace}\left((X^c)^t G(\bar{X}^c)\hat{R}^t\right) & \mathrm{trace}\left((\bar{X}^c)^t G(\bar{X}^c)\right) \end{bmatrix} \begin{bmatrix} \cos\phi \\ \sin\phi \end{bmatrix} = \mu \begin{bmatrix} \cos\phi \\ \sin\phi \end{bmatrix}. \quad (5.74)$$

Since the minimum of (5.73) equals the smallest eigenvalue μ_{min} of (5.74) it follows that

$$k\hat{\lambda} = \tan\hat{\phi} = \frac{\mu_{min} - k^2\,\mathrm{trace}\left((X^c)^t G(X^c)\right)}{-k\,\mathrm{trace}\left((X^c)^t G(\bar{X}^c)\hat{R}^t\right)}. \quad (5.75)$$

From (5.74) follows that

$$\mu_{min.} = \frac{1}{2}\left\{k^2\,\mathrm{trace}\left((X^c)^t G(X^c)\right) + \mathrm{trace}\left((\bar{X}^c)^t G(\bar{X}^c)\right)\right\}$$

$$- \frac{1}{2}\sqrt{\left(k^2\,\mathrm{trace}\left((X^c)^t G(X^c)\right) - \mathrm{trace}\left((\bar{X}^c)^t G(\bar{X}^c)\right)\right)^2 + 4k^2\,\mathrm{trace}^2\left((X^c)^t G(\bar{X}^c)\hat{R}^t\right)}. \quad (5.76)$$

Substitution of (5.76) into (5.75) then finally gives

$$\hat{\lambda} = k^{-1}\left\{\frac{k^2\,\mathrm{trace}\left((X^c)^t G(X^c)\right) - \mathrm{trace}\left((\bar{X}^c)^t G(\bar{X}^c)\right)}{2k\,\mathrm{trace}\left((X^c)^t G(\bar{X}^c)\hat{R}^t\right)} + \sqrt{1 + \left(\frac{k^2\,\mathrm{trace}\left((X^c)^t G(X^c)\right) - \mathrm{trace}\left((\bar{X}^c)^t G(\bar{X}^c)\right)}{2k\,\mathrm{trace}\left((X^c)^t G(\bar{X}^c)\hat{R}^t\right)}\right)^2}\right\} \quad (5.77)$$

The least-squares estimates $\hat{t}$ and $\hat{U}$ are found from substituting (5.71) and (5.77) into (5.69) and (5.68) respectively.

5.6. The extrinsic curvatures estimated

In general, the problem of finding the curvature behaviour of submanifold $\bar{N}$ can only be solved through actual computation of the extrinsic curvatures k_N from the normal field **B** for a chosen tangent direction **X** and normal direction **N**. But, as will be clear from (4.30) the computation of the principal curvatures entails some extra expenses. It is therefore of some importance to have ways of

finding **realistic** estimates for the extrinsic curvatures of $\bar{N}$. As we see it, there are three possibilities:

(i) Try to compute the extrinsic curvatures analytically. Those cases where this is possible will, however, be rare.
Let us take as an example the Symmetric Helmert transformation (5.17). For convenience we reparametrize (5.17) as

$$\begin{aligned} \tilde{x}_i &= au_i + bv_i + t_y, \\ \tilde{y}_i &= -bu_i + av_i + t_y, \\ \tilde{\tilde{x}}_i &= u_i, \\ \tilde{\tilde{y}}_i &= v_i, \end{aligned} \tag{5.78}$$

where $a = \lambda\cos\theta$ and $b = \lambda\sin\theta$.
We assume that the observation equations $y^I(x^\alpha)$, $I = 1,\ldots,4n$, and the parameters x^α, $\alpha = 1,\ldots,2n+4$, of model (5.78) are ordered such that the design matrix $\partial_\alpha y^I$ reads in partitioned form as

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix}, \tag{5.79}$$

where:

matrix A is 2n x 2n block-diagonal with equal 2 x 2 blocks $\begin{bmatrix} a & b \\ -b & a \end{bmatrix}$,

$$B_{2n\times 4} = \begin{bmatrix} u_1 & v_1 & 1 & 0 \\ v_1 & -u_1 & 0 & 1 \\ \vdots & \vdots & \vdots & \vdots \\ u_n & v_n & 1 & 0 \\ v_n & -u_n & 0 & 1 \end{bmatrix}, \quad C = I_{2n} \text{ and } D_{2n\times 4} = O_{2n\times 4}.$$

We also assume that the observation space has the standard metric, i.e. $g_{IJ} = \delta_{IJ}$. It follows then that the induced metric $g_{\alpha\beta}$ reads in partitioned form as

$$\begin{bmatrix} A^tA + I_{2n} & A^tB \\ B^tA & B^tB \end{bmatrix}, \tag{5.80}$$

with:

$$A^tA = \lambda^2 I_{2n},$$

$$B^tA_{4x2n} = \begin{pmatrix} au_1-bv_1 & bu_1+av_1 & \cdots\cdots & au_n-bv_n & bu_n+av_n \\ av_1+bu_1 & bv_1-au_1 & \cdots\cdots & av_n+bu_n & bv_n-au_n \\ a & b & \cdots\cdots & a & b \\ -b & a & \cdots\cdots & -b & a \end{pmatrix}, \text{ and}$$

$$B^tB = \begin{pmatrix} \sum_{i=1}^{n}(u_i^2+v_i^2) & 0 & nu_c & nv_c \\ 0 & \sum_{i=1}^{n}(u_i^2+v_i^2) & nv_c & -nu_c \\ nu_c & nv_c & n & 0 \\ nv_c & -nu_c & 0 & n \end{pmatrix}.$$

Furthermore it follows from (5.78) that the non-zero second derivatives of $y^I(x^\alpha)$ are given by:

$$\frac{\partial^2 y^{I=2i-1}}{\partial x^{\alpha=2i-1}\partial x^{\alpha=n+1}} = 1, \qquad \frac{\partial^2 y^{I=2i-1}}{\partial x^{\alpha=2i}\partial x^{\alpha=n+2}} = 1,$$

$$\frac{\partial^2 y^{I=2i}}{\partial x^{\alpha=2i-1}\partial x^{\alpha=n+2}} = -1 \qquad \text{and} \qquad \frac{\partial^2 y^{I=2i}}{\partial x^{\alpha=2i}\partial x^{\alpha=n+1}} = 1,$$

for i = 1,...,n.

Hence for an arbitrary unit normal vector **N**, i.e.

$$\langle y_*(\partial_\alpha), \mathbf{N}\rangle_M = 0, \tag{5.81}$$

the matrix $\langle \mathbf{B}(\partial_\alpha, \partial_\beta), \mathbf{N}\rangle_M$ reads in partitioned form as

$$\begin{pmatrix} E & F \\ F^t & G \end{pmatrix}, \tag{5.82}$$

where:

$$E_{2nx2n} = O_{2nx2n}, \quad G_{4x4} = O_{4x4} \text{ and}$$

$$F^t_{4x2n} = \begin{pmatrix} N^{I=1} & N^{I=2} & \cdots\cdots & N^{I=2n-1} & N^{I=2n} \\ -N^{I=2} & N^{I=1} & \cdots\cdots & -N^{I=2n} & N^{I=2n-1} \\ 0 & 0 & \cdots\cdots & 0 & 0 \\ 0 & 0 & \cdots\cdots & 0 & 0 \end{pmatrix}.$$

In order to determine the with the normal direction $\mathbf{N}$ corresponding principal curvatures, we need to solve the general eigenvalue problem

$$| \langle B(\partial_\alpha, \partial_\beta), \mathbf{N} \rangle_M - k_{\mathbf{N}}\, g_{\alpha\beta} | = 0.$$

With (5.80) and (5.82) this gives

$$\begin{vmatrix} -k_{\mathbf{N}}(1+\lambda^2) I_{2n} & F - k_{\mathbf{N}} A^t B \\ (F - k_{\mathbf{N}} A^t B)^t & -k_{\mathbf{N}} B^t B \end{vmatrix} = 0. \tag{5.83}$$

