Quality assessment of satellite-based global gravity field models

J. Bouman

NCG Nederlandse Commissie voor Geodesie Netherlands Geodetic Commission

Delft, June 2000

Colophon

Quality assessment of satellite-based global gravity field models
J. Bouman
Publications on Geodesy 48
ISBN 90 6132 270 7
ISSN 0165 1706

Publications on Geodesy is the continuation of Publications on Geodesy New Series

Published by: NCG Nederlandse Commissie voor Geodesie Netherlands Geodetic Commission, Delft, The Netherlands
Printed by: Meinema Drukkerij, Delft, The Netherlands
Cover: Geoid height errors caused by model errors

NCG Nederlandse Commissie voor Geodesie
P.O. Box 5030, 2600 GA Delft, The Netherlands
Tel.: +31 (0)15 278 28 19
Fax: +31 (0)15 278 17 75
E-mail: ncg@geo.tudelft.nl
Website: www.ncg.knaw.nl

The NCG Nederlandse Commissie voor Geodesie Netherlands Geodetic Commission is an institute of the Royal Netherlands Academy of Arts and Sciences (KNAW).

Contents

Summary

Quality assessment of satellite-based global gravity field models

A global model of the Earth's gravity field may be derived from satellite observations such as satellite tracking observations or satellite gravity gradiometry (SGG). Due to the satellite's altitude above the Earth's surface, the gravity field recovery from satellite data is an ill-posed problem since it lacks stability. In addition, the spatial data distribution is heterogeneous in general. A physically sensible solution may therefore only be obtained by imposing constraints on the solution or by using additional surface data. The former is called regularisation.

The regularisation methods used here are all filtered least-squares solutions. Several methods are discussed as well as how to choose the filter or regularisation parameter. Regularisation yields stable solutions at the expense of introducing bias into the solution. The more the least-squares solution is filtered, the more stable the solution is and the more bias is introduced. The solution error is described by the mean square error matrix (MSEM) which includes propagated data noise and bias as well. The optimal choice for the regularisation parameter will minimise the trace of the MSEM, the so-called mean square error.

The MSEM plays a key role in quality assessment. It describes not only the accuracy of the estimated unknowns but can also be used for error propagation to derived products, computation of signal-to-noise ratio (SNR), and so on. A number of quality measures based on the MSEM is discussed. Of the quality measures, the error propagation and the ratio measures such as SNR, appear to be useful, whereas the contribution measures, which reflect the importance of the observations relative to the constraints, are less useful.

In order to compare the different regularisation methods a simulation study is conducted for SGG. Specifically, Tikhonov regularisation (TR) with a number of different constraints is tested as well as ordinary and generalised biased estimation. A global gravity field was determined with a spatial resolution of approximately 1° using simulated SGG data leaving the polar areas uncovered. The regularisation methods perform equally well in the observation area. TR with signal constraint, however, seems to give the smallest mean square error when the unsurveyed areas are taken into account as well. Furthermore, if the error model of the observations is correct, then error analysis and simulation yield comparable conclusions on the quality.

Therefore, TR with signal constraint in combination with error analysis are used to investigate the effect of additional data as well as the effect of a change in resolution. The gradiometric observations are combined with high-low satellite-to-satellite tracking (SST) data and airborne gravimetry. The SST data are especially useful to compensate for the SGG coloured noise errors and are slightly less sensitive to data gaps. If surface gravity data are available in the uncovered areas then the ill-conditioning is largely reduced. These data, however, are not suited to overcome the ill-conditioning due to downward continuation.

The most important results are that, if only satellite data are available, the bias cannot be neglected and should be accounted for in the quality assessment. Furthermore, the bias is concentrated in the unsurveyed areas but many gravitational potential coefficients may be affected. A full understanding of the quality of a gravity field model can only be obtained by considering several quality measures. Moreover, conclusions about the quality may depend on the desired final product. Gravity anomalies, for example, are less affected by the SGG coloured noise than geoid heights.

Samenvatting

Kwaliteitsbeschrijving van op satellietmetingen gebaseerde mondiale zwaartekrachtsveldmodellen

Een mondiaal model van het aardse zwaartekrachtsveld kan worden bepaald met behulp van satellietmethoden als satellietbaanbeschrijving en satellietgravitatiegradiometrie (SGG). Vanwege de hoogte van een satelliet boven het aardoppervlak, kan de zwaartekrachtsveldbeschrijving uit satellietmetingen worden gekarakteriseerd als een slecht gesteld invers probleem omdat het instabiel is. Bovendien is de ruimtelijke dataverdeling in het algemeen heterogeen. Daarom kan een fysisch zinvolle oplossing alleen worden verkregen door randvoorwaarden aan de oplossing op te leggen of door additionele oppervlaktedata te gebruiken. Het eerste wordt regularisatie genoemd.

Alle hier gebruikte regularisatiemethoden zijn gefilterde kleinstekwadratenoplossingen. Een aantal methoden wordt besproken alsook de bepaling van de filter- of regularisatieparameter. Door middel van regularisatie worden stabiele oplossingen verkregen ten koste van de zuiverheid van de oplossing. Hoe meer de kleinstekwadratenoplossing wordt gefilterd, des te stabieler en onzuiverder de oplossing wordt. Door middel van de gemiddelde-kwadratischefoutenmatrix ('mean square error matrix' MSEM), die bestaat uit zowel onzuiverheid als voortgeplante meetfout, worden de fouten in de oplossing beschreven. De optimale keuze van de regularisatieparameter minimaliseert het spoor van de MSEM.

De MSEM vervult een sleutelrol in kwaliteitsbeschrijving. Niet alleen beschrijft deze de nauwkeurigheid van de geschatte onbekenden maar ze kan ook worden gebruikt voor foutenvoortplanting naar afgeleide producten, berekening van signaal-ruisverhouding (SNR), enzovoort. Een aantal op de MSEM gebaseerde kwaliteitsmaten wordt behandeld. Van deze blijken de foutenvoortplanting en de verhoudingsmaten zoals de SNR bruikbaar, terwijl de bijdragematen, die de invloed van de waarnemingen ten opzichte van de randvoorwaarden beschrijven, minder bruikbaar zijn.

De verschillende regularisatiemethoden worden vergeleken met behulp van een simulatiestudie voor SGG. Tikhonov regularisatie (TR) met verschillende randvoorwaarden en gewone en gegeneraliseerde onzuivere schatting worden getest. Daartoe wordt een mondiaal zwaartekrachtsveld bepaald met een ruimtelijke resolutie van ongeveer 1° uit gesimuleerde SGG-data die de poolgebieden niet bedekken. De regularisatiemethoden geven gelijke resultaten in het gebied met metingen, maar in de niet bedekte gebieden presteert TR met signaalrandvoorwaarde het best. Bovendien kan geconcludeerd worden dat, indien het foutenmodel van de waarnemingen correct is, foutenanalyse en simulatie vergelijkbare conclusies aangaande de kwaliteit opleveren.

Dat is de reden dat het effect van additionele data en het effect van een veranderende resolutie wordt onderzocht met TR met signaalrandvoorwaarde in combinatie met foutenanalyse. De gradiometrische waarnemingen worden gecombineerd met satelliet-naar-satellietmetingen (SST) en vliegtuiggravimetrie. De SST data zijn speciaal geschikt ter compensatie van de gekleurde ruis op de SGG-waarnemingen. Bovendien zijn ze iets minder gevoelig voor een heterogene bedekking. De slechte conditionering wordt grotendeels gereduceerd indien oppervlaktezwaartekrachtdata beschikbaar zijn in de niet bedekte gebieden. Deze data zijn echter niet geschikt om de slechte conditionering ten gevolge van de neerwaartse voortzetting op te heffen.

De belangrijkste resultaten zijn dat, als alleen satellietdata beschikbaar zijn, de onzuiverheid niet verwaarloosd kan worden en meegenomen moet worden in de kwaliteitsbeschrijving. De onzuiverheid concentreert zich in de gebieden zonder metingen, maar vele zwaartekrachtspotentiaalcoëfficiënten kunnen worden aangetast. Een volledig inzicht in de kwaliteit van een zwaartekrachtsveldmodel kan alleen worden verkregen door meerdere kwaliteitsmaten te beschouwen. Voorts kunnen de conclusies aangaande de kwaliteit afhangen van het eindproduct. Zo zijn bijvoorbeeld zwaartekrachtsanomaliën minder gevoelig voor de gekleurde ruis van SGG dan geoïdehoogtes.

Acknowledgements

Many people contributed to this study either by collaboration or moral and pratical support. First of all I thank Roland Klees for the discussions we had and his remarks on earlier versions of this thesis. My deep gratitude goes to Radboud Koop for his daily supervision, wisdom and our discussions, scientific or not. He also made major contributions to the SGG synthesis and analysis software. José van den IJssel computed the satellite orbits and contributed to the synthesis and analysis software. I'm especially thankful for our daily contact and the discussions on results and practical problems. Martin van Gelderen wrote the interpolation part of the SGG synthesis and commented on a draft version of the study. With Rune Floberghagen I worked with much pleasure on the 'Moon Project'. Pieter Visser provided the SGG coloured noise errors and commented on an earlier version of this thesis. Nico Sneeuw provided the SST normal matrix and some background information. The information on airborne gravimetry by Klaus Peter Schwarz and Christian Tscherning is gratefully acknowledged. Clare Macfarlane and Job Oostveen read a draft version and made useful remarks. Parts of the software are written in C++ which is facilitated by the package `newmat` developed by Robert Davies. This work is supported by the Delft University of Technology's Centre for High Performance and Applied Computing (HPαC). Special thanks to my parents whose support throughout the years is indispensible. Last but not least I thank Kyra van Onselen for sharing a room for so many years and all other colleagues at DEOS who contributed to a nice working environment.

Abbreviations

BE	biased estimation
BNR	bias-to-noise ratio
BSR	bias-to-signal ratio
CPU	central processing unit
DGSVD	damped generalised singular value decomposition
DSVD	damped singular value decomposition
DTM	digital terrrain model
EGM96	Earth gravity model 1996
FFT	fast Fourier transform
GBE	generalised biased estimation
GCV	generalised cross validation
GOCE	gravity field and steady-state ocean circulation explorer
GPS	global positioning system
GRACE	gravity recovery and climate experiment
GRS80	geodetic reference system 1980
GSVD	generalised singular value decomposition
JGM-3	joint gravity model 3
l.s.	least squares
MDB	minimal detectable bias
MSE	mean square error
MSEM	mean square error matrix
OSU91A	Ohio State University gravity model 91A
POD	precise orbit determination
PSD	power spectral density
RMS	root mean square
SA	selective availability
SGG	satellite gravity gradiometry
SNR	signal-to-noise ratio
SST	satellite-to-satellite tracking
SVD	singular value decomposition
TF	timewise approach in the frequency domain
TGSVD	truncated generalised singular value decomposition
TR	Tikhonov regularisation
tr	trace
TSVD	truncated singular value decomposition
TT	timewise approach in the time domain

1

Introduction

Background and problem description

A model of the Earth's gravity field is used in several geosciences, and throughout the years many of such models have been computed (for an overview see e.g. Bouman, 1998c). Since the beginning of the space age satellites have been used for the determination of global gravity field models. Satellites in free fall around the Earth move under the influence of the Earth's gravity field. Hence, tracking these satellites may yield gravity field information, and the models thus determined are called satellite-only models. At first visible objects were photographed against the star background, whereas later Doppler measurements and satellite laser ranging were used. With time satellite tracking improved, and because of the increased number of satellites and tracking stations, the spatial distribution of the measurements improved as well (e.g. Reigber, 1989). Despite this progress, the distribution still is non-homogeneous. Another major drawback is the inherent instability of gravity field determination from satellite tracking data. In other words, a small error in the data may lead to a large error in the gravity field model. This is a consequence of the strong damping of the high gravity field frequencies at satellite altitude. Vice versa, the measurement noise is amplified through downward continuation (cf. Rummel *et al.*, 1979).

The standard method to overcome such instabilities is to add prior information to the solution. The quality description of the solution starts with interpreting the method in the framework of unbiased least-squares collocation, sometimes also called constraint least squares (Marsh *et al.*, 1988; Schwintzer *et al.*, 1997). It is, however, generally recognised that the satellite-only models are biased, but that the bias is usually not accounted for (Marsh *et al.*, 1988; Xu, 1992b). In addition to the standard stabilisation method a number of alternative methods exists, of which most have not been studied thoroughly in geodetic literature (cf. Louis, 1989; Engl *et al.*, 1996; Hansen, 1997).

It is well known that only the long wavelengths (about 600 km at the equator, corresponding to spherical harmonic degree and order 70) of the gravity field are revealed by the available satellite tracking data (Nerem *et al.*, 1994; Schwintzer *et al.*, 1997). The combination of satellite tracking data with terrestrial gravimetry and satellite altimetry allows for resolving shorter wavelengths down to about 100 km at the equator, corresponding to spherical harmonic degree and order 360 (Rapp *et al.*, 1991; Gruber *et al.*, 1995), although the accuracy of such combined models at different frequencies and locations is far from homogeneous. The combined models as well as the satellite-only models suffer from an improper quality description. On the one hand this is due to model errors, e.g. insufficient modelling of atmospheric drag, the aliasing of oceanographic signals in altimetry, and systematic errors in terrestrial gravity data

(Nerem *et al.*, 1994; Heck, 1990). On the other hand, as said before, the satellite-only models are biased and these are the basis for the combined models.

In the near future several dedicated gravity field missions are likely to be launched. 'New' measurements techniques will be applied, specifically satellite gravity gradiometry (SGG) and satellite-to-satellite tracking (SST). Two of these missions are GRACE, using high-low and low-low SST (Tapley, 1996), and GOCE, using a combination of high-low SST and SGG (ESA, 1999). The purpose of these missions is to very accurately determine high resolution stationary gravity field models (GOCE), and time-varying gravity signatures (GRACE). The expected measurement precision of the new techniques is rather high and the required resolution and accuracy of the derived solution is much better than state-of-the-art gravity field models. However, as noted above, it is unclear as to how the accuracy should be described. When the bias is taken into account the accuracy description might be different from the conventional accuracy description. Furthermore, it is of interest to consider alternative estimation methods and to compare those with the standard solution method without ignoring the bias.

Since a mission such as GOCE leaves small parts of the Earth at the poles un-surveyed, the mission might be supported by gravity data obtained from airborne gravimetry in order to cover the whole Earth with measurements. Heretofore it is unknown how these data influence the quality of the solution, especially the impact on the unknown bias. Moreover, the resolution of the gravity field model is not fixed beforehand. A resolution increase has effect on the quality of the global model.

Objectives

The above may be summarised by stating that the determination of the Earth's gravity field from satellite observations (satellite tracking, SGG) is an ill-posed problem and that the solution requires regularisation (e.g. Rummel *et al.*, 1979). Several methods of regularisation exist which stabilise the solution at the expense of introducing a bias. *The objective of this thesis* is to study the effect of different measurement types, parameter estimation methods and resolutions on the quality of Earth gravity field models based on satellite measurements with emphasis on the bias.

To this end a simulation study is conducted for SGG as well as an error analysis for the combination of SGG with SST and gravimetry. The quality is assessed by showing differences between missions, parameter estimation methods, etc. in terms of bias, geoid errors, mean square error, and so on. Although supported by numbers such as "the RMS geoid height error is 2 cm", the quality should mainly be understood in a relative sense, for example, one method introduces less bias than another method.

Further context and limitations

Notwithstanding the fact that the measurements are inherently discrete and finite, and although only a finite number of unknowns can be solved for, the characteristics of ill-posed problems are discussed by studying continuous functions and operators defined on infinite dimensional spaces. The standard example of an unstable ill-posed problem is the integral equation of the first kind with compact operator (e.g. Kress, 1989). The problems solving this equation carry over to the finite dimensional problem, and therefore the integral equation of the first kind provides the underlying concept for the discrete case. However, as soon as quality description is involved, always finite dimensions are used since the measurements dealt with are discrete and finite, and the number of estimated parameters is finite as well.

The observational model, relating the unknowns to the observations, is assumed to be linear or linearised. In our case the approximate values for the linearisation come from the reference model GRS80, iteration could account for the non-linear effects. In this thesis the unknowns consist only of gravitational coefficients of a spherical harmonic series. The observations will include gravity gradiometry, orbit perturbations derived from satellite-to-satellite tracking and gravimetry. Note that, although the observation model is linear, a solution method such as conjugated gradients is not. This will hamper the quality

description, and therefore the non-linear methods are not discussed here although they may be of interest too (cf. Schuh, 1996).

The analysis of a real gravity mission would require the inclusion of many more parameters than just spherical harmonic coefficients, such as station coordinates, tides, polar motion, time biases, variation in time of the harmonic coefficients, etc. In the actual data processing the estimation of the unknowns is conveniently split up into two steps (cf. Reigber, 1989). First, the so-called internal parameters are estimated, that is, the unknowns related to specific data arcs, such as initial state vectors. Secondly, the external parameters are estimated, such as the spherical harmonics and tidal terms. The orbit is assumed to be known with high enough accuracy relative to the accuracy of the model and the noise in the observations. The precise orbit determination (POD) is the first step (e.g. Davis, 1997). As far as the second step is concerned, all but the spherical harmonic coefficients are considered to be nuisance parameters. Undoubtedly, the inclusion of more unknowns influences the quality of the gravity field solution. However, the estimation of the harmonics fit into the context of the regularisation of ill-posed inverse problems. Therefore, the influence of the parameter estimation methods on the estimation of the spherical harmonics is investigated, whereas the effect of estimating other parameters than spherical harmonics is not.

This study is confined to *global* gravity field models using spherical harmonics. For local gravity field determination using splines, for example, it is referred to (Thalhammer, 1995; Schneider, 1997). Furthermore, the surface of the Earth is assumed to be a sphere of radius R, which is a severe assumption that neglects complications such as the downward continuation to the actual topographic surface (see e.g. Hotine, 1967). However, the interest is in a qualitative analysis supported by a quantitative analysis. It is the purpose to compare regularisation methods and satellite missions, which, to a certain extent, eliminates questions as how to refer to the actual topographic surface.

The quality description is limited to the precision or mean square error of the individual coefficients. It may include bias and propagated noise. The errors are evaluated in several ways. The signal-to-noise ratio is considered, the errors will be propagated to other gravity field functionals such as geoid heights or gravity anomalies, and the contribution of the observations to the solution is studied with respect to the constraints or a priori information. A more complete quality description could include, for example, internal and external reliability. However, these measures are derived under the assumption of unbiased estimation, and are therefore not discussed.

Outline

The outline of this thesis is as follows. First, the instability of the least-squares solution is clarified by looking at the singular value decomposition of the linear model. Moreover, parameter estimation methods (the term ‘solution methods’ is sometimes used as well) to overcome the instability are discussed as well as the corresponding mean square error matrix in chapter 2. In chapter 3 several quality measures will be presented, that is, the mean square error matrix is looked upon from different view points which gives an idea of the overall quality of the solution. Then the observations are discussed as well as their relation to the unknowns in chapter 4. The necessary simplifications leading to a linear model and a block-diagonal normal matrix are treated.

Within the framework of these three chapters it is possible to compare the solution methods using simulations. Chapter 5 deals with representative SGG examples. First, a polar circular orbit is studied and secondly a GOCE-like orbit is considered. Then gradiometric measurements are combined with satellite-to-satellite tracking and (airborne) gravimetry using error analysis in chapter 6. Finally, the main conclusions and recommendations can be found in chapter 7.

Related work and references

Schwarz (1979) and Neyman (1985) treat ill-posed problems and their regularisation in a geodetic con-

text. Rauhut (1992) compares regularisation methods for local gravity anomaly determination from satellite altimeter data, which is known as the inverse Stokes problem. However, no attention is given to the bias. Biased estimation in connection with geopotential fields has been studied by Xu (1992a,b); Xu and Rummel (1994a). Xu and Rummel (1994b) compare several biased estimators for local gravity anomaly determination from SGG. Bouman (1998b) compares a number of regularisation methods and regularisation parameter choices by means of the downward continuation of airborne gravimetric data. There is a vast amount of non-geodetic literature on regularisation methods: e.g. Phillips (1962); Tikhonov (1963a,b); Tikhonov and Arsenin (1977); Nashed (1976); Vinod and Ullah (1981); Morozov (1984); Groetsch (1984, 1993); Kress (1989); Louis (1989); Wahba (1990); Wing (1991); Engl *et al.* (1996); Hansen (1997). This selection is not complete, further references are given later on. Parts of this thesis have been published earlier, the corresponding references will be given where appropriate.

2

Parameter estimation and the associated mean square error

2.1 Introduction

The problem of gravity field determination at the Earth's surface using data acquired at satellite height is known to be an inverse problem which is ill-posed (e.g. Rummel *et al.*, 1979). The classical example of such an inverse problem is the integral equation of the first kind, which gives the linear relation between a continuous function representing the observations and a continuous function representing the unknowns via a compact operator. It is well known that such inverse problems are unstable because the operator is compact. The whole idea becomes transparent in the frequency domain, i.e. using singular value decomposition, and it can easily be shown why the least-squares (l.s.) method fails.