Now assume that $k_{\mathbf{N}} \neq 0$. Then we can apply the following well-known result:
If U is a regular square matrix, then

$$\begin{vmatrix} U & Y \\ X & V \end{vmatrix} = | U | \, | V - X U^{-1} Y | \, . \tag{5.84}$$

This applied to (5.83) gives

$$| F^t F - k_{\mathbf{N}} F^t A^t B - k_{\mathbf{N}} B^t A F + k_{\mathbf{N}}^2 B^t A^t A B - k_{\mathbf{N}}^2 (1+\lambda^2) B^t B | = 0. \tag{5.85}$$

Since $\mathbf{N}$ is a normal vector it follows from (5.79) and (5.81) that

$$\sum_{i=1}^{n} (N^{2i-1} u_i + N^{2i} v_i) = 0, \quad \sum_{i=1}^{n} (N^{2i-1} v_i - N^{2i} u_i) = 0,$$

$$\sum_{i=1}^{n} N^{2i-1} = 0 \text{ and } \sum_{i=1}^{n} N^{2i} = 0.$$

Hence $F^t A^t B = 0$. With $A^t A = \lambda^2 I_{2n}$, this gives for (5.85)

$$| F^t F - k_{\mathbf{N}} B^t B | = 0,$$

or

$$\begin{vmatrix} \sum_{I=1}^{2n} (N^I)^2 - k_{\mathbf{N}}^2 \sum_{i=1}^{n} (u_i^2 + v_i^2) & 0 & -k_{\mathbf{N}}^2 n u_c & -k_{\mathbf{N}}^2 n v_c \\ 0 & \sum_{I=1}^{2n} (N^I)^2 - k_{\mathbf{N}}^2 \sum_{i=1}^{n} (u_i^2 + v_i^2) & -k_{\mathbf{N}}^2 n v_c & k_{\mathbf{N}}^2 n u_c \\ -k_{\mathbf{N}}^2 n u_c & -k_{\mathbf{N}}^2 n v_c & -k_{\mathbf{N}}^2 n & 0 \\ -k_{\mathbf{N}}^2 n v_c & k_{\mathbf{N}}^2 n u_c & 0 & -k_{\mathbf{N}}^2 n \end{vmatrix} = 0. \tag{5.86}$$

We can now apply the following variant of (5.84):

$$\begin{vmatrix} U & Y \\ X & V \end{vmatrix} = |V| \, |U - Y V^{-1} X| .$$

This gives for (5.86)

$$| \{ \sum_{I=1}^{2n} (N^I)^2 - k_N^2 \sum_{i=1}^{n} (u_i^2 + v_i^2) \} I_2 + n k_N^2 (u_c^2 + v_c^2) I_2 | = 0 ,$$

or

$$\sum_{I=1}^{2n} (N^I)^2 = k_N^2 \{ \sum_{i=1}^{n} (u_i^2 + v_i^2) - n u_c^2 - n v_c^2 \} = k_N^2 \sum_{i=1}^{n} \{ (u_i^c)^2 + (v_i^c)^2 \} .$$

Thus the two non-zero principal curvatures of model (5.78) read

$$k_N = \pm \sqrt{\frac{\sum_{I=1}^{2n} (N^I)^2}{\sum_{i=1}^{n} \{ (u_i^c)^2 + (v_i^c)^2 \}}} . \tag{5.87}$$

(ii) Try to estimate the extrinsic curvatures with the help of the information which comes available during the iteration. Recall from section 3.6 that the numerical examples clearly betray the curvature involved.
In order to estimate the curvature during the iteration we need a manageable formula. Formula (4.37a) does therefore not apply, since it needs $\hat{x}$ in advance. The following formula, however, can be used:

$$\lim_{q \to \infty} \frac{|| y_*(\text{grad } E) ||_{y(x_{q+1})}}{|| y_*(\text{grad } E) ||_{y(x_q)}} =$$

$$= \left(\frac{(x^\alpha_{q+2} - x^\alpha_{q+1}) \, g_{\alpha\beta}(x_{q+1}) (x^\beta_{q+2} - x^\beta_{q+1})}{(x^\alpha_{q+1} - x^\alpha_{q}) \, g_{\alpha\beta}(x_q) (x^\beta_{q+1} - x^\beta_{q})} \right)^{\frac{1}{2}} \leq \max . \{ |k_N^1| \, || y_s - \hat{y} ||_M , |k_N^n| \, || y_s - \hat{y} ||_M \} \tag{5.88}$$

The proof of (5.88) goes similar to that of (4.37.a).

(iii) Try to obtain rigorous bounds on the extrinsic curvatures. From Gauss' decomposition formula (4.18) follows that

$$\langle D_{y_*(X)} y_*(Y), N \rangle_M = \langle B(X,Y), N \rangle_M , \tag{5.89}$$

for $X, Y \in T_x N$, $N \in T^{\perp}_{y} \tilde{N}$.

Let k denote the in absolute value largest principal curvature for the normal direction $\mathbf{N} = \hat{\mathbf{e}}/||\hat{\mathbf{e}}||_M$, with $\hat{\mathbf{e}} = \mathbf{y}_s - \hat{\mathbf{y}}$. According to the eigenvalue problem (4.30) we have then

$$|\langle k\, \mathbf{N}, \hat{\mathbf{e}} \rangle_M| = ||\langle g^{\beta\alpha} \mathbf{B}(\partial_\alpha, \partial_\gamma), \hat{\mathbf{e}} \rangle_M||_2 , \tag{5.90}$$

where the matrix norm $||.||_2$ is the spectral norm.
With (5.89) and the Cauchy-Schwarz inequality we obtain the following upperbound:

$$\begin{aligned} |\langle k\, \mathbf{N}, \hat{\mathbf{e}} \rangle_M| &\leq ||\, g^{\beta\alpha}(\xi)\, ||_2\, ||\, \langle \mathbf{B}(\partial_\alpha, \partial_\gamma), \hat{\mathbf{e}} \rangle_M\, ||_2 \\ &\leq g^{\gamma\gamma}(\xi) ||\partial^2_{\alpha\beta} y^i(\xi)||_2 ||g_{ij}\hat{e}^j||_2, \\ \text{with}\quad g^{\gamma\gamma}(\xi) &= \text{trace}\,(g^{\alpha\beta}(\xi)). \end{aligned} \tag{5.91}$$

To estimate the spectral radius of $\partial^2_{\alpha\beta} y^i$ one can make use of the various exclusion theorems known from the literature. For instance, one of the simplest exclusion theorems is: For all eigenvalues μ of a matrix $A_{\alpha\beta}$ one has

$$|\mu| \leq \max_{x^\beta \neq o} \frac{||A_{\alpha\beta} x^\beta||}{||x^\beta||} , \tag{5.92}$$

where $||.||$ is a chosen vector norm.
For the max-norm $||x^\beta||_\infty = \max_\beta |x^\beta|$ this becomes then

$$|\mu| \leq \max_\alpha \sum_\beta |A_{\alpha\beta}|, \qquad \text{i.e. the largest row of } A_{\alpha\beta}. \tag{5.93}$$

For a diagonal dominant matrix one could take Gershgorin's theorem, which says that the union of all discs

$$D_\alpha = \{\mu \in \mathbb{C} \,|\, |\mu - A_{\alpha\alpha}| \leq \sum_{\substack{\beta=1 \\ \beta\neq\alpha}} |A_{\alpha\beta}|\}, \text{ (no summation over } \alpha), \tag{5.94}$$

contains all eigenvalues of the n x n matrix $A_{\alpha\beta}$.
Instead of using exclusion theorems one could also try to compute the spectral radius of $\partial^2_{\alpha\beta} y^i$ directly. This can turn out to be feasible especially when per observation equation only a few parameters are involved.

As an alternative to (5.91) we could make use of condition equations if they are available. Let $\mathbf{Y} \in T_x N$ and $\bar{\mathbf{N}} \in T^\perp_x N$. Then $\langle \mathbf{y}_*(\mathbf{Y}), \bar{\mathbf{N}} \rangle_M = 0$. Hence,

$$0 = D_{y_*(X)}\langle y_*(Y), \bar{N}\rangle_M = \langle D_{y_*(X)} y_*(Y), \bar{N}\rangle_M + \langle y_*(Y), D_{y_*(X)}\bar{N}\rangle_M ,$$

and with Gauss' decomposition formula (4.18) this gives

$$\langle B(X,Y), \bar{N}\rangle_M = - \langle D_{y_*(X)}\bar{N}, y_*(Y)\rangle_M . \tag{5.95}$$

Now assume that

$$\bar{N} = P_{T\bar{N}^{\perp}, T\bar{N}}(y_s - y) = P_{T\bar{N}^{\perp}, T\bar{N}} e .$$

With the condition equations $u^{\rho}(y) = 0$, $\rho = 1,\ldots,(m-n)$, we can write then

$$\bar{N} = (g^{jk}\partial_k u^{\rho} g_{\rho\tau}\partial_l u^{\tau} e^l)\,\partial_j , \quad j,k,l=1,\ldots m; \quad \rho,\tau=1,\ldots,(m-n),$$

where

$$g^{\rho\tau} = \partial_i u^{\rho} g^{ij} \partial_j u^{\tau} .$$

And with $X = \partial_\alpha$, $Y = \partial_\beta$ this gives for (5.95),

$$\langle B(\partial_\alpha, \partial_\beta), \bar{N}\rangle_M = - \partial_\alpha y^i \partial^2_{ij} u^{\rho} \partial_\beta y^j g_{\rho\tau} \partial_k u^{\tau} e^k .$$

Hence,

$$\begin{aligned} &|\langle k\, N, \hat{e}\rangle_M| = \|\langle g^{\beta\alpha} B(\partial_\alpha, \partial_\gamma), \hat{e}\rangle_M\|_2 = \\ &\|g^{\beta\alpha}(\xi)\partial_\alpha y^i(\xi)\partial^2_{ij} u^{\rho}(\eta)\partial_\gamma y^j(\xi) g_{\rho\tau}(\eta)\partial_k u^{\tau}(\eta)\hat{e}^k\|_2 , \\ &\alpha,\beta,\gamma = 1,\ldots,n \ , \ i,j,k = 1,\ldots,m \ , \ \rho,\tau = 1,\ldots,(m-n) . \end{aligned} \tag{5.96}$$

Expression (5.96) still looks horribly complicated. But we can simplify it somewhat by recalling the well known result that for two arbitrary matrices A^{α}_i, B^i_{α}, $i=1,\ldots,m$, $\alpha=1,\ldots,n$ their products $A^{\alpha}_i B^j_{\alpha}$ and $B^i_{\alpha} A^{\beta}_i$ have the same non-zero eigenvalues with the same multiplicities. Application of this result to (5.96) gives

$$\begin{aligned} |\langle k\, N, \hat{e}\rangle_M| &= \|g^{\beta\alpha}(\xi)\partial_\alpha y^i(\xi) g_{il} \cdot g^{lm}\partial^2_{mn} u^{\rho}(\eta) g_{\rho\tau}(\eta)\partial_k u^{\tau}(\eta)\hat{e}^k \partial_\gamma y^n(\xi)\|_2 \\ &= \|g^{lm}\partial^2_{mn} u^{\rho}(\eta) g_{\rho\tau}(\eta)\partial_k u^{\tau}(\eta)\hat{e}^k \partial_\beta y^n(\xi) g^{\beta\alpha}(\xi)\partial_\alpha y^i(\xi) g_{il}\|_2 \\ &\le \|g^{lm}\partial^2_{mn} u^{\rho}(\eta) g_{\rho\tau}(\eta)\partial_k u^{\tau}(\eta)\hat{e}^k\|_2 \|\partial_\beta y^n(\xi) g^{\beta\alpha}(\xi)\partial_\alpha y^i(\xi) g_{il}\|_2 , \end{aligned}$$

or

$$\boxed{|\langle k\ \mathbf{N}, \hat{\mathbf{e}}\rangle_{M}| \leq ||g^{lm}\partial^{2}_{mn}u^{\rho}(\xi)g_{\rho\tau}(\xi)\partial_{k}u^{\tau}(\xi)\hat{e}^{k}||_{2}}\ , \tag{5.97}$$

since the eigenvalues of a projector equal one or zero.

5.7. Some two dimensional networks

(i) Recall that Gauss' method (4.5) has a local quadratic convergence behaviour in case submanifold $\tilde{N}$ is totally geodesic. A typical example of a type of geodetic network for which this holds is a planar geodetic triangulation chain if the parameters are cartesian coordinates (see figure 33).

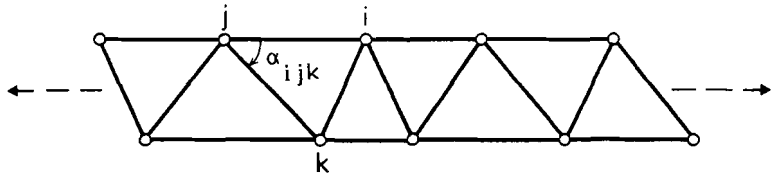


figure 33

(ii) In the previous subsection we observed that it may be worthwhile to try to compute the spectral radii of the matrices $\partial^{2}_{\alpha\beta}y^{i}$, $i=1,\ldots,m$, directly if every observation equation only contains a few parameters. Fortunately, this is precisely the case in geodetic network theory. In case of a two dimensional trilateration network for instance only four parameters are involved in each observation.

By expanding the distance function l_{pq} connecting the network points P_p and P_q we get

$$l_{pq} = l^{o}_{pq} + \frac{1}{l^{o}_{pq}}\left(-x^{o}_{pq}\ x^{o}_{pq} - y^{o}_{pq}\ y^{o}_{pq}\right)\begin{pmatrix}\Delta x_p\\ \Delta x_q\\ \Delta y_p\\ \Delta y_q\end{pmatrix} +$$

$$+\frac{1}{2(l^{o}_{pq})^{3}}\left(\Delta x_p\ \Delta x_q\ \Delta x_p\ \Delta y_q\right)\begin{pmatrix} (y^{o}_{pq})^{2} & -(y^{o}_{pq})^{2} & -x^{o}_{pq}y^{o}_{pq} & x^{o}_{pq}y^{o}_{pq}\\ -(y^{o}_{pq})^{2} & (y^{o}_{pq})^{2} & x^{o}_{pq}y^{o}_{pq} & -x^{o}_{pq}y^{o}_{pq}\\ -x^{o}_{pq}y^{o}_{pq} & x^{o}_{pq}y^{o}_{pq} & (x^{o}_{pq})^{2} & -(x^{o}_{pq})^{2}\\ x^{o}_{pq}y^{o}_{pq} & -x^{o}_{pq}y^{o}_{pq} & -(x^{o}_{pq})^{2} & (x^{o}_{pq})^{2}\end{pmatrix}\begin{pmatrix}\Delta x_p\\ \Delta x_q\\ \Delta y_p\\ \Delta y_q\end{pmatrix} + \ldots$$

And it is not too difficult to verify that the maximum eigenvalue of its 4x4 Hessian is given by

$$\lambda_{max} = \frac{2}{l^{o}_{pq}} \quad . \tag{5.98}$$

Since in practice the observations are usually assumed to be uncorrelated with equal variance, we can write

$$g_{ij} = \text{diag}\,(\ldots\ \sigma_y^{-2}\ \ldots) \tag{5.99}$$

Now if we also assume that all distances in the network are about the same, i.e. $l^{o}_{pq} \simeq l$, and that the variances of the estimated parameters do not differ too much, we get with (5.98) and (5.99) for (5.91):

$$|\langle k\ \mathbf{N}, \hat{\mathbf{e}}\rangle_{M}| \leq 2n\ \frac{\sigma_x^2}{\sigma_y^2}\ \sum_{i=1}^{m} \frac{|\hat{e}^i|}{l} \quad . \tag{5.100}$$