Since l.s. fails, alternative solution methods are considered in this thesis. Several so-called regularisation methods are discussed here and the associated mean square error matrix is derived as well. This matrix describes the accuracy of the estimated parameters and includes propagated observation noise and bias. The latter is due to the fact that the regularisation methods yield biased estimates. The regularisation methods studied here turn out to be filtered least-squares solutions and the filter may be tuned by adjusting one or more regularisation parameters. Several parameter choice rules to tune the filters are discussed as well as their relation with the mean square error.

Although the measurements are inherently discrete and finite and although only a finite number of unknowns can be solved for in practice, the characteristics of ill-posed problems will be discussed by means of (non discrete) functions and operators defined on infinite dimensional spaces. The idea is that the original inverse problem can be described by an integral equation of the first kind. Solving for more and more unknowns then leads to a discrete inverse problem which more and more resembles the original problem (e.g. Engl *et al.*, 1996).

In the current chapter only the mean square error of the solution will be discussed. However, other errors could have been taken into account too. Model errors, for example, can be important. These errors are touched upon briefly in chapter 5. Another error source comes from the finite sampling of the signal. If, for example, one wants to determine the Earth's gravity field from a dedicated satellite gravity field mission with a certain mission length and sampling rate, the maximum frequency one can solve for is limited by these factors. Of course, higher frequencies will be present in the measurements (since they are also present in the signal which is measured), and the power of these higher frequencies

will be projected onto the lower frequencies which are solved for, which is called aliasing (Oppenheim *et al.*, 1983). A few first preliminary results on this matter for gradiometry are presented in Van Gelderen (1998) and Schuh (1999). Further study is certainly required but is outside the scope of this thesis.

This chapter summarises chapters 4 and 5 of Bouman (1998b). In turn, Bouman (1998b) is largely based on Kress (1989); Louis (1989); Groetsch (1993); Engl *et al.* (1996); Hansen (1997) and other references cited therein.

2.2 Inverse problems and regularisation

In this section the abstract linear model is given. The continuous model is an integral equation, whereas the discrete model is a matrix-vector equation. The determination of the unknowns from the observations can be classified as an inverse problem which is ill-posed, and which may result in a solution which has no (physical) meaning. This becomes especially clear when the linear equations are decomposed in spectral form. The singular value decomposition (SVD), therefore, is discussed and the least-squares solution is shown to be of little use. Finally, the general technique that does give meaningful solutions is outlined, as well as the conditions such a technique must fulfil.

2.2.1 Ill-posed problems

Let the sphere in $\mathbb{R}^3$ with radius R around the origin be given by

$$\Omega_R = \{x \in \mathbb{R}^3 | |x| = R\}.$$

Then

$$\|\Omega_R\| = \int_{\Omega_R} d\omega_R = 4\pi R^2$$

with $d\omega_R$ the surface element on Ω_R. The linear integral equation

$$\int_{\Omega_R} K(x,y) f(y) d\omega_R(y) = g(x), \quad x \in \Omega_r \tag{2.1}$$

is a *Fredholm integral equation of the first kind*, where the function $f \in L^2(\Omega_R)$ is unknown, Ω_r is the sphere with radius r, $r > R$, and the square integrable kernel $K \in C(\Omega_r) \times C(\Omega_R)$ and the right hand side $g \in L^2(\Omega_r)$ are given functions. A geodetic example is the downward continuation of the gravitational potential at satellite height to the Earth's surface: the potential $g(x)$ at height r is known and the potential at height R is to be determined. The integral equation (2.1) refers to a spherical assumed Earth and the orbit is assumed to be circular. Schneider (1997) considers the generalisation of the integral equation to non-spherical geometries.

Symbolically, (2.1) is written as

$$Af = g. \tag{2.2}$$

The operator $A : F \to G$ is linear and compact, and it is a single-valued mapping with domain F and the range is contained in G, that is, for every $f \in F$ the mapping A assigns unique elements $Af \in G$. (Appendix A summarises a few definitions from functional analysis). Note that for simplicity the L^2 spaces where f and g live are denoted as F and G respectively from now on.

In real life the number of measurements is finite and so is the number of unknowns. The discrete version of (2.2) is denoted as

$$Ax = y$$

with $x \in \mathbb{R}^n, y \in \mathbb{R}^m$ and $m \geq n$. Matrix A is often assumed to have full column rank, that is, $\text{rank}(A) = n$.

Observations contaminated with noise get a superscript ε:

$$g^\varepsilon = g + \epsilon, \quad \|\epsilon\|_G \leq \varepsilon$$
$$y^\varepsilon = y + \epsilon, \quad \|\epsilon\|_2 \leq \varepsilon$$

with $\varepsilon \in \mathbb{R}_0^+$. Furthermore, $E\{g^\varepsilon\} = g, E\{y^\varepsilon\} = y$ and $D\{y^\varepsilon\} = \sigma^2 P^{-1}$, with E the expectation operator, D the dispersion operator and P^{-1} the error variance-covariance matrix. Without loss of generality P is scaled such that the variance of unit weight σ^2 is set equal to one in most cases from here on.

The determination of the unknown function f from the observed function g^ε and known operator A is called an *inverse problem*. A desirable property of the estimate $\hat{f}$ is of course that it is close to f if g^ε is close to g. Moreover, a unique solution should exist for all $g \in G$. These are the conditions of continuity, existence and uniqueness. When those three conditions are met, the inverse problem is said to be well-posed.

Definition 2.1 (well-posed, ill-posed). Let $A : F \to G$ be an operator from a normed space F into a normed space G. The equation

$$Af = g \tag{2.3}$$

with $f \in F, g \in G$ is called *well-posed* if A is bijective and the inverse operator $A^{-1} : G \to F$ is continuous. Otherwise the problem is said to be *ill-posed*. □

According to this definition three types of ill-posedness can be distinguished (Kress, 1989). If A is not surjective then (2.3) is not solvable for all $g \in G$ (*non-existence*). If A is not injective then (2.3) may have more than one solution (*non-uniqueness*). Finally, if A^{-1} exists but is not continuous then the solution f of eq. (2.3) does not depend continuously on the data g (*instability*).

Often at least one of the conditions is not satisfied in inverse problems, therefore inverse problems are often ill-posed. From theorems A.3 and A.4 (appendix A) we know, for example, that a compact operator cannot have a bounded inverse. The inverse problem is unstable, that is, small changes in the data g result in large changes in the solution f. A typical example from satellite geodesy is the downward continuation of the satellite data to the Earth's surface. At satellite height, higher and higher frequencies are damped more and more, and will eventually be overwhelmed by the measurement noise. The inverse problem, the downward continuation, causes noise amplification, resulting in an unrealistic solution. This will be elaborated in more detail later on.

The existence of the solution will not be a matter of great concern in this thesis. Naturally, it is an important requirement that a solution exists for exact data, but for perturbed data the problem has to be changed (regularised) and the notion of a solution can be relaxed, that is, the existence of an *approximate* solution is required. As side remark it is noted that even in the presence of exact data no exact solution may exist since every model contains simplifications and approximations.

The amount of data that is available for the determination of the solution f is usually finite. Because f is a continuous function the approximate solution is never unique in practice. Also when continuous exact data is available the null space of A may not be zero. However, injectivity is assumed unless stated otherwise, and the non-uniqueness due to the finite amount of data will not be discussed, see for example Backus and Gilbert (1967, 1968); Parker (1994); Trampert and Snieder (1996). Schreiner (1994) studies uniqueness in a gravity gradiometric context.

Spectral decomposition

If the determination of f from noise corrupted data g^ε is a well-posed problem, then minimising

$$\|Af - g^\varepsilon\|_G^2$$

yields the (stable) least-squares solution

$$\hat{f} = (A^*A)^{-1}A^*g^\varepsilon,$$

which is denoted as A^+g^ε. This is the unique best-approximate solution, that is, the least-squares solution of minimal norm. The set of all least-squares solutions is $\hat{f} + N(A)$ (e.g. Engl *et al.*, 1996). However, the problem investigated in the framework of this dissertation is ill-posed and the least-squares solution is not stable. This becomes particularly clear when considering the SVD of the compact operator A (2.2):

$$Af = \sum_{n=1}^{\infty} \sigma_n \langle v_n, f\rangle u_n \tag{2.4}$$

where u_n and v_n are the eigenvectors of AA^* and A^*A respectively, and σ_n are the singular values, cf. appendix A. In finite dimensions (2.4) is

$$Ax = U\Sigma V^T x \tag{2.5}$$

with U and V orthogonal matrices and the singular values on the diagonal of Σ. An important property of the singular values σ_n is that they accumulate at 0, compare appendix A:

$$\lim_{n\to\infty} \sigma_n = 0.$$

The inverse operations of (2.4) and (2.5) are

$$\hat{f} = A^+g^\varepsilon = \sum_{n=1}^{\infty} \frac{\langle u_n, g^\varepsilon\rangle}{\sigma_n} v_n \tag{2.6}$$

and

$$\hat{x} = A^+y^\varepsilon = V\Sigma^{-1}U^Ty^\varepsilon$$

respectively. Equation (2.6) shows that the inverse is unstable. Errors in g^ε corresponding to high frequencies, that is large n, are amplified by large factors $1/\sigma_n$. If dim $R(A) < \infty$ the amplification stays bounded, but might be unacceptably large. However, if dim $R(A) = \infty$, then $\sigma_n \to 0$ as $n \to \infty$, so data errors of a fixed size are amplified without bounds.

This also suffices to show that the discretisation of the original problem $Af = g$ leads to regularisation, that is, the problem is no longer ill-posed. The original equation is approximately solved by a projection method, that is, $Af = g$ is replaced by the discrete counterpart $Ax = y$. Let $F_n \subset F$, $G_n \subset G$ and let $P_n : G \to G_n$ be projection operators. For given $g \in R(A)$ the projection method approximates the solution $f \in F$ of $Af = g$ by the solution $f_n \in F_n$ of the projected equation

$$P_nAf_n = P_ng \quad (\text{ or } Ax = y)$$

compare Kress (1989). The condition number of the discrete linear system will grow with the dimension n of the subspace used for the projection method. Increasing n will make the discretisation error smaller but errors in the data will be amplified more. The discrete system has a finite number of singular values and the error stays bounded. However, the error might be unacceptably large, and for increasing n the problem becomes more and more ill-posed.

Regularisation schemes

Since the generalised inverse A^+ does not give stable solutions, the idea is to replace $\hat{f} = A^+g^\varepsilon$ by a continuous approximate solution

$$\hat{f}_\alpha = A_\alpha^+ g^\varepsilon$$

such that $\hat{f}_\alpha \to \hat{f}$, $\alpha \to 0$. Therefore one has:

Definition 2.2 (regularisation scheme). Let F and G be normed spaces and let $A : F \to G$ be an injective bounded linear operator. A *regularisation scheme* consists of a family of bounded linear operators $A_\alpha^+ : G \to F$, $\alpha > 0$, with the property of point-wise convergence

$$\lim_{\alpha \to 0} A_\alpha^+ A f = f \tag{2.7}$$

for all $f \in F$. The positive parameter α is called the *regularisation parameter*. □

Property (2.7) is equivalent to $A_\alpha^+ g \to A^+ g$, $\alpha \to 0$, for all $g \in R(A)$ (see Kress, 1989).

One would like that the regularised solution converges to the exact solution when the error level goes to zero:

Definition 2.3 (regular). The choice of the regularisation parameter $\alpha = \alpha(\varepsilon)$ depending on the error level ε is called *regular* if for all $g \in R(A)$ and all $g^\varepsilon \in G$ with $\|g^\varepsilon - g\|_G \leq \varepsilon$ it holds

$$\lim_{\varepsilon \to 0} A_{\alpha(\varepsilon)}^+ g^\varepsilon = A^+ g.$$

Thus $\alpha(\varepsilon) \to 0$, $\varepsilon \to 0$. □

A different perspective on regularisation schemes comes from the SVD of the inverse operator A^+. The ill-posedness of an integral equation of the first kind with compact operator stems from the behaviour of the singular values, $\sigma_n \to 0$, $n \to \infty$. An obvious idea, therefore, is to filter out the influence of the factor $1/\sigma_n$ and to restore continuity. To this end, consider the filter $\delta : (0, \infty) \times (0, \|A\|) \to \mathbb{R}$ which is defined as a bounded function satisfying the conditions:

1. For each $\alpha > 0$ there exists a positive constant $c(\alpha)$ such that

$$|\delta(\alpha, \sigma)| \leq c(\alpha)\sigma \tag{2.8}$$

   for all $0 < \sigma \leq \|A\|$.

2. It holds

$$\lim_{\alpha \to 0} \delta(\alpha, \sigma) = 1 \tag{2.9}$$

   for all $0 < \sigma \leq \|A\|$.

Then the operator $A_\alpha^+ : G \to F, \alpha > 0$, defined by

$$A_\alpha^+ g := \sum_{n=1}^{\infty} \frac{\delta(\alpha, \sigma_n)}{\sigma_n} \langle g, u_n \rangle v_n$$

for all $g \in G$, describes a regularisation scheme with

$$\|A_\alpha^+\| \leq c(\alpha).$$

Thus, A_α^+ is a bounded linear operator with bound c (Kress, 1989). It is not allowed to use any arbitrary filter since conditions (2.8) and (2.9) have to be satisfied. Tikhonov regularisation, biased estimation and SVD methods will be discussed, and the corresponding filter functions δ will be derived.

These methods are all 'global' methods, that is, the regularisation acts on the solution as a whole. In appendix B three alternatives are briefly discussed that try to adapt the regularisation such that it only works for specific parameters or in specific areas. These 'local' methods are spherical wavelets, the Konopliv-Sjogren method, and additional constraints.

Not discussed in this thesis are the so-called iteration methods like Landweber iteration or conjugate gradients. The former is not very useful due to its slow convergence, while the latter has extensively been studied by Schuh (1996). Moreover, the latter is a non-linear method, obscuring error propagation and, therefore, quality description.

Mean square error matrix

The difference between the solution f from the error free data g and the regularised solution $\hat{f}_\alpha$ from the erroneous data g^ε is

$$\hat{f}_\alpha - f = A_\alpha^+(g^\varepsilon - g) + (A_\alpha^+ - A^+)g.$$

The first term on the right hand side represents the influence of the *data error*, the second is due to the approximation error between A_α^+ and A^+, i.e. the *regularisation error* or *bias*.

A proper quality description of the regularised solution $\hat{f}_\alpha$ takes both errors into account. The mean square error matrix (MSEM) is the error variance-covariance matrix describing the error in the solution. Let the vector of estimated parameters be

$$\hat{x}_\alpha = A_\alpha^+ y^\varepsilon$$

with y^ε the discrete observations. The propagated data error then is

$$Q_x = \left(A_\alpha^+\right) P^{-1} \left(A_\alpha^+\right)^T$$

with $D\{y^\varepsilon\} = P^{-1}$. The MSEM is the sum of the propagated error and the bias term

$$MSEM := Q_x + \Delta A x x^T \Delta A^T \tag{2.10}$$

with $\Delta A := (A_\alpha^+ - A^+)A$.

2.2.2 Global regularisation methods

Global regularisation methods, as opposed to local regularisation methods, act on the set of unknown parameters as a whole. (A few local regularisation methods are briefly discussed in appendix B.) Three kinds of global regularisation methods are discussed: Tikhonov regularisation, biased estimation and SVD methods. The specific form of the filters are given as well as the explicit expressions for A_α^+. Note that all regularisation methods involve the choice of one or more regularisation parameters, denoted as α, α_i or k. Parameter choice rules are discussed in section 2.3.

Tikhonov regularisation

The idea of Tikhonov regularisation (TR) is to minimise the quadratic functional

$$J_\alpha(f) := \|Af - g^\varepsilon\|_G^2 + \alpha\|f\|_F^2.$$

The norm of the solution, therefore, has to be bounded: $\|f\|_F \leq c < \infty$. The positive regularisation parameter α is a compromise between data misfit (the first term) and the power of the solution (the second term). In finite dimensions, minimising

$$J_\alpha(x) := \|Ax - y^\varepsilon\|_P^2 + \alpha\|x\|_K^2 \tag{2.11}$$

yields

$$\hat{x}_\alpha = (A^T P A + \alpha K)^{-1} A^T P y^\varepsilon,$$

which means that $A_\alpha^+ = (A^T P A + \alpha K)^{-1} A^T P$. If $P = I$ and $K = I$, the minimisation problem (2.11) is said to be in *standard form*. The transformation to the standard form

$$J_\alpha(x) := \|Ax - y^\varepsilon\|_2^2 + \alpha\|x\|_2^2$$

is always possible (Eldén, 1982), and the solution is

$$\hat{x}_\alpha = (A^T A + \alpha I)^{-1} A^T y^\varepsilon = \sum_{i=1}^{n} \delta(\alpha, \sigma_i) \frac{\langle y^\varepsilon, u_i \rangle}{\sigma_i} v_i$$

with the filter

$$\delta(\alpha, \sigma_i) = \frac{\sigma_i^2}{\sigma_i^2 + \alpha}.$$

More generally, one can minimise

$$J_{\alpha,p}(f) := \|Af - g^\varepsilon\|_G^2 + \alpha \|f\|_p^2 \tag{2.12}$$

with

$$\|f\|_p^2 = \sum_{i=0}^{p} \|c_i f^{(i)}\|_F^2, \quad p \in \mathbb{N}$$

where $f^{(i)}$ is the i-th derivative of f and $c_i \in C^i[a,b]$ are given positive functions (Groetsch, 1984). This is called higher order Tikhonov regularisation and $C^i[a,b]$ denotes the space of all i-times continuously differentiable functions on $[a,b]$. The constraint of the derivatives is in the H^p-norm, and the space $H^p[a,b]$ is the *Sobolev space of order p.*

A disadvantage of general TR may be that (2.12) always contains a term which tends to minimise the mean of the approximate solution, which may be undesirable, for example, if the unknown function does not have zero mean. Instead, one could also minimise

$$J_{\alpha,L}(f) := \|Af - g^\varepsilon\|_G^2 + \alpha \|Lf\|_F^2, \quad f \in D(L) \tag{2.13}$$

with L a differential operator. It differs from TR in that the regularisation term is a semi norm rather than a norm. If $N(A) \cap N(L) = \{0\}$ then the minimiser $\hat{f}_\alpha$ of (2.13) is unique and satisfies

$$A^* A \hat{f}_\alpha + \alpha L^* L \hat{f}_\alpha = A^* g^\varepsilon.$$

In finite dimensional space, A usually has full rank, hence $N(A) = \{0\}$ and therefore $N(A) \cap N(L) = \{0\}$.

Using the generalised SVD (GSVD) of (A, L), compare appendix A, the solution is

$$\hat{x}_\alpha = \sum_{i=1}^{p} \delta(\alpha, \gamma_i) \frac{\langle y^\varepsilon, u_i \rangle}{\sigma_i} x_i + \sum_{i=p+1}^{n} \langle y^\varepsilon, u_i \rangle x_i$$

with the generalised singular values γ_i, the generalised singular vectors x_i, and filter

$$\delta(\alpha, \gamma_i) = \frac{\gamma_i^2}{\gamma_i^2 + \alpha}, \quad i = 1, \ldots, p \tag{2.14}$$

compare Hansen (1990); Bouman (1998b).

Biased estimation

The idea of biased estimation (BE) is to add a positive-definite matrix to the system of normal equations. The matrix is chosen such that the total error, i.e. bias and noise, is minimal.

The unstable least-squares solution of

$$\min_x J(x) := \min_x \|Ax - y^\varepsilon\|_2^2$$

is $\hat{x} = (A^TA)^{-1}A^Ty^\varepsilon$. One obtains a stable solution with

$$\hat{x}_\alpha = (A^TA + \alpha I)^{-1}A^Ty^\varepsilon \tag{2.15}$$

where $\alpha \geq 0$. The method (2.15) is called *biased estimation* or *ridge regression*. Of course, eq. (2.15) is the solution of minimising (2.11) in standard form. Consequently, TR and BE lead to the same type of equations. The bias is denoted as

$$\Delta x := E\{\hat{x}_\alpha - x\} = -(A^TA + \alpha I)^{-1}\alpha x$$

where $-(A^TA + \alpha I)^{-1}\alpha I = \Delta A$.

A generalisation of biased estimation is called *generalised biased estimation* (GBE), and it is derived as follows. Let the SVD of A be $U\Sigma V^T$. The l.s. solution then is

$$\begin{aligned} \hat{x} &= (V\Sigma^2V^T)^{-1}V\Sigma U^Ty^\varepsilon \\ &= V(\Sigma^2)^{-1}\Sigma U^Ty^\varepsilon. \end{aligned}$$

The GBE solution is defined as

$$\hat{x}_g = V(\Sigma^2 + \Delta)^{-1}\Sigma U^Ty^\varepsilon, \tag{2.16}$$

where Δ is a diagonal matrix with positive elements $\alpha_1, \ldots, \alpha_n$ to be chosen appropriately. The corresponding filter is

$$\delta(\alpha_i, \sigma_i) = \frac{\sigma_i^2}{\sigma_i^2 + \alpha_i}.$$

In non-spectral form (2.16) is

$$\hat{x}_g = (A^TA + M)^{-1}A^Ty^\varepsilon$$

where $M = V\Delta V^T$, which in general is a full and positive-definite matrix. This is different from BE since M is a scaled unit matrix there. The GBE solution minimises

$$\|Ax - y^\varepsilon\|_2^2 + \|Lx\|_2^2, \quad M = L^TL.$$

However, L can no longer be identified as a differential operator. The elements of Δ are not arbitrary, but will be chosen such that the trace of the MSEM is minimal, cf. section 2.3. The operator $A_\alpha^+ = (A^TPA + M)^{-1}A^TP$, including data weights.