(iii) As an example of how to apply (5.97) we take a two dimensional closed polygon in which every two neighbouring points are connected by one measured distance l and one measured azimuth A. The two condition equations read then:

$$\sum_{i=1}^{p} l_i \cos A_i = 0, \qquad \text{and} \qquad \sum_{i=1}^{p} l_i \sin A_i = 0 \quad . \tag{5.101}$$

If we assume that the observations are uncorrelated and the variances satisfy

$$\sigma_{l_i}^2 = l_i^2\ \sigma_{A_i}^2 , \tag{5.102}$$

then

$$g^{\rho\tau} = \left(\sum_{i=1}^{p} \sigma_{l_i}^2\right) \delta^{\rho\tau}, \qquad \rho,\tau = 1,2 \quad , \tag{5.103}$$

and

$$g_{\rho\tau}\,\partial_k u^{\tau} e^{k} = \begin{cases} \left(\sum_{i=1}^{p} \sigma_{l_i}^2\right)^{-1} \sum_{j=1}^{n/2} \left(e^{2j-1} \cos A_j - e^{2j} l_j \sin A_j\right) \overset{\text{call}}{=} a\ , & \text{for } \rho=1 \\ \left(\sum_{i=1}^{p} \sigma_{l_i}^2\right)^{-1} \sum_{j=1}^{n/2} \left(e^{2j-1} \sin A_j + e^{2j} l_j \cos A_j\right) \overset{\text{call}}{=} b\ , & \text{for } \rho=2 \end{cases} \tag{5.104}$$

where the odd numbered residuals refer to the distance residuals and the even numbered residuals refer to the azimuth residuals.

Furthermore it follows that the following two 2n x 2n matrices

$g^{lm}\partial^2_{mn}u^{\rho=1}$ and $g^{lm}\partial^2_{mn}u^{\rho=2}$ are block-diagonal with blocks of respectively

$$\begin{pmatrix} 0 & -\sigma^2_{l_i}\sin A_i \\ -\sigma^2_{A_i}\sin A_i & -\sigma^2_{A_i} l_i \cos A_i \end{pmatrix} \text{ and } \begin{pmatrix} 0 & \sigma^2_{l_i}\cos A_i \\ \sigma^2_{A_i}\cos A_i & -\sigma^2_{A_i} l_i \sin A_i \end{pmatrix}. \tag{5.105}$$

From (5.102), (5.103), (5.104) and (5.105) follows then that the 2n x 2n matrix

$$g^{lm}\partial^2_{mn}u^{\rho} g_{\rho\tau}\partial_k u^{\tau} e^k, \tag{5.106}$$

is block-diagonal with blocks

$$\begin{pmatrix} 0 & \sigma^2_{l_i}(-a\sin A_i + b\cos A_i) \\ \sigma^2_{l_i} l_i^{-2}(-a\sin A_i + b\cos A_i) & -\sigma^2_{l_i} l_i^{-1}(a\cos A_i + b\sin A_i) \end{pmatrix}.$$

The eigenvalues λ_i of matrix (5.106) read

$$\lambda_i = \frac{-(a\cos A_i + b\sin A_i) \pm \sqrt{4(a^2+b^2) - 3(a\cos A_i + b\sin A_i)^2}}{2 l_i \sigma^{-2}_{l_i}}.$$

And from this follows that

$$|\lambda_i| \leq \frac{3}{2}\frac{\sigma^2_{l_i}}{l_i}(|a| + |b|). \tag{5.107}$$

Hence, with (5.104) we find for (5.97):

$$|\langle k\, \mathbf{N}, \hat{\mathbf{e}}\rangle_M| \leq \tag{5.108}$$

$$\frac{3}{2}\cdot\frac{\sigma^2_{l_k}}{\sum_{i=1}^{n}\sigma^2_{l_i}}\cdot\frac{|\sum_{j=1}^{n/2}\hat{e}^{2j-1}\cos\hat{A}_j - \hat{e}^{2j}\hat{l}_j\sin\hat{A}_j| + |\sum_{j=1}^{n}\hat{e}^{2j-1}\sin\hat{A}_j + \hat{e}^{2j}\hat{l}_j\cos\hat{A}_j|}{\hat{l}_k}$$

where $\hat{l}_k$ is the distance for which $\sigma^2_{l_i}/\hat{l}_i$, $i=1,\ldots,n$, is the greatest.

6. Some statistical considerations

In the previous sections we dealt with the problem of finding the least-squares solution $\hat{\mathbf{y}}$ to

$$\min_{\mathbf{y}\in\bar{N}} . || \mathbf{y}_s - \mathbf{y} ||^2_M . \tag{6.1}$$

But it is of course not enough to compute a vector $\hat{\mathbf{y}} \in \bar{N}$ and state that this is the estimated value of the unknown $\tilde{\mathbf{y}} \in \bar{N}$. The step following the actual adjustment process is equally important. That is, one also needs to find the statistical properties of the estimators involved and formulate ways of testing statistical hypotheses. Unfortunately we are not able yet to present a complete treatment of the statistical theory dealing with non-linear geodesic estimation, although it will be clear that in considering non-linear models one cannot expect a well working theory as we know it for linear models. In the following we will restrict ourselves therefore to a few general remarks.

As we have seen, Gauss' method enabled us, given the observation point $\mathbf{y}_s \in M$, to compute the least-squares estimate $\hat{\mathbf{x}}$ of $\mathbf{x}$. And with the map $\mathbf{y}: N \to M$ this gives the least-squares estimate $\hat{\mathbf{y}} = \mathbf{y}(\hat{\mathbf{x}})$ of $\tilde{\mathbf{y}} \in \bar{N} \subset M$. In this way the least-squares estimation method defines, at least implicitly, two non-linear maps $\mathbf{P}: M \to \bar{N}$ and $\mathbf{y}^{-1} \circ \mathbf{P}: M \to N$ such that

$$\hat{\mathbf{y}} = \mathbf{P}(\mathbf{y}_s) \quad \text{and} \quad \hat{\mathbf{x}} = \mathbf{y}^{-1} \circ \mathbf{P}(\mathbf{y}_s), \tag{6.2}$$

where $\mathbf{y}^{-1}$ is a leftinverse of $\mathbf{y}: N \to M$.
If the observation process which yielded our data were to be repeated, we would obtain different values for $\mathbf{y}_s$. And application of the maps $\mathbf{P}$ and $\mathbf{y}^{-1} \circ \mathbf{P}$ to the new data would yield different values of $\hat{\mathbf{y}}$ and $\hat{\mathbf{x}}$ respectively. Thus the $\hat{\mathbf{y}}$ and $\hat{\mathbf{x}}$ are themselves samples drawn from certain probability distributions which depend both on the nature of the maps involved and on the assumed normal distribution of the observations. For making statistical inferences it is therefore important to know the statistical properties of the estimators involved.
In case the coordinate functions $y^i(x^\alpha)$ of the map $\mathbf{y}$ are linear, it is not difficult to derive the precise distribution of the least-squares estimators. The following distributional properties are well known:

$$\begin{aligned}
&\text{(a)} \quad \hat{x}^\alpha \curvearrowright N(x^\alpha, \sigma^2 g^{\alpha\beta}) && \text{(c)} \quad \hat{e}^i \curvearrowright N(0, \sigma^2 (g^{ij} - \partial_\alpha y^i g^{\alpha\beta} \partial_\beta y^j)) \\
&\text{(b)} \quad \hat{y}^i \curvearrowright N(\tilde{y}^i, \sigma^2 \partial_\alpha y^i g^{\alpha\beta} \partial_\beta y^j) && \text{(d)} \quad || \mathbf{y}_s - \hat{\mathbf{y}} ||^2_M \curvearrowright \sigma^2 \chi^2_{m-n}
\end{aligned} \tag{6.3}$$

However, these results do **not** carry over to the non-linear case. Only in the exceptional case that one is dealing with a totally geodesic submanifold $\bar{N}$ will the last three distributional properties of (6.3) still hold. Of course, a similar complete theory as we know it for linear models can hardly be expected. Essential properties which are used repeatedly in the development of the linear theory break down completely in the non-linear case. Take for instance the mathematical expectation

operator $E\{\ .\}$. If z is a random variable and g is a non-linear map, then

$$E\{\ g(z)\} \neq g(E\{\ z\}\ ), \tag{6.4}$$

i.e., the mean of the image differs generally from the image of the mean. Hence, we can hardly expect our least-squares estimators to be unbiased in the non-linear case. Consequently, one cannot justify least-squares estimation anymore by referring to the Gauss-Markov theorem. Of course this by no means implies that one should do away with the least-squares estimators. Under the usual assumption of normality the least-squares estimators are namely still maximum likelihood estimators. Besides, when one overemphasizes the importance of exactly unbiased estimators, one can find oneselves in an impossible situation. Very often namely we have a natural estimator which is, however, slightly biased. For example, if z is a good unbiased estimator of $\tilde{z}$, and if it is required to estimate $g(\tilde{z})$, then it seems natural to estimate $g(\tilde{z})$ by $g(z)$, although this estimator will nearly always be biased.