Methods based on SVD

The most simple form of the SVD methods is the truncated SVD (TSVD). The smallest singular values are neglected and the TSVD solution is

$$\hat{x}_k = \sum_{i=1}^{n} \delta(k)\frac{\langle y^\varepsilon, u_i\rangle}{\sigma_i}v_i \tag{2.17}$$

with filter $\delta(k)$ defined as

$$\text{TSVD} \quad \delta(k) := \begin{cases} 1 & \text{for } i = 1, \ldots, k \\ 0 & \text{for } i = k+1, \ldots, n \end{cases}.$$

TSVD solves the problem

$$\min_x \|A_kx - y^\varepsilon\|_2 \text{ subject to } \min \|x\|_2,$$

with $A_k = U\Sigma_kV^T$ and $\Sigma_k = \text{diag}(\sigma_1, \ldots, \sigma_k, 0, \ldots, 0)$. Moreover, $A_k^+ = (A_k^TA_k)^{-1}A_k^T$.

A generalisation of TSVD is to solve

$$\min_x \|A_k x - y^\varepsilon\|_2 \text{ subject to } \min \|Lx\|_2.$$

This leads to the truncated generalised singular value decomposition (TGSVD). The solution is given by (2.17) with filter function

$$\text{TGSVD} \quad \delta(k,\sigma_i) := \begin{cases} 0 & \text{for } i = 1,\ldots,p-k \\ 1 & \text{for } i = p-k+1,\ldots,p \\ \sigma_i & \text{for } i = p+1,\ldots,n \end{cases}$$

where the σ_i are the values from the GSVD of A, compare appendix A.

Depending on the severeness of the ill-posedness of a problem, one might want to introduce less filtering. Moreover, the TSVD filter is an ideal low-pass filter, that is, the filter factors are either zero or one leaving all low frequent parts unchanged, which may lead to oscillations around the true solution in the spatial domain. The filter factors zero and one in TSVD and TGSVD could be replaced by more smooth filter factors to obtain damped SVD (DSVD) and damped GSVD (DGSVD). They are given as

$$\text{DSVD} \quad \delta(\alpha,\sigma_i) = \frac{\sigma_i}{\sigma_i + \sqrt{\alpha}}$$

and

$$\text{DGSVD} \quad \delta(\alpha,\gamma_i) = \frac{\gamma_i}{\gamma_i + \sqrt{\alpha}}$$

respectively, with γ_i the generalised singular values (Hansen, 1997). These filter factors decay slower than the Tikhonov filter factors and thus introduce less filtering. Schneider (1997) considers smoothed TSVD with filter

$$\delta(k) := \begin{cases} 1 & \text{for } i = 1,\ldots,p-k \\ \tau(i) & \text{for } i = p-k+1,\ldots,p \\ 0 & \text{for } i = p+1,\ldots,n \end{cases}$$

with $\tau(i)$ a strict monotonously decreasing function which satisfies $\tau(p-k+1) = 1$ and $\tau(p) = 0$.

2.3 Choice of regularisation parameters

All regularisation methods involve one or more regularisation parameter(s) to be determined. Several methods to choose a single parameter are discussed in this section as well as one method to determine multiple parameters in case of generalised biased estimation. The relation of the different parameter choice rules with the (minimum) mean square error (MSE) is treated as well.

2.3.1 Minimum MSE

The trace of the MSEM is called the *mean square error* and it is the expected squared difference between $\hat{x}_\alpha$ and x (Hoerl and Kennard, 1970):

$$\begin{aligned} MSE &:= E\{\|\hat{x}_\alpha - x\|_2^2\} = \text{tr}(MSEM) \\ &= \text{tr}(Q_x) + \text{tr}(\Delta x \Delta x^T) = \text{tr}(Q_x) + \Delta x^T \Delta x \end{aligned}$$

with the bias $\Delta x := E\{\hat{x}_\alpha - x\}$. In table 2.1 the MSE is given for the regularisation methods in standard form. They have been derived using the SVD, compare Bouman (1998b). Here σ^2 is the variance of unit weight and λ_i are the eigenvalues of $A^T A$ or $\sigma_i^2 = \lambda_i$, with σ_i the singular values of A.

For the moment let's concentrate on one regularisation parameter α, see section 2.3.3 for multiple α_i. The smallest squared distance between $\hat{x}_\alpha$ and x equals the minimum MSE and, since x is 'known'

Table 2.1: *Propagated error, bias term and mean square error of several regularisation methods.*

Method	$\mathrm{tr}(Q_x)$	$\Delta x^T \Delta x$	MSE
TR/BE	$\sum_{i=1}^{n} \frac{\sigma^2 \lambda_i}{(\lambda_i+\alpha)^2}$	$\sum_{i=1}^{n} \frac{\alpha^2 \langle x, v_i\rangle^2}{(\lambda_i+\alpha)^2}$	$\sum_{i=1}^{n} \frac{\sigma^2 \lambda_i + \alpha^2 \langle x, v_i\rangle^2}{(\lambda_i+\alpha)^2}$
GBE	$\sum_{i=1}^{n} \frac{\sigma^2 \lambda_i}{(\lambda_i+\alpha_i)^2}$	$\sum_{i=1}^{n} \frac{\alpha_i^2 \langle x, v_i\rangle^2}{(\lambda_i+\alpha_i)^2}$	$\sum_{i=1}^{n} \frac{\sigma^2 \lambda_i + \alpha_i^2 \langle x, v_i\rangle^2}{(\lambda_i+\alpha_i)^2}$
TSVD	$\sum_{i=1}^{k} \frac{\sigma^2}{\lambda_i}$	$\sum_{i=k+1}^{n} \langle x, v_i\rangle^2$	$\sum_{i=1}^{n} \left[\frac{\sigma^2 \delta_i}{\lambda_i} + (1-\delta_i)\langle x, v_i\rangle^2 \right]$
DSVD	$\sum_{i=1}^{n} \frac{\sigma^2}{(\sqrt{\lambda_i}+\sqrt{\alpha})^2}$	$\sum_{i=1}^{n} \frac{\alpha \langle x, v_i\rangle^2}{(\sqrt{\lambda_i}+\sqrt{\alpha})^2}$	$\sum_{i=1}^{n} \frac{\sigma^2 + \alpha\langle x, v_i\rangle^2}{(\sqrt{\lambda_i}+\sqrt{\alpha})^2}$

in a simulation, the α or k which really yields the minimum MSE may really be computed. For all the errors given here it holds that tr(Q_x) is a monotonic decreasing function of α (or monotonic decreasing function of $1/k$), while tr($\Delta x \Delta x^T$) is a monotonic increasing function of α (or monotonic increasing function of $1/k$). Hoerl and Kennard (1970) show for BE that the sum of these two functions, the MSE, always has a minimum smaller than that of the least-squares estimate, compare figure 2.1.

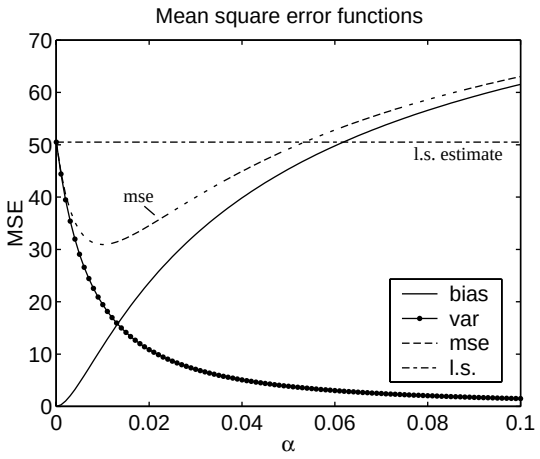

Figure 2.1: *Example of the mean square error, the eigenvalues λ_i are $1/i$, $n = 100$, $\sigma^2 = 10^{-2}$ and $\langle x, v_i\rangle = 1$. The l.s. value is equal to the value of the total* MSE *for $\alpha = 0$ which is tr$(A^T P A)^{-1}$.*

Of course, in a real life situation one does not know x. Xu (1998) therefore suggests to use $\hat{x}$ or an initial $\hat{x}_\alpha$ instead and then to minimise for this approximation of x. The problem, however, can be ill-posed such that one cannot determine $\hat{x}$ due to numerical instability. In that case, $\hat{x}$ cannot be used. Also, using $\hat{x}_\alpha$, α may be too large yielding too smooth solutions or α may be too small which will lead to solutions oscillating around the mean with large amplitude. In the former case, the bias is underestimated which results in a too large α since the emphasis will be on minimising tr(Q_x). In the latter case, if the initial α is too small, $\|\hat{x}_\alpha\|$ tends to be unrealistically large and the emphasis of minimising the MSE will be on minimising the bias: the 'optimal' α will be too small. In practice, however, it may turn out that several iterations with new initial $\hat{x}_\alpha$ selected 'randomly' within a certain bound, lead to an unambiguous determination of α (which hopefully is close to the optimal α). This would be a kind of Monte Carlo simulation which is, however, outside the scope of this thesis.

2.3.2 Single regularisation parameter

The methods discussed to determine a single regularisation parameter are

- quasi-solutions (Ivanov, 1962),
- discrepancy principle (Morozov, 1984),
- L-curve (Hansen, 1992),
- generalised cross validation (GCV) (Wahba, 1990),
- quasi-optimality (Morozov, 1984).

These methods can be divided into two groups, the a posteriori methods and the heuristic methods. The first two parameter choice rules belong to the first group and the last three choice rules to the second. It can be shown for the a posteriori methods that α goes to zero as ε goes to zero, that is, the parameter choice is called regular (definition 2.3). This is formally not the case for the heuristic methods. In practice, however, these methods may work well (cf. Engl *et al.*, 1996; Engl, 1997).

The parameter choice rules are given here with emphasis on Tikhonov regularisation. The application of these rules to other regularisation methods is straightforward. For more details refer to Bouman (1998b).

A posteriori methods

Quasi-solutions. Given a perturbed g^ε of $g \in G$, with $\|g - g^\varepsilon\|_G \leq \varepsilon$, choose α such that

$$(\alpha I + A^*A)\hat{f}_\alpha = A^*g^\varepsilon \tag{2.18}$$

satisfies $\|\hat{f}_\alpha\|_F = c$, where c is an a priori bound on the norm of the exact solution (Kress, 1989).

Numerically the regularisation parameter can be obtained by solving

$$Z(\alpha) = \|\hat{f}_\alpha\|_F^2 - c^2 = 0$$

with *Newton iteration*

$$\alpha_{n+1} = \alpha_n - \frac{Z(\alpha_n)}{Z'(\alpha_n)}, \quad n = 0, 1, \ldots$$

(see Press *et al.*, 1992). The derivative of Z is given by

$$Z'(\alpha) = 2\langle \frac{d\hat{f}_\alpha}{d\alpha}, \hat{f}_\alpha \rangle$$

since $\|\hat{f}_\alpha\|^2 = \langle \hat{f}_\alpha, \hat{f}_\alpha \rangle$, and

$$\frac{d\hat{f}_\alpha}{d\alpha} = -(A^*A + \alpha I)^{-1}\hat{f}_\alpha \tag{2.19}$$

as can be derived from (2.18).

Kress (1989) derives a starting value for the iteration to find the desired α. Suppose that $\|\hat{f}_\alpha\|_F = c$. Since $\hat{f}_\alpha$ satisfies $\alpha\hat{f}_\alpha + A^*A\hat{f}_\alpha = A^*g^\varepsilon$, it holds that

$$\begin{aligned} \alpha\|\hat{f}_\alpha\| &= \|A^*g^\varepsilon - A^*A\hat{f}_\alpha\|_F \\ \alpha c &\leq \|A\|\|g^\varepsilon - A\hat{f}_\alpha\|_G \\ &\leq \|A\|\|g^\varepsilon - AA^+g\|_G \leq \|A\|\varepsilon. \end{aligned}$$

Therefore, provided that $\|A^+g\|_F \leq c$, one has the estimate

$$\alpha c \leq \|A\|\varepsilon$$

In practice rescaling of c to assure convergence may be necessary (e.g. Bouman, 1998b).

Discrepancy principle. Given a perturbed g^ε of $g \in G$ with a known error level $\|g^\varepsilon - g\|_G \leq \varepsilon < \|g^\varepsilon\|_G$, choose α such that $\|A\hat{f}_\alpha - g^\varepsilon\|_G = \varepsilon$.

The regularisation parameter can be obtained by solving

$$Z(\alpha) = \|A\hat{f}_\alpha - g^\varepsilon\|_G^2 - \varepsilon^2 = 0$$

with Newton's method. Rewriting the above norm using (2.18), we obtain

$$\begin{aligned} \|g^\varepsilon - A\hat{f}_\alpha\|_G^2 &= \langle g^\varepsilon - A\hat{f}_\alpha, g^\varepsilon - A\hat{f}_\alpha\rangle \\ &= \langle g^\varepsilon - A\hat{f}_\alpha, g^\varepsilon\rangle - \langle A^*(g^\varepsilon - A\hat{f}_\alpha), \hat{f}_\alpha\rangle \\ &= \|g^\varepsilon\|_G^2 - \langle \hat{f}_\alpha, A^*g^\varepsilon\rangle - \alpha\|\hat{f}_\alpha\|_F^2. \end{aligned}$$

Thus

$$Z(\alpha) = \|g^\varepsilon\|_G^2 - \langle \hat{f}_\alpha, A^*g^\varepsilon\rangle - \alpha\|\hat{f}_\alpha\|_F^2 - \varepsilon^2$$

and

$$Z'(\alpha) = -\langle \frac{d\hat{f}_\alpha}{d\alpha}, A^*g^\varepsilon\rangle - \|\hat{f}_\alpha\|_F^2 - 2\alpha\langle \frac{d\hat{f}_\alpha}{d\alpha}, \hat{f}_\alpha\rangle$$

where the derivative $d\hat{f}_\alpha/d\alpha$ is given by (2.19).

Provided that $\|g^\varepsilon\|_G > \varepsilon$ (SNR > 1), one has the estimate

$$\alpha(\|g^\varepsilon\|_G - \varepsilon) \leq \|A\|^2\varepsilon$$

which may serve as starting value for the iteration (until $\|A\hat{f}_\alpha - g^\varepsilon\|_G = \varepsilon$), see (Kress, 1989; Groetsch, 1984).

The discrepancy principle is widely used. Louis (1989), for instance, exclusively applies the discrepancy principle as parameter choice rule. It has to be mentioned that often the criterion $\|A\hat{f}_\alpha - g^\varepsilon\|_G = R\varepsilon$, with $R > 1$ and fixed, is used.

Schwintzer (1990) developed an algorithm to determine α, in the framework of collocation, very similar to the discrepancy principle. Let the least-squares collocation estimate be

$$\begin{aligned} \hat{x}_c &= (\sigma^{-2}A^TQ_y^{-1}A + C_{xx}^{-1})^{-1}\sigma^{-2}A^TQ_y^{-1}y^\varepsilon \\ &= (A^TQ_y^{-1}A + \sigma^2C_{xx}^{-1})^{-1}A^TQ_y^{-1}y^\varepsilon \end{aligned}$$

where $\sigma^{-2}Q_y^{-1}$ is the weight matrix of the observations y^ε with σ^2 the variance of unit weight and C_{xx} is the signal covariance matrix of x. Determining α, means scaling of σ^2 here. Schwintzer's idea is as follows. In a least-squares context it holds that

$$E\{\hat{e}^TP\hat{e}\} = m - n \tag{2.20}$$

with $\hat{e} = y^\varepsilon - A\hat{x}$ the vector of residuals, $\hat{x}$ the least-squares estimate, $P = \sigma^{-2}Q_y^{-1}$, m the number of observations and n the number of unknowns, $m > n$. Generally, (2.20) does not hold when instead of $\hat{e}$ the residuals $\hat{e}_c = y^\varepsilon - A\hat{x}_c$ are used. The correct variance of unit weight σ^2 is assumed to be that σ^2 for which

$$(\hat{e}_c)^T P\,(\hat{e}_c) = m - n$$

or

$$\|A\hat{x}_c - y^\varepsilon\|_P^2 = m - n$$

is true. It is, therefore, a discrepancy principle-like method.

The method of quasi-solutions and the discrepancy principle can be summarised as follows (Kress, 1989):

- For given $c > 0$ minimise the defect $\|Af - g\|_G$ subject to the constraint that the norm is bounded by $\|f\|_F \leq Rc$.
- For given $\varepsilon > 0$ minimise the norm $\|f\|_F$ subject to the constraint that the defect is bounded by $\|Af - g\|_G \leq R\varepsilon$.

Heuristic methods

A disadvantage of the above two methods is the necessity of a priori bounds on either the signal or the measurement error. Dealing with gravity field determination of the Earth, a few signal models exist, e.g. Kaula (1966) or Tscherning and Rapp (1974). However, these are approximations and the power of the models differs from one model to another (e.g. Rapp, 1972; Jekeli, 1978; Rapp, 1979). Consequently, for quasi-solutions the regularisation parameter may be too large or too small, resulting in a too smooth or too rough solution.

The information on the noise level may also be unreliable. Typically, the worst-case bound will be a severe overestimation, while the a priori standard deviation might underestimate the true error (Engl *et al.*, 1996). Therefore, it is necessary to consider alternative a posteriori parameter choice rules that avoid knowledge of the noise level or the signal energy, and to determine a regularisation parameter on the basis of the actual performance of the regularisation method. Examples are the L-curve, generalised cross validation (GCV) and the quasi-optimality criterion. Strictly speaking these heuristic parameter choice rules cannot provide a convergent regularisation method because they are not regular in the sense of definition 2.3, (Engl *et al.*, 1996), but in practice they may work well.

L-curve. Let $\hat{x}_\alpha$ be the solution of minimising

$$\|Ax - y^\varepsilon\|_P^2 + \alpha\|Lx\|_K^2.$$

Then the L-curve is a plot of the (semi)norm

$$\eta(\alpha) = {}^{10}\log \|L\hat{x}_\alpha\|_K$$

of the regularised solution versus the corresponding residual norm

$$\xi(\alpha) = {}^{10}\log \|A\hat{x}_\alpha - y^\varepsilon\|_P$$

as a function of the regularisation parameter. For discrete ill-posed problems it turns out that this curve, when plotted in *log-log* scale, has an L-shaped appearance with a distinct corner separating the vertical and horizontal parts of the curve (Hansen, 1997), see figure 2.2. Originally the use of the L-curve was suggested by Lawson and Hanson (1974). Ilk (1986, 1993) discusses a few choices of the regularisation parameter very similar to the L-curve in the context of unbiased estimators.

The L-curve behaviour can be explained by considering the two error components, that is the data error ϵ and the regularisation error Δx. The vertical part of the L-curve corresponds to solutions where $\|L\hat{x}_\alpha\|_K$ is very sensitive to changes in the regularisation parameter because the data error ϵ dominates $\hat{x}_\alpha$ and because ϵ does not satisfy the discrete Picard condition (Hansen, 1997), see also appendix A. Stated otherwise, the vertical part corresponds to smaller α. The emphasis of minimising $J(\alpha)$ is on $\|A\hat{x}_\alpha - y^\varepsilon\|_P$, allowing $\|L\hat{x}_\alpha\|_K$ to become large. The horizontal part of the L-curve corresponds to solutions where the residual norm $\|A\hat{x}_\alpha - y^\varepsilon\|_P$ is most sensitive to the regularisation parameter because $\hat{x}_\alpha$ is dominated by the regularisation error Δx, as long as y satisfies the discrete Picard condition (ibid).

The corner of the L-curve can be found by considering the point $C = (\xi(\alpha_c), \eta(\alpha_c))$ where the L-curve is concave and the tangent at C has slope -1. The concave condition is necessary, because the slope may also be -1 near the endpoints of the curve, compare figure 2.2. It turns out that point C is a corner of the L-curve if and only if the function

$$\psi(\alpha) = \|L\hat{x}_\alpha\|_K \|A\hat{x}_\alpha - y^\varepsilon\|_P$$

has a local minimum at $\alpha = \alpha_c$ (Regińska, 1996; Engl *et al.*, 1996).

Although the L-curve method seems to work well in a number of applications, it still lacks a sound mathematical foundation, see (Engl *et al.*, 1996, section 4.5) and (Vogel, 1996).