Another property that fails to carry over to the non-linear case, is the property of estimability. Recall that with respect to a linear model $\tilde{y} \in \tilde{N} = AN \subset M$, a linear function (x^*, x), $x^* \in N^*$, $x \in N$, is usually defined to be an estimable function if it admits an unbiased linear estimator. However, this definition cannot be used for a non-linear model. First of all since a restriction to linear estimators is not reasonable anymore, and secondly since non-linear estimators are almost always biased. Thus what we need is a more general definition of estimability, one which for linear models reduces to the above given one. The answer is given by the dual relation

$$A^{-1}(0) = (A^* M^*)^{\circ}. \tag{6.5}$$

This dual relation implies namely that either

$$x^* = A^* y^* \quad \text{for some } y^* \in M^* \quad \text{or} \quad Ax = 0 \quad \text{and} \quad (x^*, x) \neq 0,$$

but not both hold. Hence, asking for an unbiased linear function (x^*, x) is equivalent to asking for a linear function (x^*, x) which is invariant to solutions of $Ax = 0$ (see e.g. Grafarend and Schaffrin, 1974). Therefore in general it would seem more appropriate to couple the definition of estimability to the property of invariance.

Since it is impossible in general to derive precise formulae for the distributional properties of the non-linear estimators, the best we can do seems to be to find approximations. Three approaches suggest themselves:

When one has a non-linear model it is natural to hope that it is only moderately non-linear so that application of the linear theory is justified. In practical applications the first step taken should therefore be to prove whether a linear(ized) model is sufficient as approximation, because then the statistical treatment is much more simple. And since the origin of all complications in non-linear adjustment lies in the presence of curvatures, it seems reasonable to take the mean curvature as a measure of non-linearity. Let us therefore Taylorize the expressions in (6.2) about the true values $\tilde{y} = y(x)$. With $e = y_s - \tilde{y}$ this gives:

$$\hat{y}^k = (P(y_s))^k = \tilde{y}^k + \partial_i (P(\tilde{y}))^k e^i + \frac{1}{2} \partial^2_{ij} (P(\tilde{y}))^k e^i e^j + \ldots ,$$

and

$$\hat{x}^\alpha = (y^{-1} \circ P(y_s))^\alpha = x^\alpha + \partial_i (y^{-1} \circ P(\tilde{y}))^\alpha e^i + \frac{1}{2} \partial^2_{ij} (y^{-1} \circ P(\tilde{y}))^\alpha e^i e^j + \ldots .$$

By taking the expectation we find to an approximation of the order σ^4:

$$E\{\hat{y}^k - \tilde{y}^k\} = \frac{1}{2} \sigma^2 \partial^2_{ij} (P(\tilde{y}))^k g^{ij}$$

and (6.6)

$$E\{\hat{x}^\alpha - x^\alpha\} = \frac{1}{2} \sigma^2 \partial^2_{ij} (y^{-1} \circ P(\tilde{y}))^\alpha g^{ij} .$$

And with the definitions of the unique mean curvature normal $\bar{\mathbf{N}}$ (see (4.33)) and the Christoffel symbols of the second kind $\Gamma^\alpha_{\beta\gamma}$ (see (4.17)), and by using the fact that $\mathbf{y}_s - \mathbf{P}(\mathbf{y}_s) \in T^\perp_{\hat{\mathbf{y}}} \bar{N}$, one will find that one can rewrite (6.6) as

(a) $E\{\hat{\mathbf{y}} - \tilde{\mathbf{y}}\} = \frac{1}{2} \sigma^2 n \, \bar{k}_{\mathbf{N}_p} \mathbf{N}_p = \frac{1}{2} \sigma^2 n \, \bar{\mathbf{N}}$,

and

(b) $E\{\hat{x}^\alpha - x^\alpha\} = -\frac{1}{2} \sigma^2 g^{\beta\gamma} \Gamma^\alpha_{\beta\gamma}$, (6.7)

where $\mathbf{N}_p$, $p = 1, \ldots, (m-n)$, is an orthonormal basis of $T^\perp_{\tilde{\mathbf{y}}} \bar{N}$ and $\alpha, \beta, \gamma = 1, \ldots, n$.

(see also Teunissen, 1984c).

Thus the first moments of the parameters depend on the connection coefficients of N, whereas the first moment of the residual vector depends on the mean curvature of submanifold $\mathbf{y}(N)$.

Hence, the first moments of the parameters can be manipulated by a change of parameter-choice, whereas the first moment of the residual vector is invariant to such a change of parameters.

As an example, let us apply (6.7) to the two dimensional Symmetric Helmert transformation (5.17). We assume that the observation space has the standard metric.

According to (5.87) the non-zero principal curvatures of model (5.17) for an arbitrary normal direction $\mathbf{N}$ read

$$k_{\mathbf{N}} = \pm \sqrt{\frac{\sum_{I=1}^{2n} (N^I)^2}{\sum_{i=1}^{p} ((u_i^c)^2 + (v_i^c)^2)}} .$$

Hence, the corresponding mean curvature reads $\bar{k}_{\mathbf{N}} = 0$. With (6.7.a) follows then that the adjusted coordinates $\hat{x}_i, \hat{y}_i$ and $\hat{\tilde{x}}_i, \hat{\tilde{y}}_i$, $i = 1, \ldots, n$, are unbiased.

The bias in the parameters follows if one applies (6.7.b.) For the Symmetric Helmert transformation one will then find that

$$
\left.
\begin{aligned}
E\{\hat{\lambda}-\lambda\} &= \frac{\sigma^2}{\sum_{i=1}^{n}\{(u_i^c)^2+(v_i^c)^2\}} \cdot \frac{1}{2} \cdot (\lambda^{-1}+\lambda) \\
E\{\hat{\theta}-\theta\} &= 0, \quad E\{\hat{t}_x - t_x\} = E\{\hat{t}_y - t_y\} = 0 .
\end{aligned}
\right\} \qquad (6.8)
$$

Similar estimates as given by (6.7) can also be derived for the higher order moments of the non-linear estimators.

Fortunately our rather pessimistic estimates in section 5 indicate that the application of the theory of linear statistical inference is generally justified in geodetic network adjustments. But, we must admit that it is not clear to us yet what to do when the model is significantly non-linear and therefore much more research needs to be done in this area. Such being the case one may be surprised to realize how little developed is the statistical theory of non-linear estimation for practical applications. See for instance the survey papers (Cox, 1977), (Bunke, 1980); the book (Goldfeld and Quandt, 1972) and the very recent book (Humak, 1984).

An alternative way to estimate the properties of the distribution of the estimators involved, would be to use computer simulation. One could replicate the series of experiments as many times as one pleases, each time with a new sample of errors drawn from the prescribed normal distribution and so obtain the relevant distributional properties by averaging over all replications. Although this approach could give us valuable insight into the effect of non-linearity, it must be carried out on a system whose parameters are known in advance, and such a system may not always be realistic. But then again, since the distributions of the estimators involved depend on the actual distribution of the observational data which on its turn depends on the "true" values **x** which are generally unknown, one is almost always faced with the problem that even when one can derive exact formulae for the distributions one can evaluate only the approximation obtained by substituting the estimated parameters for the true ones.

Finally we mention the possibility to rely on results from asymptotic theory. The central idea of asymptotic theory is that when the number m of observations is large and errors of estimation corresponding small, simplications become available that are not available in general. The rigorous mathematical development involves limiting distributional results holding as $m \to \infty$ and is closely related to the classical limit theorems of probability theory. In recent years many researchers have concentrated on developing an asymptotically theory for non-linear least-squares estimation. In (Jennrich, 1969) a first complete account was given of the asymptotic properties of non-linear least-squares estimators. And in (Schmidt, 1982) it was shown how the asymptotic theory can be utilized to formulate asymptotic exact test statistics. See also the very recent book (Bierens, 1984). Roughly speaking one can say that under suitable conditions one gets the same asymptotic results for the non-linear model as for the linear one. Unfortunately, we doubt whether the results obtained up to now can satisfy the requirements of applications in practice. In particular, the theory still seems to lack statements concerning the accuracy of the approximations by limit distributions.

7. Epilogue

In this chapter we have tried to show how contemporary differential geometry can be used to improve our understanding of non-linear adjustment. We have seen that unfortunately one can very seldom extend the elegant formulations and solution techniques from linear to non-linear situations. For most non-linear problems one will therefore have to recourse, in practice, to methods which are iterative in nature. As our analysis showed, the Gauss' method is pre-eminently suited for **small** extrinsic curvature non-linear adjustment problems. On the whole, one could say that solutions to linear problems are prefabricated, while exact solutions to non-linear problems are custom made. An important example is our inversion-free solution to the Symmetric Helmert transformation.
Although we have treated a number of new aspects of non-linear adjustment, we must recognize that we are only on the brink of understanding the complex of problems of non-linear adjustment. Many problems and topics were left untouched or were not further elaborated upon.

For instance, in our proof of the global convergence theorem (4.66) we made use of the line search strategy known as the minimization rule. However, its practical application is limited by the fact that the line search must be exact, i.e., it requires that the exact minimum point of the function $E(\mathbf{c}_q(t))$ be found in order to determine $\mathbf{x}_{q+1}$. Therefore in practice the exact minimization is replaced by an inexact line search, in particular by a finite search process (see e.g. Ortega and Rheinboldt, 1970).

In our discussion of Gauss' method, we assumed the non-linear map $\mathbf{y}$ to be injective. However, in many practical applications the matrix of first derivatives $\partial_\alpha y^i$ becomes of non-maximum rank (see e.g. chapter III) and the required inverses cannot be calculated. A way out of this dilemma is suggested by the theory of inverse linear mapping. Instead of an ordinary inverse of $g_{\alpha\beta}$, one then takes a generalized inverse, $\bar{g}^{\beta\alpha}$ say, of $g_{\alpha\beta}$. To show that $\Delta x^\beta = -\bar{g}^{\beta\alpha}\partial_\alpha E = -(\mathrm{grad}\, E)^\beta$ is still in a descent direction, note that

$$-\langle \mathbf{grad}\, E,\ \boldsymbol{\Delta x} \rangle_N = (y_s^i - y^i(x))g_{ik}\partial_\beta y^k(x)\bar{g}^{\beta\alpha}(x)\partial_\alpha y^l(x)g_{lj}(y_s^j - y^j(x)).$$

Hence, if $\boldsymbol{\Delta x} \neq \mathbf{0}$ then $-\langle \mathbf{grad}\, E,\ \boldsymbol{\Delta x} \rangle_N > 0$, which shows that $\boldsymbol{\Delta x}$ has a positive component along the negative gradient and so is downhill.
As to the local rate of convergence of Gauss' method, recall that the extrinsic curvatures are a property of the submanifold $\bar{N}$. Therefore, the local convergence results obtained for Gauss' method will remain unchanged if $\partial_\alpha y^i$ has non-maximum but local constant rank.

Of the many iteration methods available, we only discussed Gauss' method. We did not mention any of the possible alternative iteration methods such as, for instance, Newton's method, Levenberg-Marquardt's compromise or the method of conjugate-gradients (see e.g. Ortega and Rheinboldt, 1970). Although more intricate, these methods can become quite attractive in case of **large** curvature problems since they take care, in one way or the other, of the curvature behaviour of $\bar{N}$.

Also did we not discuss the interesting point of view which is provided if one interprets the iteration process as a dynamical system. Consider namely Gauss' method

$$(a) \quad \Delta x^{\beta}_{q} = g^{\beta\alpha}(x_q)\partial_{\alpha} y^{i}(x_q) g_{ij}(y^{j}_{s} - y^{j}(x_q)),$$

$$(b) \quad x^{\beta}_{q+1} = x^{\beta}_{q} + t_q \Delta x^{\beta}_{q} ,$$

and assume that the positive scalar t_q is taken infinitesimally small in each iteration step. We obtain then the autonomous dynamical system

$$\frac{dx^{\beta}}{dt} = g^{\beta\alpha}(x)\partial_{\alpha} y^{i}(x)\, g_{ij}(y^{j}_{s} - y^{j}(x)) = - (\text{grad. } E(x))^{\beta} . \tag{7.1}$$

Its solution is a curve $\mathbf{c}(t)$ which passes through the initial value $\mathbf{x}_0$ at time t=0 and which has its velocity given by the value of the vector field - **grad** E. Although the uniqueness theorem for systems of differential equations implies that $\mathbf{c}(t)$ is never a critical point of E, this should not bother us too much since one can show that under suitable conditions $\lim_{t\to\infty} \mathbf{c}(t) = \hat{\mathbf{x}}$ with $\mathbf{grad}.E(\hat{\mathbf{x}}) = \mathbf{0}$. This is like the pendulum paradox, which says that the pendulum once it is in motion can never come to a state of rest, but only approximate one arbitrary closely. Thus, given an initial guess $\mathbf{x}_0$ which is not a critical point of E, one can try to solve our non-linear adjustment problem by solving the system of differential equations (7.1), using one of the many numerical integration methods available.

In connection with the above dynamical interpretation we also mention the potential value which a study of the qualitative theory of the global behaviour of dynamical systems and of Morse theory, can have for a betterment of our understanding of non-linear adjustment. This qualitative theory is namely concerned with the existence of equilibrium behaviour of a dynamical system, together with questions of local and global stability (see e.g. Chillingworth, 1976; Hirsch and Smale, 1974). And Morse theory studies, amongst other things, the equilibrium configuration of a gradient system. The Morse inequalities, for instance, place restrictions on the number of critical points that a function E can have due to the topology of the manifold on which it is defined (see e.g. Hirsch, 1976).

Finally we note that we omitted the important case of an implicitly defined submanifold $\bar{N}$. This would correspond to a non-linearly constrained adjustment problem. Although the geometry of the problem is not too different from the one discussed in this chapter, the various methods for actually solving a constrained problem can become quite involved (see e.g. Hestenes, 1975). The usual way to go about is, to prolong the original constrained problem with the aid of the Langrange multiplier rule to one which is unconstrained. It is interesting to point out that although the Lagrange multipliers are often thought of as being merely dummy variables, which are just needed to prolong the constrained problem into an unconstrained one, they actually have an important interpretation of their own. In fact, there is a very rich duality theory connected with the Lagrangian formulation (see e.g. Rockafellar, 1969). It goes back to the Legendre transformation of classical mechanics. The Lagrangian formulation has namely the physical significance that it replaces the given (kinematical) constraints by forces which maintain those constraints. As a result the multipliers equal the forces of

reaction (see e.g. Krarup, 1982b). The multipliers can therefore be used as test statistics. For linear models one can show that the standardized Lagrangian multiplier equals Baarda's w-test statistic (see Teunissen, 1984b).

That many more problems and topics related to non-linear adjustment can be brought forward is indisputable. Many questions are still open for future research and it will probably take some time before we understand non-linear geodesic adjustment as well as we understand linear adjustment. We therefore conclude by expressing the wish that the rather unsurveyed area of non-linear adjustment and statistical inference will receive more serious attention than it has received hitherto.