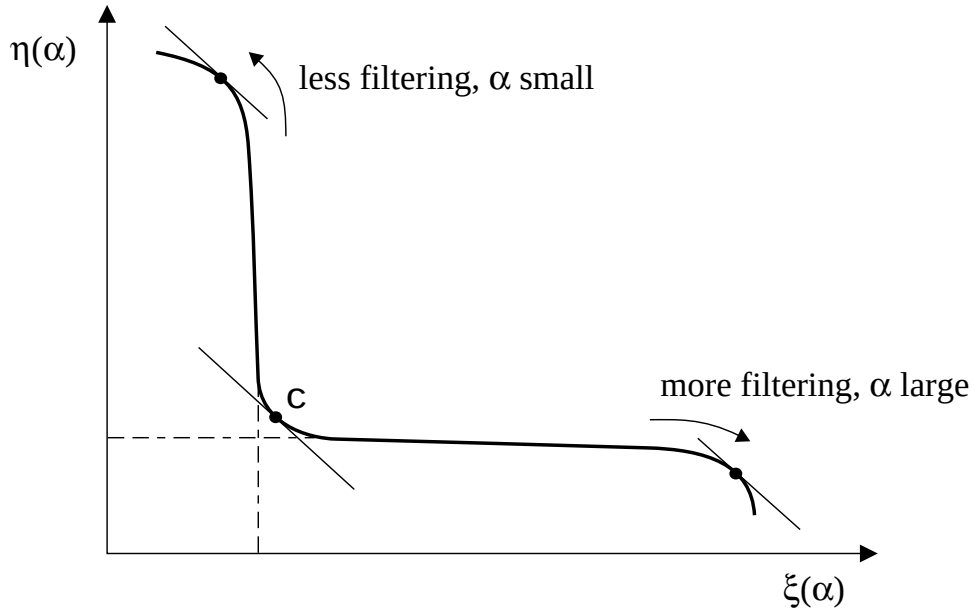

Figure 2.2: *The L-curve in log-log scale (from Hansen (1997)).*

Generalised cross validation. The idea of GCV is that if an arbitrary element y_i^ε of y^ε is left out, then the corresponding regularised solution should predict this observation well. Moreover, let $\tilde{y}^\varepsilon := Qy^\varepsilon$, with Q an orthogonal matrix. The problem of estimating x from

$$\tilde{y}^\varepsilon = \tilde{A}x + \tilde{\epsilon}$$

where $\tilde{A} = QA$, $\tilde{\epsilon} = Q\epsilon$, is the same problem of estimating x from

$$y^\varepsilon = Ax + \epsilon$$

if the errors are normally distributed. Therefore, the choice of the regularisation parameter should be independent of an orthogonal transformation of y^ε (Wahba, 1990). This leads to the minimisation of:

$$J(\alpha) := \frac{\|A\hat{x}_\alpha - y^\varepsilon\|_P^2}{(\mathrm{tr}(I_m - AA_\alpha^+))^2}. \tag{2.21}$$

The denominator can be expressed in terms of filter factors:

$$\mathrm{tr}(I_m - AA_\alpha^+) = m - (n-p) - \sum_{i=1}^{p} \delta(\alpha, \gamma_i)$$

with the filter δ defined in (2.14), see (Hansen, 1997). If $\alpha = 0$ the trace in (2.21) equals $m - n$, compare (2.14).

The range, $R(A)$, has finite dimension, since the foundation of generalised cross-validation originates from statistical considerations and depends on the assumption that the data perturbation is finite-dimensional white noise (Engl *et al.*, 1996):

$$E\{y - y^\varepsilon\} = 0 \quad \text{and} \quad E\{(y - y^\varepsilon)(y - y^\varepsilon)^T\} = \sigma^2 I.$$

This implies that $E\{\|y - y^\varepsilon\|_P^2\} = m\sigma^2$, hence $\varepsilon = \sqrt{m}\sigma$. The restriction to finite dimensions of $R(A)$, therefore, is a must. The assumption of *white* noise is indeed essential as (Hansen and O'Leary, 1993) show. In case of coloured noise no minimum is found with the GCV method whereas the L-curve works well.

Quasi-optimality. The third heuristic parameter choice rule discussed is the quasi-optimality method (Morozov, 1984; Engl *et al.*, 1996; Hansen, 1997). This rule also tries to compromise between the data error and the regularisation error by minimising the change in the regularised solution with respect to α.

Let $\hat{f}_{\alpha,j}$ be the solution of the problem of minimising

$$J_\alpha(f, f_{\alpha,j-1}) := \|Af - g^\varepsilon\|_G^2 + \alpha\|f - f_{\alpha,j-1}\|_F^2, \quad j \in \mathbb{N} \tag{2.22}$$

with $f_{\alpha,0} = 0$. This is called iterated Tikhonov regularisation. If $j \to \infty$ then $\hat{f}_{\alpha,j} \to \hat{f}$, that is, it converges to the unstable least-squares solution (Engl *et al.*, 1996), see also section 4.2. The smaller α is, the faster iterated Tikhonov regularisation converges to the unstable l.s. solution and therefore $\hat{f}_{\alpha,j+1}$ will be a worse approximation of f, the true solution, than $\hat{f}_{\alpha,j}$. The solution $\hat{f}_{\alpha,j+1}$ is, so to say, less regularised than the solution $\hat{f}_{\alpha,j}$ and the data error ε dominates. The smaller α gets, the larger the difference $\|\hat{f}_{\alpha,j+1} - \hat{f}_{\alpha,j}\|$ will be. For large values of α, however, the bias dominates $\hat{f}_{\alpha,j} - f$ and $\hat{f}_{\alpha,j+1}$ will be a better approximation of f, and the absolute difference $\|\hat{f}_{\alpha,j+1} - \hat{f}_{\alpha,j}\|$ will decrease as α becomes smaller (Engl *et al.*, 1996). Altogether, $\|\hat{f}_{\alpha,j+1} - \hat{f}_{\alpha,j}\|$ considered as a function of α will in general decrease as long as α is small, and increase for larger values of α. Intuitively, it seems reasonable to choose a value $j = j_0$ for which

$$\|\hat{f}_{\alpha,j+1} - \hat{f}_{\alpha,j}\|_F$$

is minimised, which should correspond to balancing the data error and the regularisation error. The value $\alpha = \alpha(j)$ is called the *quasi-optimal value* of the regularisation parameter.

Initial value of α

For the a posteriori parameter choice rules initial values were already given. For the heuristic methods one could use the following. Let the minimisation problem be

$$\min \|Ax - y\|_2^2 + \alpha\|Lx\|_2^2.$$

Press *et al.* (1992) suggest to firstly use

$$\alpha = \frac{\text{tr}(A^TA)}{\text{tr}(L^TL)}$$

which tends to make the two parts of the minimisation having comparable weights.

Relation between certain parameter choice rules and the mean square error

It is demonstrated in for example (Golub *et al.*, 1979; Wahba, 1990) that the GCV criterion is expected to give a regularisation parameter that results in a MSE close to the minimum MSE. Wahba (1990) remarks that the discrepancy principle does not give a minimum MSE and is likely to give too smooth solutions. The relation of the L-curve with the (minimum) MSE is not well understood, although the L-curve tends to give too smooth solutions (Hansen, 1999). The relation between the corner of the L-curve and the MSE can be described as follows. The horizontal and vertical part correspond to a large change in data error and regularisation error, respectively. The corner of the L-curve is defined as the point where the rate of change for both errors is equal. Quasi-optimality is a similar method, its relation with the MSE has not been studied so far.

Discussion on the use of the regularisation parameters

The a posteriori regularisation parameters will not be used here. The problem of determining the regularisation parameter shifts to the problem of determining the right scaling factor R for the total noise or total signal (cf. Bouman, 1998b). The L-curve and quasi-optimality method may give stable solutions

but, as stated, their relation with the MSE is not well understood. Also the application of the GCV method is cumbersome, since coloured noise is to be studied in this thesis whereas white noise is required. Furthermore, the computation of the trace in (2.21) requires a decomposition of the design matrix A (cf. Engl *et al.*, 1996, section 9.4), which is too laborious in our case.

It is, therefore, decided to rely on the minimisation of the MSE. Because the MSE is the expected squared distance between the true and computed solution, the minimum MSE is adopted as the lower bound of the accuracy. No matter what parameter choice rule is chosen, the result should preferably be close to the regularisation parameter associated with the minimum MSE. Since simulation will be used, the true solution is known and the MSE can be computed.

2.3.3 Multiple regularisation parameters

The GBE solution involves the determination of multiple regularisation parameters. Hoerl and Kennard (1970) show that, starting from the l.s. solution, one can iterate towards a set of α_i's with minimum MSE. Later, Hemmerle (1975) found an explicit expression for the set of optimal regularisation parameters with respect to the least-squares solution.

Minimum MSE. The set of α_i with minimum MSE is obtained by differentiating the MSE with respect to α_i (see for example (Xu and Rummel, 1994a) and table 2.1):

$$\frac{\partial MSE}{\partial \alpha_i} = \frac{2\lambda_i(\alpha_i\langle x, v_i\rangle^2 - \sigma^2)}{(\lambda_i + \alpha_i)^3}$$

where σ^2 is the a priori variance of unit weight. The minimum is obtained for $\alpha_i = \sigma^2/\langle x, v_i\rangle^2$. The second order derivative of the MSE with respect to α_i is

$$\frac{\partial^2 MSE}{\partial \alpha_i^2} = 2\frac{\lambda_i^2\langle x, v_i\rangle^2 - 2\lambda_i\alpha_i\langle x, v_i\rangle^2 + 3\lambda_i\sigma^2}{(\lambda_i + \alpha_i)^4}.$$

Inserting $\alpha_i = \sigma^2/\langle x, v_i\rangle^2$ yields

$$\frac{\partial^2 MSE}{\partial \alpha_i^2} = 2\frac{\lambda_i^2\langle x, v_i\rangle^2 + \lambda_i\sigma^2}{(\lambda_i + \sigma^2/\langle x, v_i\rangle^2)^4} > 0$$

because $\lambda_i > 0$ (they are the eigenvalues of a positive definite matrix). Thus a minimum is found and the minimum MSE becomes

$$\min(MSE) = \sum_{i=1}^{n} \frac{\sigma^2}{\lambda_i + \sigma^2/\langle x, v_i\rangle^2}.$$

The above equation is not very useful for practical purposes since the true solution x is unknown. Having gravity field determination in mind one could for example use approximate coefficients from an existing gravity model such as OSU91A (Rapp *et al.*, 1991), instead of the true coefficients x. Computing α_i with this approximate x and inserting these values in (2.16) gives updated x and α_i's until the change in the α_i's is considered to be small enough. Note that if $\langle x, v_i\rangle^2$ is small, a small change in x may lead to a large change in α_i. In that case, convergence may be a problem.

Hoerl and Kennard (1970) suggest to use the least-squares solution as initial value for the iteration:

$$\alpha_{i,0} = \frac{\hat{\sigma}^2}{\langle \hat{x}, v_i\rangle^2}$$

where the a posteriori variance of unit weight $\hat{\sigma}^2$ and $\hat{x}$ are the least-squares values. However, it may occur in practice that because of numerical instability it is impossible to compute the least-squares solution.

Then, one can start the iteration from any (stable) BE solution. The first part of the MSE is $\mathrm{tr}(Q_{\hat{x}})$ which is a continuous, monotonically decreasing function of $\alpha_1, \alpha_2, \ldots, \alpha_n$ (table 2.1), whereas the second part, $\|\Delta x^T\|_2^2$, is continuous, monotonically increasing as function of α_i (Xu and Rummel, 1994a). The MSE therefore has a unique minimum as function of α.

2.4 Summary

The determination of the Earth's gravity field by satellite methods may be characterised as an ill-posed inverse problem. In practical circumstances, that is, in a finite dimensional setting, this means that the l.s. solution yields an ill-conditioned system of equations. The ill-conditioning is such that small data errors may lead to unacceptably large solution errors.

A number of regularisation methods exists to overcome this ill-conditioning. All discussed methods are filtered least-squares solutions and the filter can be tuned by one or more regularisation parameters. A price one has to pay, however, is that the solution becomes biased and the mean square error matrix, which replaces the usual a posteriori error variance-covariance matrix, consists of bias and propagated noise.

A posteriori methods as well as heuristic methods to determine the regularisation parameter have been discussed. For conceptual reasons the a posteriori methods will not be used. Also the heuristic methods will not be applied due to the computational restrictions. Instead, the mean square error is minimised. The regularisation parameter is chosen such that the trace of the MSEM is minimal. This will allow the comparison of different regularisation methods in chapter 5.

3

Quality measures

3.1 Introduction

In this thesis, the gravitational potential at the Earth's surface is the unknown function to be determined. Satellite observations are used for this purpose here, and due to the downward continuation regularisation is necessary. It was shown in the preceding chapter that the regularisation methods considered in this thesis can all be written as filtered least-squares solutions. The filter differs from one method to another and so does the solution and the mean square error matrix (MSEM). However, not the potential itself is used but its expansion in a series of spherical harmonics. The spherical harmonic coefficients are the unknowns to be solved for. Therefore, the spherical harmonic expansion of the gravitational potential is treated first in section 3.2.

Using simulated data one could compare the different solution methods by looking at the difference between solved coefficients and 'true' coefficients. However, this is not possible in reality and quality measures have to be used instead. All quality measures use the MSEM whether the solution is biased or not. Concerning this bias, it might be argued that (2.15) should be interpreted as an unbiased solution. This is discussed here since it has implications for the quality description of the solution. Furthermore, a few general remarks on the bias computation are made (section 3.3).

The quality assessment of a global gravity field solution can be based on a number of tests and measures. On the one hand there are many different applications for a global gravity field model and on the other hand a better understanding of the quality is obtained considering the errors in the frequency domain (the spherical harmonic coefficients) *and* the spatial domain. In practice, the solution obtained from a dedicated gravity field mission, like GOCE, will be compared with independent data, e.g. gravimetry data, altimeter data and satellite tracking data. Since a simulation study is used here, such a comparison cannot be carried out. Instead, one has to rely on the formal error description of the estimated parameters. First of all, in section 3.4 the propagation of the errors in the potential coefficients, as described by the MSEM, to other gravity field quantities like geoid heights is discussed. Furthermore, the so-called ratio measures will be considered, such as the size of the solved parameters relative to the size of their uncertainty (signal-to-noise ratio) in section 3.5. Finally, measures for the contribution of the observations to the solution are addressed (section 3.6).

The methods used are largely based on the work done by (Haagmans and van Gelderen, 1991; Xu, 1992a,b; Bouman, 1993, 1998a).

3.2 Spherical harmonic expansion of the gravitational potential

The most common series expansion of the gravitational potential V is the one in spherical harmonics as function of geocentric polar coordinates (Heiskanen and Moritz, 1967):

$$V(r,\theta,\lambda) = \frac{GM}{R}\sum_{l=0}^{\infty}\left(\frac{R}{r}\right)^{l+1}\bar{Y}_l(\theta,\lambda) \tag{3.1}$$

with

$$\bar{Y}_l(\theta,\lambda) = \sum_{m=-l}^{l}\bar{K}_{lm}\bar{Y}_{lm}(\theta,\lambda)$$

where

$$\bar{K}_{lm} = \begin{cases} \bar{C}_{lm}, & m \geq 0 \\ \bar{S}_{l|m|}, & m < 0 \end{cases} \tag{3.2}$$

and

$$\bar{Y}_{lm}(\theta,\lambda) = \begin{cases} \bar{P}_{lm}(\cos\theta)\cos m\lambda, & m \geq 0 \\ \bar{P}_{l|m|}(\cos\theta)\sin|m|\lambda, & m < 0 \end{cases} \tag{3.3}$$

and

GM	universal gravitational constant times mass of the Earth
R	mean equatorial radius
l, m	degree, order
$\bar{K}_{lm}$	fully normalised potential coefficients
$\bar{Y}_{lm}$	fully normalised surface spherical harmonics
$\bar{P}_{lm}$	fully normalised associated Legendre functions
r, θ, λ	geocentric polar coordinates (radius, co-latitude, longitude).

Interchanging the summation over l and m leads to the Fourier series (Colombo, 1981)

$$V(r,\theta,\lambda) = \sum_{m=0}^{\infty} A_m(r,\theta)\cos m\lambda + B_m(r,\theta)\sin m\lambda \tag{3.4}$$

with coefficients

$$\left.\begin{matrix} A_m(r,\theta) \\ B_m(r,\theta) \end{matrix}\right\} = \sum_{l=m}^{\infty} H_{lm}(r,\theta)\left\{\begin{matrix} \bar{C}_{lm} \\ \bar{S}_{lm} \end{matrix}\right. \tag{3.5}$$

$$H_{lm}(r,\theta) = \frac{GM}{R}\left(\frac{R}{r}\right)^{l+1}\bar{P}_{lm}(\cos\theta).$$

Disturbing potential

Let the Earth be approximated by an ellipsoid of revolution. Introducing such a reference ellipsoid has the advantage that the deviations of the actual gravity field from the ellipsoidal normal field are so small that second order terms can often be neglected (Heiskanen and Moritz, 1967). The remaining field is called *disturbing gravity field*.

Let the reference ellipsoid be an equipotential surface of the normal gravity field, with given mass M_0, semi-axes a and b, and angular velocity ω. Then the normal potential is uniquely determined. The angular velocity of the ellipsoid, for example, is the GRS80 value (Moritz, 1980), the difference with the

angular velocity of the actual Earth are neglected. Refer to Chovitz (1988): in a very good approximation the Earth's angular velocity is $\omega(t) = \omega_0 + \dot{\omega}t$ with $\omega_0 = 7.292115 \times 10^{-5}$ rad/s (the GRS80 value) and the secular change is $\dot{\omega} = -4.6 \pm 0.4 \times 10^{-22}$ rad/s^2. The angular velocity, therefore, is accurately known.

The expansion of the normal potential in spherical harmonics is

$$U(r,\theta,\lambda) = \frac{GM_0}{b} \sum_{l=0(2)}^{\infty} \left(\frac{b}{r}\right)^{l+1} \bar{c}_{l0}\bar{P}_{l0}(\cos\theta) + \frac{1}{2}\omega^2 r^2 \sin^2\theta$$

which is a Somigliana-Pizetti reference field. The summation over l has a step size of 2. The disturbing potential T is defined as

$$T := W - U$$

where $W = V + 1/2\omega^2 r^2 \sin^2\theta$. Putting $b = R$, and taking the origin at the center of mass of the Earth ($\bar{C}_{10} = \bar{C}_{11} = \bar{S}_{11} = 0$), the expansion of T is

$$T(r,\theta,\lambda) = \frac{GM}{R} \sum_{l=2}^{\infty} \left(\frac{R}{r}\right)^{l+1} \sum_{m=-l}^{l} \Delta\bar{K}_{lm}\bar{Y}_{lm}(\theta,\lambda)$$

with

$$\Delta\bar{K}_{lm} = \begin{cases} \bar{C}_{l0} - \bar{c}_{l0}, & m = 0, \quad l \text{ even} \\ \bar{K}_{lm}, & \text{elsewhere} \end{cases}$$

where the term $G(M - M_0)/R$ has been omitted since the relative uncertainty of GM is of the order 10^{-9} (Ries *et al.*, 1992). If the estimation of spherical harmonic coefficients with respect to a reference orbit or field is mentioned, $\bar{C}_{lm}$ will be written, but it should be clear that these are the actual corrections to the reference coefficients. The reference field used here is GRS80 with even degree coefficients $\bar{q}_0$ up to degree 8 (Moritz, 1980).

3.3 Biased and unbiased estimation

In chapter 2 it was shown that a stable solution is obtained by minimising

$$\|Ax - y^\varepsilon\|_P^2 + \alpha\|x\|_K^2$$

with the semi-norm $\|x\|_K^2 = \langle x, Kx\rangle$ and solution

$$\hat{x}_\alpha = (A^T P A + \alpha K)^{-1} A^T P y^\varepsilon. \tag{3.6}$$

This solution is biased and the bias is

$$\begin{aligned} \Delta x := E\{\hat{x}_\alpha - x\} &= -(A^T P A + \alpha K)^{-1}\alpha K x \\ &= \Delta A x = (A_\alpha^+ - A^+)Ax \end{aligned}$$

see chapter 2. If $\alpha \to \infty$ then $\Delta x \to -x$ ($\hat{x}_\alpha = 0$). One obvious problem is that the true solution x is needed if one wants to estimate the bias. This issue is addressed later.

Note that (3.6) also is the *unbiased* solution of the linear model

$$E\{\begin{pmatrix} y^\varepsilon \\ z \end{pmatrix}\} = \begin{pmatrix} A \\ I \end{pmatrix} x, \quad D\{\begin{pmatrix} y^\varepsilon \\ z \end{pmatrix}\} = \begin{pmatrix} P^{-1} & 0 \\ 0 & [\alpha K]^{-1} \end{pmatrix} \tag{3.7}$$

where $z = 0$ are zero observations with variance-covariance matrix $(\alpha K)^{-1}$. Usually, K is a diagonal matrix with weights according to a signal variance model, Kaula's rule for example (Kaula, 1966). The scaling parameter α is considered as a scaling of the variance of unit weight. The error variance-covariance matrix of $\hat{x}$ is $Q_x = (A^T PA + \alpha K)^{-1}$, which is clearly different from the MSEM and Q_x in case of a biased solution, cf. eq. (2.10).

The question arises, therefore, how the solution (3.6) should be interpreted: "Is it biased or not?". It will be clear by now that the author's opinion is inclined towards biased estimation for the following reasons.

- The so-called satellite-only models are computed using (3.6). These models are derived from satellite tracking data and because of the nature of the observations (long wavelength sensitivity) in combination with the satellites altitude (typically 800 - 1500 km for the major part of the satellites) the maximum degree and order solved for is about 70. As it is generally recognised, however, the high degrees are biased towards zero, that is, have less power than one would expect (see for example Marsh *et al.*, 1988; Reigber, 1989; Nerem *et al.*, 1993). Remarkably enough, the bias is not accounted for in the subsequent quality description of the satellite-only models.