REFERENCES

Adám, J., F. Halmos and M. Varga (1982): On the Concepts of Combination of Doppler Satellite and Terrestrial Geodetic Networks, Acta Geodaet., Geophys. et Montanist. Acad. Sic. Hung. Volume 17(2), pp. 147-170.

Alberda, J.E. (1969): The Compliance of Least-Squares Estimates with the Condition Equations (In Dutch: "Het Voldoen van Kleinste-Kwadraten schattingen aan de Voorwaardevergelijkingen"), Laboratorium voor Geodetische Rekentechniek, R. 65, Delft.

Baarda, W. (1967a): Adjustment Theory, Part One (In Dutch: "Vereffeningstheory, Eerste deel"), Laboratorium voor Geodetische Rekentechniek, Delft.

Baarda, W. (1967b): Statistical Concepts in Geodesy, Netherlands Geodetic Commission, Publications on Geodesy, New Series, Vol. 2, No. 4, Delft.

Baarda, W. (1969): A Testing Procedure for Use in Geodetic Networks, Netherlands Geodetic Commission, Publications on Geodesy, New Series, Vol. 5, No. 1, Delft.

Baarda, W. (1973): S-transformations and Criterion Matrices, Netherlands Geodetic Commission, Publications on Geodesy, New Series, Vol. 5, No. 1, Delft.

Baarda, W. (1978): Mathematical Models, European Organisation for Experimental Photogrammetric Research, Publ. Off. Nr. 11, pp. 73-101.

Baarda, W. (1979): A Connection between Geometric and Gravimetric Geodesy; A first sketch, Netherlands Geodetic Commission, Publications on Geodesy, New Series, Vol. 6, No. 4, Delft.

Backus, G. and F. Gilbert (1968): The Resolving Power of Gross Earth Data, Geophys. J.R. astr. Soc., 16, pp. 169-205.

Bierens, H.J. (1984): Robust Methods and Asymptotic Theory in Nonlinear Econometrics, Lecture Notes in Economics and Mathematical Systems, Springer-Verlag, Vol. 192.

Bjerhammar, A. (1951): Rectangular Reciprocal Matrices, with Special Reference to Geodetic Calculations, Bull. Géod., 20, pp. 188-220.

Bjerhammar, A. (1973): Theory of Errors and Generalized Matrix Inverses, Elsevier, Amsterdam.

Blais, J.A.R. (1983): Duality Considerations in Linear Least-Squares Estimation, Manuscripta Geodaetica, Vol.8, No.2, pp. 199-213.

Blaha, G. (1984): Tensor Structure Applied to the Least-Squares Method, Revisited, Bull. Géod. 58, pp. 1-30.

Brouwer, F.J.J., D.T. van Daalen, F.T. Gravesteijn, H.M. de Heus, J.J. Kok and P.J.G. Teunissen (1982): The Delft Approach for the Design and Computation of Geodetic Networks, In: "Forty Years of Thought..." Anniversary edition on the occasion of the 65th birthday of Professor W. Baarda., Vol. I, pp. 202-274.

Bunke, H. (1980): Parameter Estimation in Nonlinear Regression Models, in Handbook of Statistics (P.R. Krishnaiah, ed.), Vol. 1, North-Holland Publishing Company, pp. 593-615.

Celmins, A. (1981): Least-Squares Model Fitting with Transformations of Variables, J. Statist. Comput. Simul. Vol. 14, pp. 17-39.

Celmins, A. (1982): Estimation of NMR Function Accuracies from Least-Squares Fitting, Journal of Magnetic Resonance, 50, pp. 373-381.

Chillingworth, D.R.J. (1976): Differential Topology with a View to Applications, Research Notes in

Mathematics, No. 9, Pitman Publishing.

Cox, D.R. (1977): Nonlinear Models, Residuals and Transformations, Math. Operationsforsch. Statist. Ser. Statistics, Vol. 8, No. 1, pp. 3-22.

Eeg, J. (1982): Continuous Methods in Least-Squares Theory, Bollettino di Geodesia e Scienze Affini, Nr. 4, pp. 393-407.

Engler, K., E. Grafarend, P. Teunissen, and J. Zaiser (1982): Testcomputations of Three-Dimensional Geodetic Networks with Observables in Geometry and Gravity Space. DGK, Reihe B, Heft Nr. 258/VII, München, pp. 119-141.

Flemming, W. (1977): Functions of Several Variables, Springer Verlag.

Gauss, C.F. (1827): Algemeine Flächentheorie (Disquisitiones Generales Circa Superficies Curvas), Deutsch herausgegeben von A. Wangerin, Leipzig, 1889.

Gauss, C.F. (1887): Abhandlungen zur Methode der Kleinsten Quadrate, Deutsch herausgegeben von Königl. Preussischen Geodätischen Institut, Berlin 1887.

Goldfeld, S.M. and R.E. Quandt (1972): Nonlinear Methods in Econometrics, Amsterdam, North-Holland Publishing Company.

Grafarend, E. (1970): Verallgemeinerte Methode der kleinsten Quadraten für Zyklische Variabele, ZfV. 95, Heft 4, pp. 117-121.

Grafarend, E. (1973): Attempts for a Unified Theory of Geodesy, Bull. Géod., 109, pp. 237-260.

Grafarend, E. and B. Schaffrin (1974): Equivalence of Estimable Quantities and Invariants in Geodetic Networks, ZfV 101, pp. 485-491.

Grafarend, E., H. Heister, R. Kelm, H. Kropff and B. Schaffrin (1979): Optimierung Geodätischer Messoperationen, Herbert Wichmann Verlag, Karlsruhe, Band II.

Grafarend, E.W. (1981): Kommentar eines Geodäten zu einer Arbeit E.B. Christoffels, in E.B. Christoffel, The Influence of his Work on Mathematics and the Physical Sciences. Edited by P.L. Butzer and F. Fehér, Birkhäuser Verlag.

Grafarend, E.W., E.H. Knickmeyer and B. Schaffrin (1982): Geodätische Datumstransformationen, ZfV., No. 1, pp. 15-25.

Griffiths, L.W. (1947): Introduction to the Theory of Equations, John Wiley & Sons, Inc., New York.

Heiskanen, W.A. and H. Moritz (1967): Physical Geodesy, Freeman and Co., San Francisco/London.

Helmert, F.R. (1880): Die Mathemat. und Physikal. Theorieen der Höheren Geodäsie, Leipzig.

Hestenes, M.R. (1975): Optimization Theory, The Finite Dimensional Case, John Wiley, New York.

Hirsch, M.W. (1976): Differetial Topology, Springer-Verlag.

Hirsch, M.W. and S. Smale (1974): Differential Equations, Dynamical Systems and Linear Algebra, Academic Press, New York.

Hotine, M. (1969): Mathematical Geodesy, ESSA Monograph 2, Washington.

Humak, K.M.S. (1984): Statistische Methoden der Modelbildung, Bd. II, Akademie-Verlag, Berlin.

Jackson, J. (1982): Survey Adjustment, Survey Review, Vol. 26, No. 203, pp. 248-249.

Jennrich, R.I. (1969): Asymptotic Properties of Nonlinear Least Squares Estimators, The Annals of Mathematical Statistics, 40, pp. 633-643.

Kelley, R.P. and W.A. Thompson jr. (1978): Some Results on Nonlinear and Constrained Least Squares, Manuscripta Geodaetica, Vol. 3, pp. 299-320.

Köchle, R. (1982): Die Räumliche Helmert transformation in Algebraischer Darstellung, Vermessung, Photogrammetrie, Kulturtechnik, 9, pp. 292-297.

Kooimans, A.H. (1958): Principles of the Calculus of Observations, Rapport Spécial, Neuvième Congrès International des Géomères, Pays-Bas, pp. 301-310.

Krarup, T. (1969): A Contribution to the Mathematical Foundation of Physical Geodesy, Geod. Inst. København, Medd. No. 44.

Krarup, T. (1972): On the Geometry of Adjustment, ZfV., Heft 10, 97, pp. 440-445.