- The model $E\{z\} = Ix$, $D\{z\} = (\alpha K)^{-1}$ is not correct. Kaula's rule or any other power rule, as used to determine K, represents the estimated *signal* variance of the coefficients, not their *error* variance.

  Alternative additional observations might be coefficients from an earlier model with the corresponding error matrix. The solution then becomes

  $$\hat{x} = (A^T PA + P_0)^{-1}(A^T Py^{\varepsilon} + P_0 x^0)$$

  where x^0 contains the coefficients of a global gravity field model and P_0 is the inverse of their error matrix. However, a global gravity field model is to be determined as independent as possible from previous gravity field solutions.

For further discussion on biased and unbiased estimation see (Xu, 1992a,b) and (Xu and Rummel, 1994a).

As said before, a complication in the quality description of biased estimators is the bias assessment. The bias computation involves the true solution, which is not available of course. Instead of the true values, the bias could be estimated by using $\hat{x}_\alpha$, that is, using the biased solution itself. Xu (1992b) shows, however, that this on the average will underestimate the size of the bias, since the coefficients are biased towards zero. Fortunately, the power per degree is known approximately, as in Kaula's rule for example. Thus the size of the coefficients is known approximately. The bias can therefore be estimated by taking the *sign* of the biased coefficients and their *size* according to a certain power law (Koop, 1993). Alternatively, the bias can be estimated using existing global gravity fields model like OSU91A or EGM96 (Rapp *et al.*, 1991; Lemoine *et al.*, 1999).

A further complication is that, in general, the linear model is obtained after linearisation, see also section 4.2. Different reference models yield different linear models and the bias therefore depends on the reference model. In addition, solving the minimisation problem

$$\min_x \|Ax - y^{\varepsilon}\|_P^2 + \alpha\|x - x^0\|_K^2$$

with x^0 an a priori estimate of x yields the bias

$$E\{\hat{x}_\alpha^0 - x\} = \alpha(A^T PA + \alpha K)^{-1} K(x^0 - x)$$

where

$$\hat{x}_\alpha^0 = (A^T PA + \alpha K)^{-1}(A^T Py^{\varepsilon} + \alpha K x^0).$$

The solution $\hat{x}_\alpha^0$ is therefore biased towards the a priori x^0. If the bias is computed by using OSU91A for x and, for example, JGM-3 for x^0, then the bias will be smaller compared to a computation with $x^0 = 0$. However, a proper quality description of $\hat{x}_\alpha^0$ has to take the quality of x_0 into account. If x_0 is biased itself, for example, then $\hat{x}_\alpha^0$ is biased towards the biased x_0. It is outside the scope of this thesis to assess whether the quality description of the individual coefficients of current global gravity field models is adequate. As stated in chapter 1, one of the problems of, for example, satellite-only models is that they are biased. Moreover, gravity field models suffer from systematic errors. In this thesis, therefore, $x^0 = 0$, the preferred reference model always is GRS80, and OSU91A is the 'true' gravity field.

3.4 Error propagation

The unknowns to be estimated are the coefficients of a spherical harmonic series. Not only the error variances of the coefficients are of importance, the expected error of derived products, like geoid heights, is of interest as well. To this end, error propagation can be used. The error propagation with a full error variance-covariance matrix is followed by a description of the consequences for the error propagation when the error matrix is block diagonal.

3.4.1 Full error matrix

Let the potential V at the surface of a sphere with radius R be approximated by a truncated spherical harmonic expansion, that is, the maximum degree and order in (3.4) and (3.5) is L. Thus, the number of spherical harmonic coefficients is finite and can be determined from the observations. The variance and covariance of the unknowns is described by the (full) MSEM. For the propagation of these error covariances to error covariances of for example geoid heights, $cov(P, Q)$, the propagation law of variances is used

$$\begin{aligned} f &= Bx \\ E_f &= BE_xB^T, \end{aligned}$$

where f is a linear functional of the harmonic coefficients x and E_x is the error variance matrix of x, i.e., either $E_x = MSEM$ (biased estimator) or $E_x = Q_x$ (unbiased estimator). In particular one may write

$$f(\theta,\lambda) = \sum_{m=0}^{L} \left[\left(\sum_{l=m}^{L} \beta_l \bar{C}_{lm} \bar{P}_{lm}(\cos\theta) \right) \cos m\lambda + \left(\sum_{l=m}^{L} \beta_l \bar{S}_{lm} \bar{P}_{lm}(\cos\theta) \right) \sin m\lambda \right],$$

where the eigenvalues β_l may depend on the degree. Since in this study the error propagation is restricted to geoid heights and gravity anomalies, the eigenvalues do not depend on the order. Applying the propagation law one gets a two-dimensional Fourier expansion for the error covariances $cov(P, Q)$, (Haagmans and van Gelderen, 1991):

$$\begin{aligned} cov(P,Q) &= \sum_{m=0}^{L}\sum_{k=0}^{L} [A_{mk}\cos m\lambda_P \cos k\lambda_Q + B_{mk}\sin m\lambda_P \cos k\lambda_Q \\ &\quad + C_{mk}\cos m\lambda_P \sin k\lambda_Q + D_{mk}\sin m\lambda_P \sin k\lambda_Q] \end{aligned}$$

with

$$\begin{aligned} A_{mk} &= \sum_{l=m}^{L}\sum_{n=k}^{L} \beta_l\beta_n cov(\bar{C}_{lm},\bar{C}_{nk})\bar{P}_{lm}(\cos\theta_P)\bar{P}_{nk}(\cos\theta_Q) \\ B_{mk} &= \sum_{l=m}^{L}\sum_{n=k}^{L} \beta_l\beta_n cov(\bar{S}_{lm},\bar{C}_{nk})\bar{P}_{lm}(\cos\theta_P)\bar{P}_{nk}(\cos\theta_Q) \end{aligned}$$

$$C_{mk} = \sum_{l=m}^{L}\sum_{n=k}^{L} \beta_l\beta_n cov(\bar{C}_{lm},\bar{S}_{nk})\bar{P}_{lm}(\cos\theta_P)\bar{P}_{nk}(\cos\theta_Q)$$
$$D_{mk} = \sum_{l=m}^{L}\sum_{n=k}^{L} \beta_l\beta_n cov(\bar{S}_{lm},\bar{S}_{nk})\bar{P}_{lm}(\cos\theta_P)\bar{P}_{nk}(\cos\theta_Q)$$

and A_{mk}, B_{mk}, C_{mk} and D_{mk} are the Fourier coefficients of the Fourier expansion of the error covariances. $cov(\bar{C}_{lm},\bar{C}_{nk})$ etc. are the spherical harmonic coefficient error covariances. In the sequel only point error variances are considered, i.e. the case $P = Q$ is considered.

3.4.2 Block-diagonal error matrix

In general the error matrix E_x is full and positive definite. Sometimes, however, it has a special structure such as diagonal or block diagonal. As we will see this has consequences for the spatial pattern of the propagated error. Here only the block-diagonal case is discussed, see (Bouman, 1993) for the diagonal case.

A block-diagonal error matrix implies that $B_{mk} = C_{mk} = 0$ and $m = k$. The propagated error, which now is denoted as $var(\theta,\lambda)$ since it is a function of one point only, then is

$$var(\theta,\lambda) = \sum_{m=0}^{L}\left[A_m\cos^2 m\lambda + D_m\sin^2 m\lambda\right] \tag{3.8}$$

with

$$A_m = \sum_{l=m}^{L}\sum_{n=m}^{L} \beta_l\beta_n var(\bar{C}_{lm},\bar{C}_{nm})\bar{P}_{lm}(\cos\theta)\bar{P}_{nm}(\cos\theta)$$
$$D_m = \sum_{l=m}^{L}\sum_{n=m}^{L} \beta_l\beta_n var(\bar{S}_{lm},\bar{S}_{nm})\bar{P}_{lm}(\cos\theta)\bar{P}_{nm}(\cos\theta).$$

If $A_m = D_m$ for $m = 1,\ldots,L$ then (3.8) becomes

$$var(\theta,\lambda) = \sum_{m=0}^{L} A_m$$

that is, the propagated error is independent of longitude.

Bouman and Koop (1998c) showed that north-south symmetry occurs when even and odd degrees are separated, i.e., the error covariance is zero when $|l-n|$ is odd. The property

$$P_{lm}(-t) = (-1)^{l+m}P_{lm}(t)$$

yields the following four cases:

1. m is even, l,n are even; $P_{lm}(-t) = P_{lm}(t)$ and $P_{nm}(-t) = P_{nm}(t)$,
2. m is even, l,n are odd; $P_{lm}(-t) = -P_{lm}(t)$ and $P_{nm}(-t) = -P_{nm}(t)$,
3. m is odd, l,n are even; $P_{lm}(-t) = -P_{lm}(t)$ and $P_{nm}(-t) = -P_{nm}(t)$,
4. m is odd, l,n are odd; $P_{lm}(-t) = P_{lm}(t)$ and $P_{nm}(-t) = P_{nm}(t)$.

Because l and n have the same parity, the Legendre functions for a specific m are always simultaneously symmetric or anti-symmetric with respect to the equator. The combination of two of these functions, as in A_m and D_m, is therefore always north-south symmetric: $var(\theta,\lambda) = var(\pi-\theta,\lambda)$.

3.5 Ratio measures

Ratio measures for the quality of the potential coefficients give the relative size of the signal with respect to the bias, for example. Three such measures are discussed in this section.

Signal-to-noise ratio. The signal-to-noise ratio (SNR) is a measure for the significance of a coefficient. The size of the estimated coefficient with respect to its uncertainty indicates how well the coefficient is resolved. The SNR is defined as

$$SNR_{lm} := \frac{|\bar{K}_{lm}|}{\sigma_{lm}}$$

with $\bar{K}_{lm}$ the estimated coefficient and σ_{lm} the error RMS. Ideally the SNR is as large as possible for each coefficient. If the SNR is one, then there is more signal than noise with a probability of approximately 68% given a Gaussian distribution of the errors. The SNR will not be shown itself, instead a logarithmic scale is used. The operation ${}^{10}\log(SNR)$ gives the number of significant digits. If, for example, the SNR is one, then the number of significant digits is zero.

Bias-to-signal ratio. In addition to the SNR, the bias-to-signal ratio (BSR) is of importance as well. It is defined as

$$BSR_{lm} := \frac{|\Delta\bar{K}_{lm}|}{|\bar{K}_{lm}|}$$

with $\Delta\bar{K}_{lm}$ the bias in each coefficient. It shows how severe the bias in each coefficient is.

Bias-to-noise ratio. The bias-to-noise ratio (BNR) is a measure for the significance of the bias with respect to the pure noise part of the error. The diagonal elements of the corresponding matrices in the MSEM are compared:

$$BNR_{lm} := \frac{[\Delta x \Delta x^T]_{lm}}{[Q_x]_{lm}}.$$

The smaller the BNR is, the less important is the bias. One could compute the BNR for each coefficient as above, but one could also consider the ratio of the traces of the bias and pure noise matrices

$$BNR := \frac{\mathrm{tr}(\Delta x \Delta x^T)}{\mathrm{tr}(Q_x)},$$

which is a measure for the total power of the bias with respect to the total power of the propagated error.

3.6 Contribution measures

The third and final quality measure discussed is the *contribution* of the observations to the solution. The idea is that, if model (3.7) is correct, the estimators $\hat{x}$ are partially determined by the observations $\hat{y}$ and partially by the a priori information z. Schwintzer (1990) advocates the application of the redundancy number to assess the contribution of the observations $\hat{y}$ to the solution of each individual coefficient x_i. This is discussed here. The contribution measure based on (3.7) is only valid for unbiased estimators. Bouman (1998a), therefore, developed an equivalent measure for biased estimators, which is discussed as well.

In this section contribution measures are derived for Tikhonov regularisation or biased estimation. For the other regularisation methods, it is rather straightforward to derive the contribution measure once the MSEM for each method is given.

3.6.1 Contribution measure for the unbiased solution

Schwintzer (1990) uses the redundancy number as a measure for the contribution of the observations to the solution of the unknowns. The redundancy number, therefore, is discussed first, followed by Schwintzer's interpretation. Finally, the relation of this measure with the gain matrix in recursive least-squares estimation is clarified.

Redundancy number

Consider the linear relationship $E\{y^\varepsilon\} = Ax$, where the number of observations is m and the number of unknowns is n, $m \geq n$. The redundancy is $r := m - n$ and it can be shown that (Teunissen, 1994) $E\{\hat{\epsilon}^T Q_y^{-1} \hat{\epsilon}\} = \text{tr}(Q_{\hat{\epsilon}} Q_y^{-1})$, and because $E\{\hat{\epsilon}^T Q_y^{-1} \hat{\epsilon}\} = m - n$, it holds that $\text{tr}(Q_{\hat{\epsilon}} Q_y^{-1}) = m - n = r$, with Q_y the error covariance matrix of y^ε, $Q_{\hat{\epsilon}}$ the covariance matrix of $\hat{\epsilon}$ and $\hat{\epsilon}$ the vector minimising $\epsilon^T Q_y^{-1} \epsilon$, $\epsilon = y^\varepsilon - Ax$. The least-squares solution of x is $\hat{x}$, and $\hat{y}^\varepsilon = A\hat{x},\ \hat{\epsilon} = y^\varepsilon - \hat{y}^\varepsilon$.

The elements on the diagonal of $Q_{\hat{\epsilon}} Q_y^{-1}$ are denoted as $r_i : [Q_{\hat{\epsilon}} Q_y^{-1}]_{ii} = r_i$. The sum of all r_i is

$$\sum_{i=1}^{m} r_i = r,$$

r_i is the i-th local redundancy number. It measures to what extent the observation y_i^ε contributes to the total redundancy. Because $Q_{\hat{\epsilon}} = Q_y - Q_{\hat{y}}$ (Teunissen, 1994) one may write

$$\begin{aligned} r_i &= [(Q_y - Q_{\hat{y}}) Q_y^{-1}]_{ii} = [I - Q_{\hat{y}} Q_y^{-1}]_{ii} \\ &= 1 - \frac{\sigma_{\hat{y}_i}^2}{\sigma_{y_i}^2} \end{aligned}$$

if Q_y is diagonal. Therefore, $0 \leq r_i \leq 1$ since $0 \leq \sigma_{\hat{y}_i}^2 \leq \sigma_{y_i}^2$.

Under the assumption that Q_y is diagonal (the observations are uncorrelated), the local redundancy number can be associated with internal reliability, which is a measure of the model error that can be detected with a certain probability γ_0, for example, $\gamma_0 = 80\%$. The minimal detectable bias (MDB) of an observation y_i^ε is (Teunissen, 1995)

$$|\nabla_i| = \sigma_{y_i} \sqrt{\frac{\lambda_0}{r_i}}$$

with the non centrality parameter λ_0 which depends on the choice of γ_0. The MDB tells us that an error of size $|\nabla_i|$ can be detected with a probability of γ_0, the power of the test. Therefore, the smaller r_i, the larger an error in that observation must be in order to be detectable with probability γ_0.

Schwintzer's interpretation

Consider the extended model (3.7). The redundancy number of the zero observation z_i is

$$r_{z_i} = 1 - \frac{\sigma_{\hat{z}_i}^2}{\sigma_{z_i}^2} \tag{3.9}$$

where $\sigma_{\hat{z}_i}^2$ and $\sigma_{z_i}^2$ are the i-th diagonal elements of $Q_{\hat{z}}$ and Q_z respectively, Q_z (that is K) is assumed to be diagonal. Here one has $Q_z = (\alpha K)^{-1}$ and $Q_{\hat{z}} = Q_x$ since $\hat{z} = \hat{x}_\alpha$. Schwintzer (1990) uses the local redundancy number (3.9) as a measure for the contribution to the solution. He states: "The partial redundancy r_{z_i}, (...), reflects the contribution of the a priori information to the corresponding results for $\bar{C}_{lm}$ or $\bar{S}_{lm}$ in relation to the contribution coming from the real data." (Schwintzer, 1990, page 3).

There is a certain truth in this. Specifically when the zero observations are uncorrelated and $E\{z\} = Ix$, any redundancy of an observation z_i, that is any verifiability of z_i, has to come from the observations

y^ε. If $r_{z_i} = 1$, then, because of the uncorrelated zero observations, the redundant part of z_i has to come from the 'real' observations y^ε and these observations contribute 100% to the verification of z_i. However, $E\{z_i\} = x_i$ and therefore one could say that y^ε contributes 100% to the solution of x_i. On the other hand, if $r_{z_i} = 0$ then the corresponding zero observation has poor reliability and y does not contribute to the solution of $z_i = x_i$.

Note that if the a priori information consists of, for example, correlated coefficients of an earlier solution, it is not obvious how to explain the redundancy number.

Relation with the gain matrix

The gain matrix in recursive least-squares estimation describes the contribution of one or more additional observations to the estimate. The gain matrix is (Teunissen, 1996)

$$K_k := (Q^{-1}_{\hat{x}_{k-1}} + A_k^T Q_k^{-1} A_k)^{-1} A_k^T Q_k^{-1} \tag{3.10}$$

where $Q_{\hat{x}_{k-1}}$ is the error covariance matrix of the least-squares estimate $\hat{x}_{k-1}$ based on the observations y^ε_{k-1}, Q_k^{-1} the weight matrix of the observable y^ε_k and A_k the corresponding design matrix. The gain matrix describes by how much the previous estimate $\hat{x}_{k-1}$ changes to form $\hat{x}_k$. For example, if the observable y^ε_k has a relatively low precision, K_k is 'small' and $\hat{x}_{k-1}$ will not change a lot.

Translated to our problem, it is $y^\varepsilon_k = z$, $y^\varepsilon_{k-1} = y^\varepsilon$, $A_k = I$, $Q_k^{-1} = \alpha K$ and $Q_{\hat{x}_{k-1}} = A^T P A$. Inserting this in (3.10) one obtains

$$(A^T P A + \alpha K)^{-1} \alpha K =: W_z.$$

The i-th diagonal element of W_z is the relative weight of the a priori observation z_i contributing to the solution of the unknowns, and its contribution measure is

$$W_{z_i} = \frac{\sigma^2_{\hat{z}_i}}{\sigma^2_{z_i}}$$

since αK is a diagonal matrix. The relative weight of the real observations y^ε contributing to the solution of the unknowns is defined as

$$W_{y_i} := 1 - W_{z_i} = 1 - \frac{\sigma^2_{\hat{z}_i}}{\sigma_{z_i}}$$

since the total contribution to the solution of an unknown has to be 1. Thus, if the error covariance matrix of the a priori observations is diagonal, W_{y_i} equals the local redundancy number r_{z_i}. Finally, W_y can be defined as

$$W_y := I - W_z = (A^T P A + \alpha K)^{-1} A^T P A. \tag{3.11}$$

This comparison makes sense. The larger the weight of the observations is relatively to the prior information, the larger W_y gets. Conversely, the smaller the weight matrix P with respect to αK, the smaller W_y gets.

3.6.2 Contribution measure for the biased solution

The above derivations are all based on the assumption of unbiasedness of the estimator. However, the solution might be biased and the precision of the solution can no longer be described by the propagated observation error alone, the bias has to be included as well. The contribution (3.11) compares the error covariance matrix of the regularised solution with the error covariance matrix of the least-squares solution. Therefore, Bouman (1998a) suggested to introduce a similar comparison for biased estimators.

Instead of the error covariance matrix, the MSEM should be used. The contribution measure for y now is, compare with (3.11),

$$\begin{aligned} W_y &:= MSEM \cdot MSEM^{-1}\big|_{\alpha=0} \\ &= (A^TPA + \alpha K)^{-1}(A^TPA + \alpha^2 Kxx^TK)(A^TPA + \alpha K)^{-1}A^TPA. \end{aligned} \tag{3.12}$$

If $\alpha = 0$ then $W_y = I$ as required. However, if $\alpha \to \infty$ the contribution should go to zero but it does not:

$$\begin{aligned} \lim_{\alpha\to\infty} W_y &= \lim_{\alpha\to\infty} (\frac{1}{\alpha}A^TPA + K)^{-1}(\frac{1}{\alpha^2}A^TPA + Kxx^TK)(\frac{1}{\alpha}A^TPA + K)^{-1}A^TPA \\ &= K^{-1}Kxx^TKK^{-1}A^TPA \neq 0. \end{aligned}$$

Because for $\alpha \to \infty$ the bias term remains, this is not really surprising. The contribution measure (3.12), therefore, is not satisfactory and will not be used.

An alternative to the MSEM is the spectral decomposition of the MSE (Bouman, 1993):

$$\mathrm{tr}(MSEM) = MSE = \sum_{i=1}^{n} \frac{\sigma_i^2 + \alpha^2\langle x, v_i\rangle^2}{(\sigma_i^2 + \alpha)^2}$$

with the singular value decomposition of $A = U\Sigma V^T$ and v_i is a column vector of V and σ_i is the i-th diagonal element of Σ. For a single i one obtains

$$W_{y_i} = \frac{MSE_i}{MSE_i|_{\alpha=0}} = \frac{\sigma_i^2 + \alpha^2\langle x, v_i\rangle^2}{(\sigma_i^2 + \alpha)^2}\sigma_i^2. \tag{3.13}$$

If $\sigma_i^2 \gg \alpha$ it means that the unknown x_i is represented well by the measurements and $W_{y_i} \approx 1$. On the other hand, if σ_i^2 is small and $\sigma_i^2 \ll \alpha$, $W_{y_i} \approx 0$, as required. However, also in this case the behaviour for $\alpha \to \infty$ is not satisfactory:

$$\lim_{\alpha\to\infty} \frac{\frac{\sigma_i^2}{\alpha^2} + \langle x, v_i\rangle^2}{\frac{\sigma_i^4}{\alpha^2} + 2\frac{\sigma_i^2}{\alpha} + 1}\sigma_i^2 = \langle x, v_i\rangle^2\sigma_i^2 \neq 0.$$

The contribution measure (3.13) will therefore not be applied as well.

3.7 Summary

The coefficients of the expansion of the gravitational potential in a series of spherical harmonics are the unknowns to be solved for. The estimated coefficients are biased and OSU91A will be used to compute the bias with respect to the reference model GRS80.

Three quality measures have been studied, which together should provide a clear picture of the quality of a global gravity field model: error propagation, ratio measures and contribution measures. The latter two reveal a part of the quality of the gravitational potential coefficients themselves, while error propagation is a measure for the quality of derived products such as geoid heights and gravity anomalies.

A problem with the contribution measures for biased estimators is their behaviour for large α. If the regularisation parameter is large the contribution measure should go to zero, which, however, is not true. The contribution measure for unbiased estimators does show the required properties. Although the solution methods yield biased solutions, the latter contribution measure will be applied in order to validate its usefulness for quality assessment.

4

Gravity field observations

4.1 Introduction

Three types of gravity field observation methods are of importance in this study: high-low satellite-to-satellite tracking (SST), airborne gravimetry and satellite gravity gradiometry (SGG). The latter is the most important one since, in principle, a high resolution gravity field with high precision can be obtained with these observations (Schrama, 1991; Koop, 1993; Rummel *et al.*, 1993; Visser *et al.*, 1994). The GOCE mission will serve as an example (ESA, 1999). The unknowns solved for are the coefficients of a spherical harmonic series, compare chapter 3.

Because the SGG measurements will be complemented by SST this technique is discussed shortly as well. Furthermore, airborne gravimetry is considered. The GOCE satellite will fly in a non-polar orbit yielding two polar caps without observations (see the GOCE introduction below). Hence, it is anticipated that augmenting the satellite observations with gravimetry in the polar areas gives better solutions. Before discussing the observations, however, it is explained why an iterative solution method and a so-called block-diagonal normal matrix are adopted.

The material in this chapter is mainly based on the work of Colombo (1981); Schrama (1989, 1990); Koop (1993) and Rummel *et al.* (1993).

The GOCE mission

The gravity field and steady-state ocean circulation explorer mission (GOCE for short) has been selected by ESA as the first of two Earth explorer core missions and is scheduled for launch in 2004. The foreseen lifetime of the mission is 20 months with two times six months a measurement phase, the sampling period is one second (ESA, 1999). The orbit of the satellite is almost circular at a height of approximately 250 km. Due to constraints on the satellite (power supply and disturbances due to temperature fluctuations), the orbit will most likely be a sun-synchronous dawn-dusk orbit with an inclination of about 96.6°. To eliminate the effect of the non-conservative forces as much as possible the satellite motion is drag compensated.

As far as gravity field analysis is concerned, the gravity gradients are measured in three orthogonal directions x, y, z with z radial and y cross-track and x completing the right-handed coordinate system. The triad $\{x, y, z\}$ defines an orthonormal, Earth-pointing coordinate system, and it is assumed here that the attitude control system will realise such a triad (cf. ESA, 1999). The gravity gradients are denoted

as V_{xx}, V_{yy} and V_{zz}. A GPS/GLONASS receiver will be mounted on board of the gradiometer satellite enabling high-low satellite-to-satellite tracking.

Ideally, the satellite is in a polar orbit. Suppose that the mission design is such that after a period of, for example, six months the ground track pattern repeats itself. Then one would have a global data set if the data flow is uninterrupted. The inclination of 96.6°, however, yields two polar regions without observations and these are called the *polar gaps*.

4.2 Observation model and iterative solution

In general the relation between the observations and the unknowns x is non-linear:

$$y^{\varepsilon} = A(x) + \epsilon,$$

with $y^{\varepsilon} = y + \epsilon$, the observations containing noise, which is denoted as ϵ, and A is a non-linear operator. The observations y^{ε} could for example be the second derivatives of the gravity potential, star tracker readouts or GPS phase observations. The unknowns x are the potential coefficients $\bar{K}_{lm}$ up to degree and order L, eq. (3.2), and other unknowns to be determined like the coordinates of the observation points.

Linearisation of the non-linear equation yields

$$\Delta y = \partial A \Delta x + \epsilon$$

with $\Delta x = x - x^0$, $\Delta y = y^{\varepsilon} - y^0 = y^{\varepsilon} - A(x^0)$, and $\partial A = \partial_x A(x^0)$ the partial derivatives of A with respect to the unknowns x evaluated in the approximate point x^0. From now on it is assumed that the only unknowns to be determined are the corrections to the reference potential coefficients. Other parameters are assumed to be known or determined at an earlier stage. In gradiometry, for example, the orbit determined with SST is precise enough to be considered known. The approximation x^0 is in this case the GRS80 reference potential, that is, $\bar{C}_{l0}, l = 0, 2, 4, 6, 8$.

Instead of a non-linear model, the model now is linear, which is denoted as

$$y^{\varepsilon} = Ax + \epsilon$$

where one should keep in mind that y and x are the corrections to the initial values and that ϵ consists of measurement errors. Second and higher order linearisation terms as well as other model errors are neglected. Matrix A is called the design matrix. Alternatively, one can write

$$E\{y^{\varepsilon}\} = Ax, \quad D\{y^{\varepsilon}\} = Q_y \tag{4.1}$$

where Q_y is the a priori error variance-covariance matrix of the measurements, $E\{.\}$ is the expectation and $D\{.\}$ is the dispersion operator. In (4.1) the number of observations is assumed to always be larger than, or equal to, the number of unknowns.

The weighted least-squares solution of (4.1) is

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

with $P = Q_y^{-1}$. The matrix $Q_x = (A^T P A)^{-1}$ is the error covariance matrix of the solved unknowns and $A^T P A$ is called the normal matrix. If the unknowns are spherical harmonic coefficients, the size of the normal matrix is typically $O(L^4)$ where L is the maximum degree of the spherical harmonic expansion. Thus, when $L = 180$ the size of $A^T P A$ is $3 \cdot 10^4 \times 3 \cdot 10^4$ which cannot be solved directly on the computers that were available for this investigation. Standard methods exist for the solution of the normal equations. An approximate inverse is used which gives an approximate least-squares solution. Successive iterations should yield convergence towards the true least-squares solution, see e.g. (Varga, 1962; Press *et al.*, 1992). Such an iterative procedure is adopted here.

Since the inverse of the normal matrix is needed in quality assessment, its approximate inverse should be close to the true inverse matrix, while the computation of the inverse should preferably be fast and easy. The design matrix is approximated such that the normal matrix will be block diagonal, $A = A_0 + \Delta A$, where A_0 is the major part of the model and ΔA a small correction matrix. For SGG, for example, A_0 is the circular orbit part, see section 4.4. The exact normal matrix then is

$$\begin{aligned} N &:= A^T P A \\ &= A_0^T P A_0 + A_0^T P \Delta A + \Delta A^T P A_0 + \Delta A^T P \Delta A \\ &= N_0 + \Delta N \end{aligned}$$

with N_0 block diagonal. The solution of

$$N\hat{x} = A^T P y^\varepsilon$$

is the exact least-squares solution. Substituting $N = N_0 + \Delta N$ leads to

$$\begin{aligned} N_0(I + N_0^{-1}\Delta N)\hat{x} &= A^T P y^\varepsilon \\ \hat{x} &= -N_0^{-1}\Delta N \hat{x} + N_0^{-1} A^T P y^\varepsilon \end{aligned}$$

which suggests the iteration (since ΔN is small)

$$\begin{aligned} \hat{x}_n &= -N_0^{-1}\Delta N \hat{x}_{n-1} + N_0^{-1} A^T P y^\varepsilon \\ &= -N_0^{-1}(N - N_0)\hat{x}_{n-1} + N_0^{-1} A^T P y^\varepsilon \\ &= \hat{x}_{n-1} + N_0^{-1}(A^T P y^\varepsilon - N\hat{x}_{n-1}) \\ &= \hat{x}_{n-1} + N_0^{-1} A^T P (y^\varepsilon - A\hat{x}_{n-1}) \end{aligned}$$

for $n = 1, 2, 3, \ldots$. Putting $\hat{x}_0 = 0$, the first iteration is

$$\hat{x}_1 = N_0^{-1} A^T P y^\varepsilon. \tag{4.2}$$

It can be shown that

$$\lim_{n\to\infty} \hat{x}_n = \hat{x}$$

if and only if the spectral radius $\rho(I - N_0^{-1}N) < 1$, compare (Varga, 1962; Press *et al.*, 1992). In fact, Klees *et al.* (1999) show that any matrix M instead of N_0 will do provided that $\rho(I - M^{-1}M_0) < 1$ with $M_0 = A_0^T P A$.

Instead of (4.2), however,

$$\hat{x}'_1 = N_0^{-1} A_0^T P y^\varepsilon$$

is used since $A_0^T P y^\varepsilon$ is more easy to compute, with the iteration

$$\hat{x}'_n = \hat{x}'_{n-1} + N_0^{-1} A_0^T P (y^\varepsilon - A\hat{x}'_{n-1}), \quad n = 1, 2, 3, \ldots.$$

Van Gelderen (1992) shows that this converges to

$$\hat{x}' = (A_0^T P A)^{-1} A_0^T P y^\varepsilon$$

which is *not* equal to $\hat{x}$. The approximate solutions $\hat{x}'_n$ are therefore biased. However, the expectation of the approximate solution $\hat{x}'$ is

$$E\{\hat{x}'\} = (A_0^T P A)^{-1} A_0^T P A x = x$$

which shows that the approximate solution is unbiased in the limit.

As we have seen in chapter 2, it is not the least-squares solution that is of interest but a regularised solution of the form

$$\hat{x}_\alpha = (A^T P A + \alpha I)^{-1} A^T P y^\varepsilon.$$

Iteration of

$$\hat{x}'_\alpha = (A_0^T P A_0 + \alpha I)^{-1} A_0^T P y^\varepsilon.$$

may reduce the model error. As mentioned in section 2.3.2, however, this is called iterated Tikhonov regularisation and if $n \to \infty$ then the unstable least-squares solution is obtained. If many iterations are needed to overcome the model error, then the solution may become unstable. As long as the number of iterations can be kept low then the instability will not show up. In practice this depends on the size of the model error and the size of the regularisation parameter.

4.3 Series expansion of the potential in orbital coordinates

In chapter 3 the series expansion of the gravitational potential in spherical harmonics as a function of geocentric polar coordinates has been given, eq. (3.1). Working with satellite data, however, it is more convenient to adopt a coordinate system related to the orbit. The motion of a satellite around the Earth can be expressed in terms of the Keplerian orbital elements a, e, I, Ω, ω and M, which respectively are the semi-major axis of the elliptical orbit, eccentricity, inclination, right ascension of the ascending node, argument of perigee and mean anomaly. The potential in orbital elements is (e.g. Kaula, 1966; Sneeuw, 1991)

$$V(I, a, e, \psi_{kmq}(t)) = \frac{GM}{R} \sum_{l=0}^{\infty} \left(\frac{R}{a}\right)^{l+1} \sum_{m=0}^{l} \sum_{k=-l(2)}^{l} \bar{F}_{lmk}(I) \sum_{q=-\infty}^{\infty} G_{lkq}(e) S(\psi_{kmq}(t)) \qquad (4.3)$$

where

$$S(\psi_{kmq}(t)) = \begin{pmatrix} \bar{C}_{lm} \\ -\bar{S}_{lm} \end{pmatrix}_{l-m:odd}^{l-m:even} \cos\psi_{kmq}(t) + \begin{pmatrix} \bar{S}_{lm} \\ \bar{C}_{lm} \end{pmatrix}_{l-m:odd}^{l-m:even} \sin\psi_{kmq}(t),$$

$$\psi_{kmq}(t) = k(\omega(t) + M(t)) + qM(t) + m(\Omega(t) - \vartheta(t)),$$

with $\bar{F}_{lmk}$ the normalised inclination functions, G_{lkq} the eccentricity functions and ϑ the Earth's argument of longitude (or Greenwich hour angle). The summation over k runs from $-l$ to l with step size of 2.

The orbit of a dedicated gravity mission will be nearly circular, that is, $e \approx 0$ and $a \approx r$. The index q can be restricted, with sufficient accuracy, to $-1 \leq q \leq 1$ (Schrama, 1989). Since the inclusion of these eccentricity functions does not influence the analysis method fundamentally, they are left out for simplicity (Koop, 1993). Therefore, only $q = 0$ is considered and (4.3) simplifies to

$$V(I, r, \psi_{km}(t)) = \frac{GM}{R} \sum_{l=0}^{\infty} \left(\frac{R}{r}\right)^{l+1} \sum_{m=0}^{l} \sum_{k=-l(2)}^{l} \bar{F}_{lmk}(I) S(\psi_{km}(t)) \qquad (4.4)$$

with

$$S(\psi_{km}(t)) = \begin{pmatrix} \bar{C}_{lm} \\ -\bar{S}_{lm} \end{pmatrix}_{l-m:odd}^{l-m:even} \cos\psi_{km}(t) + \begin{pmatrix} \bar{S}_{lm} \\ \bar{C}_{lm} \end{pmatrix}_{l-m:odd}^{l-m:even} \sin\psi_{km}(t),$$

$$\psi_{km}(t) = k\omega_o(t) + m\omega_e(t),$$

where $\omega_o = \omega + M$ and $\omega_e = \Omega - \vartheta$, 'o' refers to orbit and 'e' to Earth.

The interchange of the summation l, m, k to k, m, l, truncated at L leads to a Fourier series (Schrama, 1989)

$$V_L(I, r, \psi_{km}(t)) = \sum_{k=-L}^{L} \sum_{m=0}^{L} A_{km}(I, r) \cos \psi_{km}(t) + B_{km}(I, r) \sin \psi_{km}(t) \tag{4.5}$$

where

$$(A_{km}(I, r), B_{km}(I, r)) = \frac{GM}{R} \sum_{l=lmin(2)}^{L} \left(\frac{R}{r}\right)^{l+1} \bar{F}_{lmk}(I) \begin{pmatrix} \bar{C}_{lm}, & \bar{S}_{lm} \\ -\bar{S}_{lm}, & \bar{C}_{lm} \end{pmatrix}_{l-m:odd}^{l-m:even}$$

with $lmin = \max(|k|, m) + \delta$, $\delta = 0$ if k and $\max(|k|, m)$ have the same parity, otherwise $\delta = 1$. The A_{km} and B_{km} coefficients are the so-called *lumped coefficients*.

4.4 Satellite gravity gradiometry

The principle of satellite gravity gradiometry (SGG) is explained as well as the assumptions leading to a block-diagonal normal matrix.

4.4.1 Principle

If two proof masses in free fall around the Earth are close to each other, they encounter a slightly different attraction by the Earth due to their different positions in the non-homogeneous Earth's gravity field. Constraining the relative motion of these two proof masses by fixing them inside a spacecraft centred around the centre of mass of the spacecraft and measuring the fixing forces exerted on the proof masses means measuring the acceleration difference which, in good approximation, is proportional to the second derivative of the gravity potential at the spacecraft's centre of mass or the gravity gradient (e.g. Rummel, 1986; Koop, 1993). Several of these pairs in different orientations are sensitive to different projections of the gradient, and one speaks of *satellite gravity gradiometry*. An example of a proposed gradiometric mission is GOCE, see section 4.1.

Two types of methods can be distinguished for the analysis of SGG measurements. One is the time-wise approach, the other the space-wise approach. The latter considers the measurements as a function of (geographical) position only, while the former considers the consecutive observations as a time series (Rummel *et al.*, 1993). The time-wise approach is adopted here. This study is concerned with quality and in the time-wise approach it is easier to model for example coloured noise (instead of white noise), whereas it is not obvious how to implement coloured noise models in the space-wise approach (Rummel *et al.*, 1993). The space-wise approach is not discussed.

4.4.2 Time-wise approach

One can discriminate between time-wise approach in the time domain (TT) and the time-wise approach in the frequency domain (TF), eq. (4.4) and (4.5) respectively. The underlying assumptions leading to a block-diagonal matrix are that (see also Koop, 1993),

- the orbit is circular,
- the repeat is exact,
- the data flow is uninterrupted,
- the data period is an integer multiple of the repeat period.

Table 4.1: *'Eigenvalues' for gradiometry.*

obs	κ_{lk}^{obs}
xx	$\frac{GM}{R^3}\left(\frac{R}{r}\right)^{l+3}(-(l+1+k^2))$
yy	$\frac{GM}{R^3}\left(\frac{R}{r}\right)^{l+3}(k^2-(l+1)^2)$
zz	$\frac{GM}{R^3}\left(\frac{R}{r}\right)^{l+3}(l+1)(l+2)$

Time-wise approach in the time domain (TT)

The potential is expressed as a time series along the orbit in eq. (4.4). It is assumed that the total mission length T_r is exactly one repeat period. When the sampling interval is Δt the total number of observations is $N_p = T_r/\Delta t$. The mission period consists of the relative primes N_d (nodal days) and N_r (orbit revolutions), or $T_r = N_d 2\pi/\dot{\omega}_e = N_r 2\pi/\dot{\omega}_o$. In this approach, ω, M, Ω and ϑ are time dependent, whereas a, e and I are assumed to be not (semi-major axis, eccentricity and inclination).[1] The angular variable ψ embodies the time dependency

$$\begin{aligned}\psi_{km}(t) &= k\omega_o(t) + m\omega_e(t) \\ &= \dot{\psi}_{km} t \end{aligned} \tag{4.6}$$

neglecting the constant phase shift ψ_{km}^o, with

$$\dot{\psi}_{km} = k\dot{\omega}_o + m\dot{\omega}_e. \tag{4.7}$$

In the observation points j, eq. (4.6) is, using (4.7),

$$\begin{aligned}\psi_{km}(j) &= (kN_r + mN_d)\frac{2\pi}{T_r} j\Delta t = 2\pi(k + m\frac{N_d}{N_r})N_r\frac{j}{N_p} \\ &= 2\pi\beta_{km}N_r\frac{j}{N_p}, \quad j = 1,\ldots,N_p\end{aligned} \tag{4.8}$$

(compare Schrama, 1990; Rummel *et al.*, 1993). In order to determine the gravitational potential coefficients $\bar{K}_{lm}$ uniquely, no two different km combinations should yield the same β_{km} frequency. The β_{km} are unique if N_r is a prime number and $N_r > 2L$, apart from $\beta_{k0} = \beta_{-k0}$, see section 4.4.3.

The diagonal elements of the gravity gradient tensor, V_{xx}, V_{yy}, V_{zz} are measured. The time series is

$$V_{obs}(I, r, \psi_{km}(t)) = \sum_{l=0}^{L}\sum_{m=0}^{l}\sum_{k=-l(2)}^{l}\kappa_{lk}^{obs}\bar{F}_{lmk}(I)S(\psi_{km}(t)) \tag{4.9}$$

where $obs = \{xx, yy \text{ or } zz\}$, and κ_{lk}^{obs} as in table 4.1, compare (Koop, 1993).

[1] If a, e and I are time dependent, then the inclination functions would become time dependent for example, and the orbit is no longer circular.

Time-wise approach in the frequency domain (TF)

Since eq. (4.5) is a Fourier series, the coefficients A_{km} and B_{km} can be derived from the time series. These coefficients serve as pseudo observations from which the coefficients $\bar{C}_{lm}$ and $\bar{S}_{lm}$ can be computed. Per order m the Fourier coefficients are a linear combination in l of the potential coefficients, and therefore are called lumped coefficients.

Specifically, consider the diagonal elements of the gravity gradient tensor, V_{xx}, V_{yy}, V_{zz}. Their Fourier series are (Koop, 1993)

$$V_{obs}(I, r, \psi_{km}(t)) = \sum_{k=-L}^{L} \sum_{m=0}^{L} A_{km}^{obs}(I, r) \cos \psi_{km}(t) + B_{km}^{obs}(I, r) \sin \psi_{km}(t)$$

with

$$\begin{pmatrix} A_{km}^{obs}(I, r) \\ B_{km}^{obs}(I, r) \end{pmatrix} = \sum_{l=lmin(2)}^{L} \kappa_{lk}^{obs} \bar{F}_{lmk}(I) \begin{pmatrix} \alpha_{lm} \\ \beta_{lm} \end{pmatrix} \tag{4.10}$$

$$\alpha_{lm} = \begin{pmatrix} \bar{C}_{lm} \\ -\bar{S}_{lm} \end{pmatrix}_{l-m:odd}^{l-m:even}, \quad \beta lm = \begin{pmatrix} \bar{S}_{lm} \\ \bar{C}_{lm} \end{pmatrix}_{l-m:odd}^{l-m:even} \tag{4.11}$$

and $lmin = \max(|k|, m) + \delta, \delta = 0$ if k and $\max(|k|, m)$ have the same parity, otherwise $\delta = 1$. The κ_{lk}^{obs} are given in table 4.1.