Krarup, T. (1982a): Non-Linear Adjustment and Curvature, In: Daar heb ik veertig jaar over nagedacht... Feestbundel ter gelegenheid van de 65ste verjaardag van professor Baarda, Delft, pp. 145-159.

Krarup, T. (1982b): Mechanics of Adjustment, Peter Meissl - Gedenkseminar, Geodätische Institute, T.U. Graz.

Kube, R. and K. Schnädelbach (1975): Geometrical Adjustment of the European Triangulation Networks - Report of the RETrig Computing Centre München, AIG, Section I, Publication No. 11.

Kubik, K. (1967): Iterative Methoden zur Lösung des Nichtlinearen Ausgleichsproblems, ZfV., Nr. 6, pp. 214-225.

Levallois, J.J. (1960): La Réhabilitation de la Géodésie Classique et la Géodésie Tridimensionelle, Bull. Géod., No. 68, pp. 193-199.

Marussi, A. (1952): Intrinsic Geodesy, The Ohio State Research Foundation, Project No. 485, Columbus.

Meissl, P. (1973): Distortions of Terrestrial Networks caused by Geoid Errors, Bolletino di Geodesia e Scienze Affini, N. 2, pp. 41-52.

Meissl. P. (1982): Least Squares Adjustment, A Modern Approach, Mitteilungen der geodätischen Institute der Technischen Universität Graz, Folge 43.

Molenaar, M. (1981a): A Further Inquiry into the Theory of S-transformations and Criterion Matrices, Netherlands Geodetic Commission, Vol. 7, Nr. 1, Delft.

Molenaar, M. (1981b): S-transformations and Artificial Covariance Matrices in Photogrammetry, ITC Journal, No. 1, pp. 70-79.

Morduchow, M. and L. Levin (1959): Comparison of the Method of Averages with the Method of Least-Squares: Fitting a Parabola. Presented at the 557th Meeting of the American Mathematical Society, New York.

Moritz, H. (1979): The Geometry of Least-Squares, Publications of the Finnish Geodetic Institute, No. 89, pp. 134-148.

Neeleman, D. (1973): Multicollinearity in Linear Economic Models, Tilburg University Press.

Nibbelke, P. (1984): Adjustment of Geodetic Networks on the Ellipsoid (In Dutch: "Vereffening van geodetische netwerken op de ellipsoïde"), thesis.

Ortega, J.M. and W.C. Rheinboldt (1970): Iterative Solution of Nonlinear Equations in Several Variables, Academic Press.

Penrose, R. (1955): A Generalized Inverse for Matrices, Proc. Cambridge Philos. Soc., 51, pp. 406-413.

Peterson, A.E. (1974): Merging of the Canadian Triangulation Network with the 1973 Doppler Satellite Data, The Canadian Surveyor, Vol. 28, No. 5, pp. 487-495.

Pope, A. (1972): Some Pitfalls to be Avoided in the Iterative Adjustment of Nonlinear Problems, Proceedings of the 38th Annual Meeting, American Society of Photogrammetry.

Pope, A. (1974): Two Approaches to Nonlinear Least-Squares Adjustments, The Canadian Surveyor, Vol. 28, No. 5, pp. 663-669.

Rao, C.R. (1973): Linear Statistical Inference and its Applications, Wiley, New York.

Rao, C.R. and S.K. Mitra (1971): Generalized Inverse of Matrices and its Applications, J. Wiley, New York.

Rockafellar, R.T. (1969): Convex Analysis, Princeton University Press, Princeton, N.J.

Rummel, R. and P.J.G. Teunissen (1982): A Connection between Geometric and Gravimetric Geodesy - Some Remarks on the Role of the Gravity Field, In: "Forty Years of Thought..." Anniversary edition on the occasion of the 65th birthday of Professor W. Baarda., Vol. II, pp. 602-623.

Rummel, R. (1984): From the Observational Model to Gravity Parameter Estimation, Lecture Notes of the International Summer School on Local Gravity Field Approximation, Beijing, China, Aug. 21 to Sept. 4.

Sansò, F. (1973): An Exact Solution of the Roto-Translation Problem, Photogrammetria, 29, pp. 203-216.

Schmidt, W.H. (1982): Testing Hypotheses in Nonlinear Regressions, Math. Operationsforsch. Statist., Secr. Statistics, Vol. 13, 1 , pp. 3-19.

Schwidefsky, K. and F. Ackermann (1975): Photogrammetrie, Grundlagen, Verfahren, Anwendungen, B.G. Teubner, Stuttgart.

Spivak, M. (1975): Differential Geometry, Vol. 1-5, Publish or Perish Inc.

Stark, E. and E. Mikhail (1973): Least-Squares and Non-Linear Functions, Photogrammetric Engineering, pp. 405-412.

Stoker, J.J. (1969): Differential Geometry, Wiley-Interscience.

Strang van Hees, G.L. (1977): Orientation of the Ellipsoid in Geodetic Networks, Delft Progress Report, 3, pp. 35-38.

Teunissen, P.J.G. (1980): Some Remarks on Gravimetric Geodesy, Reports of the Department of Geodesy, Section Mathematical and Physical Geodesy, No. 80.2, Delft.

Teunissen, P.J.G. (1982): Anholonomity when using the Development Method for the Reduction of Observations to the Reference Ellipsoid, Bull. Géod. 56, no. 4, pp. 356-363.

Teunissen, P.J.G. (1983): A Note on Anholonomity, Paper presented at the meeting on Geometric Geodesy, IUGG/AIG general assembly, Hamburg 15-27 August 1983.

Teunissen, P.J.G. (1984a): Generalized Inverses, Adjustment, The Datum Problem and S-transformations, Lecture Notes, International School of Geodesy, 3rd Course: Optimization and Design of Geodetic Networks, Erice-Trapani-Sicily, 25 April - 10 May 1984.

Teunissen, P.J.G. (1984b): Quality Control in Geodetic Networks, Lecture Notes, International School of Geodesy, 3rd Course: Optimization and Design of Geodetic Networks, Erice-Trapani-Sicily, 25 April - 10 May 1984.

Teunissen, P.J.G. (1984c): A Note on the Use of Gauss' formulas in Non-Linear Geodesic Adjustment, Paper presented at the 16th European Meeting of Statisticians, Marburg (FRG), 3-7 Sept. 1984.

Tienstra, J.M. (1947): An Extension of the Technique of the Method of Least-Squares to Correlated Observations, Bull. Géod., 6, pp. 301-335.

Tienstra, J.M. (1948): The Foundation of the Calculus of Observations and the Method of Least-Squares, Bull. Géod., 10, pp. 289-306.

Tienstra, J.M. (1956): Theory of the Adjustment of Normally Distributed Observations, N.V.

Uitgeverij Argus, Amsterdam.

Torge, W. and H.G. Wenzel (1978): Dreidimensionale Ausgleichung des Testnetzes Westharz, DGK, Report B234, München.

Vanicek, P. (1979): Tensor Structure and the Least-Squares, Bull. Géod., Vol. 53, No. 3, pp. 221-225.

Van Mierlo, J. (1978): A Testing Procedure for Analysing Geodetic Deformation Measurements, F.I.G. Symp. on Deformation Measurements, Bonn.

Van Mierlo, J. (1979): Free Networks Adjustment and S-transformations, DGK B, Nr. 252, pp. 41-54.

Whittaker, E. and G. Robinson (1944): The Calculus of Observations, Dover Publications, Inc.

Wolf, H. (1963a): Die Grundgleichungen der Dreidimensionalen Geodäsie in elementarer Darstellung, ZfV, Nr. 6, pp. 225-233.

Wolf, H. (1963b): Geometric Connection and Re-Orientation of Three-Dimensional Triangulation Nets, Bull. Géod., No. 68, pp. 165-169.

Wolf, H. (1978): The Helmert Block Method - Its Origin and Development. Proceedings Second International Symposium on Problems Related to the Redefinition of North American Geodetic Networks, pp. 319-326.

Yeremeyev, V.F. and M.I. Yurkina (1969): On Orientation of the Reference Geodetic Ellipsoid, Bull. Géod., No. 91, pp. 13-15.