Relation between TT and TF

Let the time series y^{ε} be related to the unknown coefficients x as

$$E\{y^{\varepsilon}\} = Ax, \quad D\{y^{\varepsilon}\} = Q_y.$$

The Fourier transform of this discrete time series gives the pseudo observations y_F^{ε}:

$$y_F^{\varepsilon} = F y^{\varepsilon}$$

and

$$Q_{y_F} = F Q_y F^T$$

where F is the Fourier matrix, $F^{-1}F = I$, compare (Strang, 1986). Let $FAx := A_F x$. Thus, the least-squares solution is

$$\begin{aligned} \hat{x}_F &= \left(A_F^T Q_{y_F}^{-1} A_F\right)^{-1} A_F^T Q_{y_F}^{-1} y_F^{\varepsilon} \\ &= \left(A^T F^T F^{-T} Q_y^{-1} F^{-1} F A\right)^{-1} A^T F^T F^{-T} Q_y^{-1} F^{-1} F y^{\varepsilon} \\ &= \hat{x}. \end{aligned}$$

The normal matrix, $N = A^T P A$, and the right-hand vector $A^T P y^{\varepsilon}$ are therefore equal in the TT and TF approach.

4.4.3 Block-diagonal normal matrix

Consider the time series $V_{obs}(j)$, $j = 1, \ldots, N_p$, eq. (4.9). The time dependent variable is $\psi_{km}(j)$ as defined in (4.8). One column $[A]_{lm}$ of the design matrix corresponds to one coefficient $\bar{K}_{lm}$ and contains all j. Specifically, four types of elements exist: $[A]_{lm} =$

1. $\sum_{k=-l(2)}^{l} \kappa_{lk}^{obs} \bar{F}_{lmk}(I) \cos\psi_{km}(j)$, for C_{even}, 2. $\sum_{k=-l(2)}^{l} \kappa_{lk}^{obs} \bar{F}_{lmk}(I) \sin\psi_{km}(j)$, for C_{odd},
3. $\sum_{k=-l(2)}^{l} \kappa_{lk}^{obs} \bar{F}_{lmk}(I) \sin\psi_{km}(j)$, for S_{even}, 4. $-\sum_{k=-l(2)}^{l} \kappa_{lk}^{obs} \bar{F}_{lmk}(I) \cos\psi_{km}(j)$, for S_{odd},

where $even$ and odd stand for $l-m : even$ and $l-m : odd$ respectively. Suppose for the moment that $A^TPA = 1/\sigma^2 A^TA$, that is, equal variance for all observations and no correlation. One element of A^TA then is the inner product of the columns of A:

$$\begin{aligned}
[A^TA]_{l_1m_1l_2m_2} &= \sum_{j=1}^{N_p} \sum_{k_1} \kappa_{l_1k_1}^{obs} \bar{F}_{l_1m_1k_1} \begin{bmatrix} \cos\psi_{k_1m_1}(j) \\ \text{or} \\ \sin\psi_{k_1m_1}(j) \end{bmatrix} \sum_{k_2} \kappa_{l_2k_2}^{obs} \bar{F}_{l_2m_2k_2} \begin{bmatrix} \cos\psi_{k_2m_2}(j) \\ \text{or} \\ \sin\psi_{k_2m_2}(j) \end{bmatrix} \\
&= \sum_{k_1=-l_1(2)}^{l_1} \kappa_{l_1k_1}^{obs} \bar{F}_{l_1m_1k_1} \sum_{k_2=-l_2(2)}^{l_2} \kappa_{l_2k_2}^{obs} \bar{F}_{l_2m_2k_2} [cc \text{ or } cs \text{ or } sc \text{ or } ss]
\end{aligned}$$

with the four cases, using (4.8),

$$cc = \sum_{j=1}^{N_p} \cos\frac{2\pi}{N_p} j(\beta_{k_1m_1} N_r) \cos\frac{2\pi}{N_p} j(\beta_{k_2m_2} N_r)$$

and for cs, sc and ss, c and s have to be replaced by $\cos$ and $\sin$ respectively.

The orthogonality properties of the trigonometric functions on $[-\pi, \pi]$ yield $cs = sc = 0$ and $cc = ss = 0$ if $|\beta_{k_1m_1}| \neq |\beta_{k_2m_2}|$. Furthermore, if $|\beta_{k_1m_1}| = |\beta_{k_2m_2}| \neq 0$ then $cc = ss = N_p/2$, and if $\beta_{k_1m_1} = \beta_{k_2m_2} = 0$ then $cc = ss = N_p$. As stated earlier, Schrama (1990) shows that the number of revolutions has to be larger than $2L$ in order to avoid

$$|\beta_{k_1m_1}| = |\beta_{k_2m_2}| \text{ for } k_1 \neq k_2 \text{ and } m_1 \neq m_2.$$

If $N_r > 2L$, therefore, $|\beta_{k_1m_1}| = |\beta_{k_2m_2}|$ if $m_1 = m_2$ and $k_1 = k_2$ for $m \neq 0$. Consequently, $A^TA = 0$ for $m_1 \neq m_2$. For $m = 0$ one always has $\beta_{k_10} = \beta_{-k_20}, \forall\ k_1 = -k_2$. Thus only half of the required frequencies remain. However, also the number of unknowns is divided by two since there are no $\bar{S}_{l0}$ coefficients to be determined. If $k = m = 0$ then $\beta_{00} = 0$.

Assuming that $N_r > 2L$ one then has for $m \neq 0$

$$\begin{aligned}
[A^TA]_{l_1l_2m}^{obs} &= \frac{N_p}{2} \sum_{k=-l_{12}(2)}^{l_{12}} \kappa_{l_1k}^{obs} \kappa_{l_2k}^{obs} \bar{F}_{l_1mk}(I) \bar{F}_{l_2mk}(I) \\
&= \frac{T_r}{2\Delta t} \sum_{k=-l_{12}(2)}^{l_{12}} \kappa_{l_1k}^{obs} \kappa_{l_2k}^{obs} \bar{F}_{l_1mk}(I) \bar{F}_{l_2mk}(I)
\end{aligned}$$

with $l_{12} := \min(l_1, l_2)$ where l_1 and l_2 have the same parity as a result of $k_1 = k_2 = k$. If $m = 0$ then the multiplication factor is N_p instead of $N_p/2$.

The above derivation shows that the normal matrix is block diagonal in the TT approach. Since we know that the normal matrix of the TT and TF approach are equal, the normal matrix corresponding to the TF approach is block diagonal as well. This is easy to verify. Taking the lumped coefficients as observations and the potential coefficients as unknowns, the design matrix is as follows. A single row

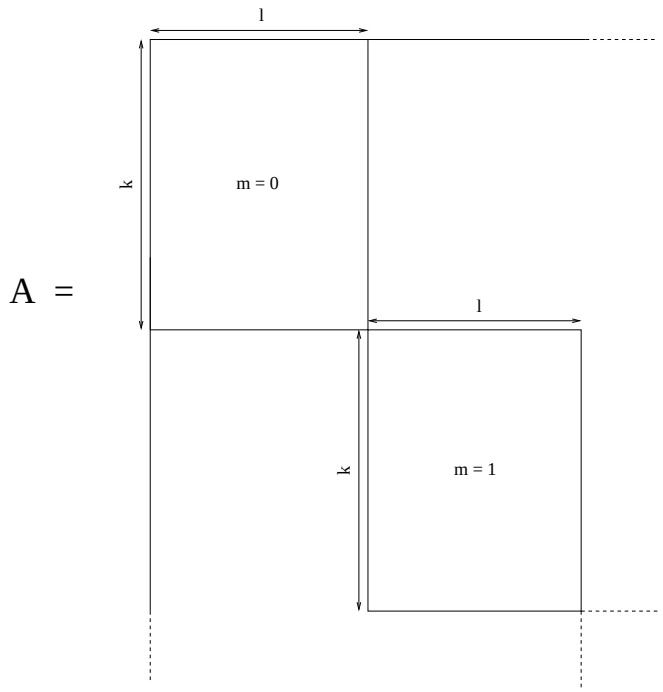

Figure 4.1: *The structure of the design matrix (*TF*).*

of A corresponds to one observation A^{obs}_{km} or B^{obs}_{km}, eq. (4.10). Thus, k and m are fixed while l runs. A single column of A corresponds to one unknown K_{lm}, l and m are fixed, while k runs. Taking the inner product of two columns of A to form $[A^TA]_{l_1m_1l_2m_2}$, therefore, results in a summation over k and $m_1 = m_2 = m$ since m is fixed for a single row and column. The design matrix consists of rectangular blocks for each m, other elements of A with the same row or column number are zero, compare fig. 4.1.

So far, only white noise was considered. One element of A^TPA is $\sigma^{-2}[A^TA]^{obs}_{l_1l_2m}$. In the TF approach, however, it is easy to include coloured noise. Suppose that the lumped coefficients are uncorrelated, but the variance has the form σ^2_{km}. This means that the noise depends on the frequency. It is a function of k and m, resulting in coloured noise. The Q_y matrix remains diagonal in this case and one element of A^TPA is

$$[A^TPA]^{obs}_{l_1l_2m} = \frac{T_r}{2\Delta t} \sum_{k=-l_{12}(2)}^{l_{12}} \kappa^{obs}_{l_1k}\kappa^{obs}_{l_2k}\bar{F}_{l_1mk}(I)\bar{F}_{l_2mk}(I)\frac{1}{\sigma^2_{km}}. \tag{4.12}$$

Due to the instrumental and environmental error sources together with the sampling rate, the error spectrum will be band limited and coloured. All frequencies below β_{min} are disturbed too much, while frequencies above β_{max} cannot be obtained because of the sampling rate (Koop, 1993). For GOCE, for example, $\beta_{min} = 2$ cpr (cycles per revolution), between 2 and 27 cpr the noise behaves like $1/\omega$, while above $|\beta_{km}| = 27$ cpr till the maximum frequency a flat spectrum is assumed. If the sampling period is 5 s, then the sampling rate is 0.2 Hz and the highest frequency is 0.1 Hz (Nyquist rate). At a height of 250 km the orbital period of the satellite is approximately 5370 s, so that $\beta_{max} = 5370 \times 0.1 = 537$ cpr. The maximum corresponding degree is $L = 505$. However, if the mission length is 30 days, then the satellite completes 481 revolutions. The maximum degree is therefore $L = 240$ in this case.

In the TT approach the Q_y matrix is full because of the error characteristics, which means that the measurements are correlated in time. The normal matrix, however, still is block diagonal under the condition that the model assumptions, circular orbit etc., hold (see Koop, 1993).

4.5 Satellite-to-satellite tracking

A GPS/GLONASS receiver will be mounted on board of the GOCE satellite enabling high-low *satellite-to-satellite tracking* (SST). The receiver provides code pseudo range observations and phase observations. Basically, a distinction can be made between two approaches for gravity field determination with these observations. In the first approach, the SST measurements are used directly, in the second approach the SST measurements are used first to obtain a best estimate of the GOCE orbit and the coordinates of this orbit are then used as pseudo observations (Visser *et al.*, 2000).

In the first approach, it is good practice to assume that for the undifferenced SST measurements, e.g. the phase measurements, the measurement errors are uncorrelated and behave as Gaussian noise. (Tiberius (1998) shows that in first approximation this is true but that it might be necessary to refine the noise model.) However, correlations between different measurements may exist:

- GPS ephemeris errors have typical periods of about 12 hours (1 orbital revolution);
- differenced measurements may be used to eliminate e.g. clock errors and the effect of selective availability (SA):
  - certain measurements show up in more than one differenced measurement;
  - in case of double differenced measurements atmospheric refraction errors may be introduced which may lead to correlated errors, etc.

In the second approach, successive estimates of the GOCE position, e.g. in the form of inertial Cartesian x, y and z coordinates may be used as pseudo observations. Depending on the precise orbit determination (POD) strategy, errors in these pseudo observations will be correlated in time differently. For example, in case of kinematic POD, for each epoch an independent estimate of the GOCE position will be made and orbit errors appear to be quite random in time. In case of dynamic POD, orbit errors will look more systematic (Visser *et al.*, 2000).

The indirect approach is adopted here. The POD is not discussed here, the orbit is assumed to be known with errors of a few cm (see Davis, 1997; Visser *et al.*, 2000). The observation equations for the pseudo observations are derived using the solution of the Hill equations, which relate the orbit disturbances to T in a local satellite frame (Schrama, 1989, 1990).

4.5.1 Hill equations

The reference potential U, cf. section 3.2, defines a circular, precessing orbit. The Hill equations describe the motion in the local x, y, z-triad and take the form

$$\begin{aligned} F_x &= \ddot{x} + 2n_0\dot{z} \\ F_y &= \ddot{y} + n_0^2 y \\ F_z &= \ddot{z} - 2n_0\dot{x} - 3n_0^2 z \end{aligned}$$

where $n_0 = \dot{\omega}_o$ represents the mean circular orbit velocity (mean motion).

The forcing functions F_x, F_y, F_z are developed as Fourier series from which particular solutions can be obtained (e.g. Schrama, 1989). The non-resonant particular solution is found by solving

$$\begin{aligned} P_x \cos\omega t + Q_x \sin\omega t &= \ddot{x} + 2n_0\dot{z} \\ P_y \cos\omega t + Q_y \sin\omega t &= \ddot{y} + n_0^2 y \\ P_z \cos\omega t + Q_z \sin\omega t &= \ddot{z} - 2n_0\dot{x} - 3n_0^2 z \end{aligned}$$

where P_x through Q_z symbolise time independent constants originating from the disturbing function. The solution of this system of equations becomes (Schrama, 1989):

$$\begin{aligned} x(t) &= \frac{(3n_0^2+\omega^2)P_x + 2n_0\omega Q_z}{\omega^2(n_0^2-\omega^2)}\cos\omega t + \frac{(3n_0^2+\omega^2)Q_x - 2n_0\omega P_z}{\omega^2(n_0^2-\omega^2)}\sin\omega t \\ y(t) &= \frac{P_y}{n_0^2-\omega^2}\cos\omega t + \frac{Q_y}{n_0^2-\omega^2}\sin\omega t \\ z(t) &= \frac{\omega P_z - 2n_0 Q_x}{\omega(n_0^2-\omega^2)}\cos\omega t + \frac{\omega Q_z - 2n_0 P_x}{\omega(n_0^2-\omega^2)}\sin\omega t \end{aligned} \tag{4.13}$$

showing that singularity occurs when $\omega = 0$ or $\omega = \pm n_0$. These cases require separate, resonant solutions, as described in Schrama (1989).

4.5.2 Observation equations

The SST observation equations are derived from (4.13), ω is replaced by $\beta_{km} n_0$ and all partial derivatives are substituted (Schrama, 1990):

$$\Delta obs(t) = \sum_{k=-L}^{L}\sum_{m=0}^{L} A_{km}^{obs}(I,r)\cos\psi_{km}(t) + B_{km}^{obs}(I,r)\sin\psi_{km}(t) \tag{4.14}$$

where $obs = x, y$ or z. Equation (4.14) relates the disturbances $\Delta obs(t)$ with respect to the reference orbit to the disturbing potential T. The observation equations become

$$\begin{pmatrix} A_{km}^{x}(I,r) \\ B_{km}^{x}(I,r) \end{pmatrix} = \sum_{l=lmin(2)}^{L} \kappa_{lmk}^{x}\bar{F}_{lm(l-k)/2}(I)\begin{pmatrix} \beta_{lm} \\ -\alpha_{lm} \end{pmatrix} \tag{4.15}$$

and

$$\begin{pmatrix} A_{km}^{y}(I,r) \\ B_{km}^{y}(I,r) \end{pmatrix} = \sum_{l=lmin(2)}^{L} \kappa_{lmk}^{y}\bar{F}_{lmk}^{*}(I)\begin{pmatrix} \beta_{lm} \\ -\alpha_{lm} \end{pmatrix} \tag{4.16}$$

and

$$\begin{pmatrix} A_{km}^{z}(I,r) \\ B_{km}^{z}(I,r) \end{pmatrix} = \sum_{l=lmin(2)}^{L} \kappa_{lmk}^{z}\bar{F}_{lm(l-k)/2}(I)\begin{pmatrix} \alpha_{lm} \\ \beta_{lm} \end{pmatrix} \tag{4.17}$$

for the along-track, cross-track and radial component respectively, with components κ_{lmk}^{obs} given in table 4.2, $\bar{F}_{lmk}^{*}$ the cross-track inclination functions and α_{lm}, β_{lm} as in (4.11), see also (Balmino *et al.*, 1996).

4.5.3 Block-diagonal normal matrix

Comparing the observation equations (4.15) - (4.17) with (4.10), it is obvious that here the normal matrix also becomes block diagonal with elements given by eq. (4.12). However, obs now is x, y or z and κ_{lk}^{obs} should be replaced by κ_{lmk}^{obs} from table 4.2. As with SGG the underlying assumptions are that the orbit is circular, the repeat is exact and the data flow is uninterrupted.

Note that the reference orbit is computed using the reference potential U, yielding a block-diagonal matrix. This approximation, however, will not be used in practice since it is not good enough. Instead, a reference field like JGM-3 or a field from the GRACE mission will be used and the approach based on the Hill equations can no longer be used, the normal matrix is full. Nevertheless the SST block-diagonal normal matrix should give a reasonable idea of the benefits of these observations.

Table 4.2: *'Eigenvalues' for satellite-to-satellite tracking.*

obs	κ_{lmk}^{obs}
x	$\frac{GM}{n_0 r^2}\left(\frac{R}{r}\right)^l \frac{2\beta_{km}(l+1)-(\beta_{km}^2+3)k}{\beta_{km}^2(\beta_{km}^2-1)}$
y	$\frac{GM}{n_0 r^2}\left(\frac{R}{r}\right)^l \frac{1}{1-\beta_{km}^2}$
z	$\frac{GM}{n_0 r^2}\left(\frac{R}{r}\right)^l \frac{\beta_{km}(l+1)-2k}{\beta_{km}(\beta_{km}^2-1)}$

4.6 Airborne gravimetry

A more homogeneous gravity field solution, in terms of precision, would be obtained having any other kind of gravity related observations in the polar gap regions in addition to the SGG and SST data. Airborne gravimetry is the most likely candidate to support a gradiometric mission. In this section the characteristics of airborne gravimetric observations are discussed, as well as the assumptions necessary to obtain a block-diagonal matrix.

4.6.1 Observation model

Scalar gravimetry in the usual sense provides point values of the magnitude of the gravity vector. Using a spring gravimeter, the difference in spring length between two locations corresponds to a gravity difference between the two locations. The situation is more complicated, however, if the gravimeter is attached to a moving vehicle like an airplane. The measured specific force is the sum of gravity and other accelerations due to the change in motion, airplane vibrations, etc. Furthermore, the orientation of the gravity sensor has to be known or stabilised (see Schwarz and Li, 1997).

Define the local-level coordinate system as $\{n, e, u\}$ where the axis n is pointing north, e is pointing east and u is up along the normal of the ellipsoid. Assume that gravity is measured in an airplane with the gravity sensor mounted on a local-level stable platform system. By means of GPS position (ϕ, λ, h) and velocity (v_n, v_e, v_u) are determined. The model for the scalar gravimetry then is

$$f_u + g_u = \dot{v}_u + E$$

where f_u vertical component of the measured specific force,
g_u vertical component of the gravitational vector,
$\dot{v}_u$ vertical vehicle acceleration,
E Eötvös effect.

The Eötvös effect is the sum of the Coriolis acceleration and centrifugal acceleration which are caused by the motion of the aircraft and the expression of the inertial vehicle acceleration in the rotating Earth-fixed coordinate system (Rummel, 1988). E is a function of the Earth's angular velocity ω_e and of ϕ, λ, h, v_n and v_e. It is, therefore, possible to obtain gravity information from these measurements, knowing the Earth's angular rate, compare Schwarz and Li (1997).

Meaningful results are obtained when the measurements have been low pass filtered. The current RMS values for gravity anomalies are 2-6 mGal with half wavelengths of 5-10 km (ibid). From now on it is assumed that a regular grid of gravity anomaly values at ground level is available, see also section 6.4. The observation model is as follows.

Let f be an L^2 function on the sphere which can be expressed in a convergent series of spherical harmonics, truncated at degree L

$$f(\theta,\lambda) = \sum_{l=0}^{L} \sum_{m=-l}^{l} \kappa_l \bar{K}_{lm} \bar{Y}_{lm}(\theta,\lambda)$$

with $\bar{K}_{lm}$ and $\bar{Y}_{lm}$ as in (3.2) and (3.3) respectively, and κ_l an eigenvalue. The coefficients $\bar{K}_{lm}$ are the unknowns to be determined from measurements $f(\theta,\lambda)$ in a regular grid. Let this grid be equidistant in both directions, that is, $\Delta\theta = \Delta\lambda$. The rows θ_i are numbered first, the columns λ_j are numbered next. If i runs from 0 to $N-1$ then j runs from 0 to $2N-1$, and there are $N \times 2N$ points. The linear model relating the measurements y to the unknowns x is $y = Ax$, with $y = f(\theta_i,\lambda_j), i = 0,\ldots,N-1, j = 0,\ldots,2N-1$, $x = \bar{K}_{lm}$. The columns of A consist of successive values $\kappa_l \bar{Y}_{lm}(\theta_i,\lambda_j)$ corresponding to the unknowns $\bar{K}_{lm}$ at the points (θ_i,λ_j) of the grid. Specifically:

$$\begin{bmatrix} \Delta g_{0,0} \\ \Delta g_{0,1} \\ \Delta g_{0,2} \\ \vdots \\ \Delta g_{N-1,2N-1} \end{bmatrix} = \begin{bmatrix} \kappa_l \bar{Y}_{lm}(\theta_0,\lambda_0) & \text{for all } l \text{ and } m \\ \kappa_l \bar{Y}_{lm}(\theta_0,\lambda_1) & \text{for all } l \text{ and } m \\ \kappa_l \bar{Y}_{lm}(\theta_0,\lambda_2) & \text{for all } l \text{ and } m \\ \vdots & \vdots \\ \kappa_l \bar{Y}_{lm}(\theta_{N-1},\lambda_{2N-1}) & \text{for all } l \text{ and } m \end{bmatrix} \begin{bmatrix} \vdots \\ \bar{K}_{lm} \\ \vdots \end{bmatrix}$$

where the observations f are gravity anomalies Δg.

4.6.2 Structure of the normal matrix

Colombo (1981) shows that the normal matrix becomes block diagonal for equidistant point values and a precision of the gravity anomaly observables independent of longitude. This is repeated here in a more comprehensive form.

The least-squares solution is $\hat{x} = (A^T Q_y^{-1} A)^{-1} A^T Q_y^{-1} y^{\varepsilon}$. Assuming that the measurements are uncorrelated and that the precision is the same for each latitude, Q_y is a diagonal matrix with N times $2N$ equal elements:

$$Q_y = \text{diag}(\sigma_0^2,\ldots,\sigma_0^2,\sigma_1^2,\ldots,\sigma_1^2,\ldots,\sigma_{N-1}^2,\ldots,\sigma_{N-1}^2).$$

The normal matrix, $A^T Q_y^{-1} A$, is formed by multiplying the columns of A and a row-wise summation with weights σ_i^{-2}. Every row corresponds to a unique point (θ_i,λ_j), and because Q_y is diagonal there is no correlation. Denoting the degree and order of the elements of A^T with l and m and that of A with n and k, the product $A^T Q_y^{-1} A$ for an arbitrary column of A with an arbitrary row of A^T gives:

$$\sum_{i=0}^{N-1} \sum_{j=0}^{2N-1} \kappa_l \bar{Y}_{lm}(\theta_i,\lambda_j) \kappa_n \bar{Y}_{nk}(\theta_i,\lambda_j) \sigma_i^{-2} = \kappa_l \kappa_n \sum_{i=0}^{N-1} \sigma_i^{-2} \sum_{j=0}^{2N-1} \bar{Y}_{lm}(\theta_i,\lambda_j) \bar{Y}_{nk}(\theta_i,\lambda_j).$$

Using the definition of $\bar{Y}_{lm}$, eq. (3.3), yields

$$cov(\bar{C}_{lm},\bar{C}_{nk}) = \kappa_l \kappa_n \sum_{i=0}^{N-1} \sigma_i^{-2} \bar{P}_{lm}(\cos\theta_i) \bar{P}_{nk}(\cos\theta_i) \sum_{j=0}^{2N-1} \cos m\lambda_j \cos k\lambda_j$$

and similar expressions for $cov(\bar{C}_{lm},\bar{S}_{n|k|})$ and $cov(\bar{S}_{l|m|},\bar{S}_{n|k|})$. The distance in λ-direction is constant, $\Delta\lambda = 2\pi/2N$. The respective λ-terms, therefore, are of the form

$$\sum_{j=0}^{2N-1} \cos mj\Delta\lambda \cos kj\Delta\lambda, \quad m,k < N.$$

The non-zero elements of $cov(C,C)$ and $cov(S,S)$ are those for which $m = k$: $cov(C,S)$ is always zero. The summations

$$\sum_{j=0}^{2N-1} \cos^2 mj\Delta\lambda \quad \text{and} \quad \sum_{j=0}^{2N-1} \sin^2 |m|j\Delta\lambda, \quad m < N$$

are the discrete counterparts of the integrals from 0 to 2π of $\sin^2 |m|\lambda$ and $\cos^2 m\lambda$, N corresponds with π, and it is

$$\sum_{j=0}^{2N-1} \cos^2 mj\Delta\lambda = \begin{cases} 2N, & m = 0 \\ N, & m > 0 \end{cases} \quad \text{and} \quad \sum_{j=0}^{2N-1} \sin^2 |m|j\Delta\lambda = N, \quad m < 0$$

which gives

$$\begin{aligned} cov(\bar{C}_{lm}, \bar{C}_{nm}) &= \kappa_l \kappa_n \sum_{i=0}^{N-1} \sigma_i^{-2} \bar{P}_{lm}(\cos\theta_i) \bar{P}_{nm}(\cos\theta_i) \begin{cases} 2N, & m = 0 \\ N, & m > 0 \end{cases} \\ cov(\bar{S}_{l|m|}, \bar{S}_{n|m|}) &= \kappa_l \kappa_n \sum_{i=0}^{N-1} \sigma_i^{-2} \bar{P}_{l|m|}(\cos\theta_i) \bar{P}_{n|m|}(\cos\theta_i) N. \end{aligned}$$

This shows that the elements of the diagonal blocks of the normal matrix are non-zero, whereas other elements are zero.

Remarks. The condition $L < N$ must hold to obtain an over-determined system of equations. The sampling in longitude should be regular so one can write $\lambda_j = j\Delta\lambda$, a regular sampling in latitude direction is not strictly necessary to obtain a block-diagonal structure. To preserve the block diagonality, the σ of the observations must not depend on longitude since it must be outside the j-summation.

North-south symmetry

Colombo (1981) also shows that if $\sigma_i^2 = \sigma_{N-1-i}^2$ and the grid is symmetric with respect to the equator, the even and odd degrees are separated, that is, $cov = 0$ if $l - n$ is odd.

4.7 Summary

The combination of computing a solution for the system of observation equations and the quality assessment of the solution, dictates the need for the availability of the inverse of the (regularised) normal matrix. Since the number of unknowns is large, the direct computation of the true inverse is avoided and replaced by an approximate inverse. Iteration will lead to the correct solution. In addition, if the approximate inverse is close to the true inverse then the former may be used in the quality assessment.

The observation types in this study are SGG, SST and airborne gravimetry. The observation equations as well as the assumptions necessary to obtain a block-diagonal normal matrix have been discussed for these observables. A block-diagonal matrix allows for an easy inverse computation, while the quality description based on this matrix is expected to be close to the 'true' quality.

5

Gravity field models from SGG *only*

5.1 Introduction

Gravity field determination from satellite gravity gradiometry is an inverse problem which is ill-posed: (i) the downward continuation of the data amplifies noise (stability), (ii) the polar areas are not covered because of the orbit inclination (uniqueness). Properties (i) and (ii), however, hold for the continuous case. Because the linear system of equations is discrete, the best-approximate solution with minimal norm is unique. Furthermore, the discrete system is stable. However, the original properties (i) and (ii) may lead to a severely ill-conditioned problem. Consequently, if data errors are present, a least-squares solution tends to lead to strong oscillations in the solution with large amplitude. Regularisation, therefore, is mandatory and various regularisation methods discussed in chapter 2 will be applied here. The measures of chapter 3 will be used to assess the quality of the different solutions and the block-diagonal approach of chapter 4 is adopted.

In this chapter gravity field determination from only SGG measurements is considered. It was decided to leave out global SST or gravimetry at first since treating gradiometry data alone gives a better idea of the possibilities and limitations of these type of data. The combination of all three data types is discussed in chapter 6. In this chapter the orbit is assumed to be known. Schrama (1990) shows that at a height of 200 km orbit errors of 10 m give errors in the gradient at the 0.01 E level, whereas orbit errors of 1 m give errors of 0.001 E (rule of thumb). The expected accuracy of the SGG measurements is at the 1 mE level in the measurement bandwidth. The expected orbit determination precision for GOCE is a few cm's (Visser *et al.*, 2000), and therefore, the precise orbit determination (POD) from GPS observations is accurate enough in view of the expected precision of the SGG observations.

The outline of the chapter is as follows. First, the gravity field recovery is tested with faultless measurements for a gradiometric mission in a circular polar orbit. This will fix the lower bound of the expected precision. Secondly, the inclination is changed to a GOCE mission inclination of 96.6°, which introduces two polar gaps without measurements, resulting in a degree versus order scheme in a 'wedge' of badly determined coefficients (low orders, all degrees), (Van Gelderen and Koop, 1997). Both a circular and a non-circular orbit are considered. The latter, referred to as GOCE orbit, gives information on the effect of model errors: the orbit is not circular but the normal matrix becomes block diagonal under the assumption of a circular orbit. Finally, noise is added to the observations and Tikhonov regularisation as well as biased estimation are used to compute a gravity field solution. The discussion of the gravity field recovery results is preceded by a few remarks on the computation of the gravity gradients, which

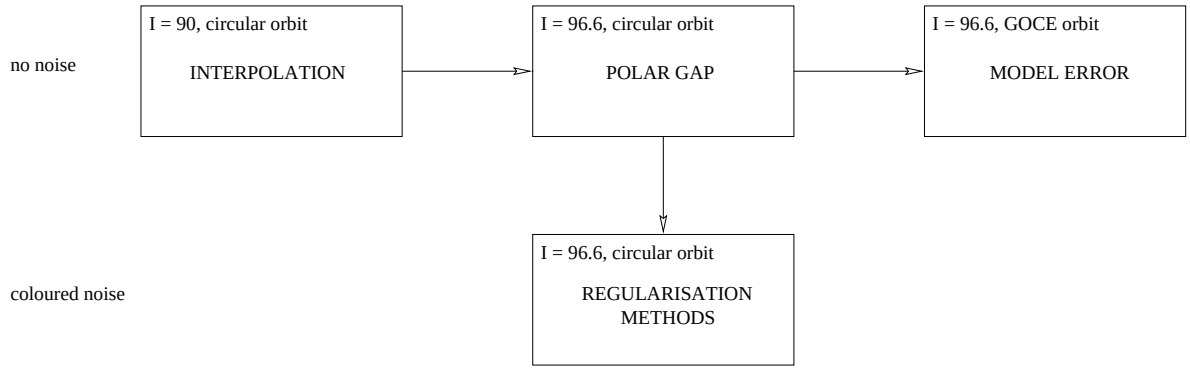

Figure 5.1: *Outline of chapter 5.*

is the direct problem referred to as synthesis, as well as the analysis which is the inverse problem. The outline of this chapter is summarised in figure 5.1.

Earlier comparisons of regularisation methods applied to gradiometry are based on error propagation. Bouman and Koop (1998a), for example, compare Tikhonov regularisation (TR) and generalised biased estimation (GBE) for several gravity gradiometry mission scenarios, while Bouman and Koop (1998b) compare TR with signal and second derivative constraint. Bouman (1998b) compares regularisation methods and parameter choices using airborne gravimetry (with simulated observations), while Floberghagen and Bouman (1998) tested several parameter choice rules in gravity field determination of the Moon.

5.2 Synthesis and analysis

5.2.1 Synthesis

The synthesis of the gradiometric observations (the direct problem) should preferably be fast. As before, let y be the observations and x the unknowns. The computation of

$$y = Ax \quad \text{or} \quad y^{\varepsilon} = Ax + \epsilon$$

has to be as efficient as possible since several missions, including iterations, are studied. Therefore, the direct computation of the V_{xx}, V_{yy} and V_{zz} gradients along the orbit is avoided. Instead, grid points with gradient values are computed using FFT methods and the correct values are obtained by interpolation, compare appendix C. The interpolation error turned out to be smaller than 10^{-5} E for the non-circular GOCE orbit for a maximum degree of $L = 360$. The non-circular orbit was generated with the GEODYN software (Eddy *et al.*, 1990).

For all missions discussed hereafter, the following specifications have been chosen (cf. ESA, 1999). The height of the satellite orbit is approximately 250 km, the mission length $N_d = 29$ days, the sampling period is 5 s, the inclination is either $I = 90^{\circ}$ or $I = 96.6^{\circ}$. The minimum degree is $l_{min} = 2$ while the maximum degree is $L = 180$. The expected GOCE observation window is two times six months. The maximum resolvable degree will be somewhere between $L = 240$ and $L = 300$. For the current study a mission length of one month is chosen to keep the computations manageable.

Measurements with and without noise are generated. The measurements without noise are not perfect due to the interpolation errors and round off errors. The error covariance matrix of these measurements is chosen as a scaled unit matrix with scale factor 10^{-10}. The noise added to the measurements is coloured, the error PSD for SGG is depicted in figure 5.2. The dashed line indicates white noise, which for V_{yy} is at the level of $1.5 \times 10^{-3}\text{E}/\sqrt{\text{Hz}}$. For the coloured noise PSD $\beta_{min} = 2$ cpr, there is a $1/\omega$ behaviour for $2 \leq \beta_{km} \leq 27$ cpr and a flat spectrum for $\beta_{km} > 27$ cpr (ω stands for frequency here). Because of the

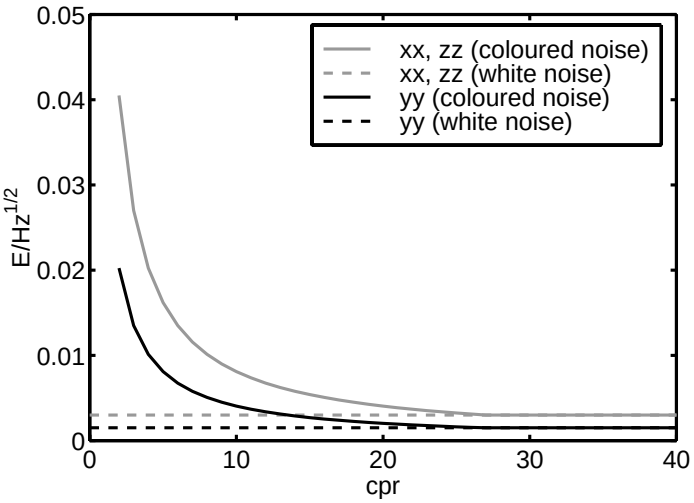

Figure 5.2: *The square root of the error* PSD *for* SGG.

sampling rate of 0.2 Hz, $\beta_{max} = 537$ cpr, cf. section 4.4.3. The error PSD's for V_{xx} and V_{zz} are a factor of two more pessimistic than that of V_{yy}, see also ESA (1999).

5.2.2 Analysis

In chapter 2 several methods to analyse observations when the inverse problem is ill-posed are listed. Of these methods the SVD methods will not be used because they involve the decomposition of the design matrix which requires too much CPU-time and memory. There are approximately half a million observations in 30 days with a 5 second sampling, while the number of unknowns is roughly 32400. Furthermore, none of the discussed parameter choice rules has been used, the minimisation of the trace of the MSEM is used instead.

Since the maximum degree solved for is $L = 180$, the downward continuation is compensated by the higher short wavelength sensitivity of the gravity gradients. In the space domain the corresponding resolution is a $1^\circ \times 1^\circ$ grid. In fact the discretisation as well as the truncation act as regularisers. Would L be much larger, then the downward continuation instability would show up, compare figure 5.3 which depicts the amplification per degree for V_{zz}. It should be noted that the amplification or attenuation of V_{xx} and V_{yy} differs from that of V_{zz}. Moreover, the power of the higher degree coefficients decreases $\sim 1/l^2$, that is, the higher degrees have less power.

A first idea of the relative influence of a polar gap compared to downward continuation one gets from figure 5.4. The maximum condition number of the normal matrix is shown for the maximum degrees $L = 120, 180, \ldots, 360$. The dash-dot line shows the maximum condition numbers for SGG observations at a height of 0 km and an inclination of 96.6°, $N_d = 6$ months. The full line shows the maximum condition numbers for SGG observations in a polar orbit at a height of 250 km, $N_d = 6$ months. The effect of coloured noise is excluded in this example by assuming white noise. Clearly, the polar gaps yield extremely large condition numbers, up to 10^{19} for $L = 360$. Although the downward continuation condition numbers are also large, they are 3-10 orders smaller.

It should be noted that not only the condition number is important, the size of the eigenvalues is important as well. If, for example, the condition number is small, say 10^2, then the eigenvalues may be small, say $10^{-8} \leq \lambda_i \leq 10^{-6}$. Since the condition number is small, the solution then will be numerically stable but the absolute precision will be small. On the other hand, if the condition number is small and the eigenvalues are large, say $10^3 \leq \lambda_i \leq 10^5$, then the precision of the solution is high.

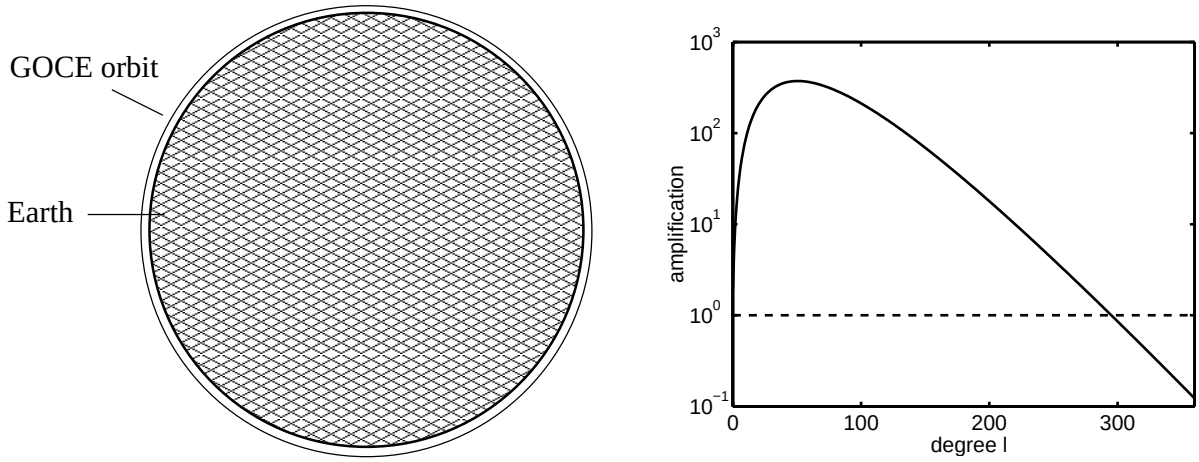

Figure 5.3: *Circular orbit at 250 km and a spherical assumed Earth with a radius of 6378 km displayed at the same scale (left). The amplification factor per degree for V_{zz} (right). Shown is the function $(\frac{R}{R+h})^{l+1}(l+1)(l+2)$, the height h is 250 km.*

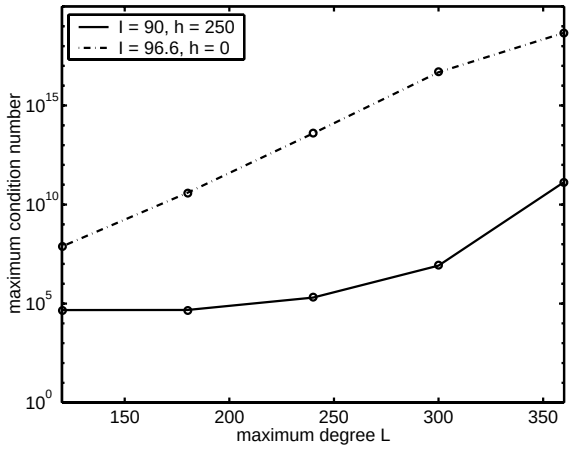

Figure 5.4: *The maximum condition number of the normal matrix for increasing degree. The full line gives the maximum condition numbers for* SGG*, polar orbit and a height of 250 km. The dash-dot line gives the maximum condition numbers for* SGG*, inclined orbit ($I = 96.6^{\circ}$) and a height of 0 km. For both cases white noise is assumed.*

5.3 Observations without noise

First of all a few bench marks are set, testing the effect of model and interpolation errors. Therefore, observations without noise will be used at the first stage and noise will be added to them later. The approach is as follows. The effect of the interpolation error and a small model error (no exact repeat) is tested solving gravity potential coefficients using SGG observations in a polar circular orbit. A least-squares solution is feasible because $L = 180$, and the error variance-covariance matrix is a scaled unit matrix. Next the effect of a polar gap is studied by looking at a circular inclined orbit. Even though there are two polar gaps, regularisation is not necessary due to the very small data errors. Finally, a 'true' GOCE orbit is studied. The orbit is no longer circular, and therefore larger model errors are present. It is shown that these model errors can be overcome by iteration, compare section 4.2. Van den IJssel *et al.* (2000) present results similar to those here.

5.3.1 Circular polar orbit

If a satellite collects its measurements along a polar orbit, then after a number of revolutions these measurements are more or less evenly distributed in a shell close to the Earth, see figure 5.3. The distribution depends, of course, on the sampling rate and on the distance between the tracks (the repeat period). However, there will certainly be no 'polar gaps' and, given the GOCE mission parameters, provided the sampling period is short enough ($\leq$ 14 s) and the repeat period long enough ($\geq$ 23 days), all coefficients up to degree and order $L = 180$ may be estimated uniquely (Koop, 1993; Rummel *et al.*, 1993). Certainly, the attenuation effect limits the maximum resolvable L, but observing V_{xx}, V_{yy} and V_{zz} together the coefficients up to degree and order $L = 180$ are estimable.

The results for the circular polar orbit serve as a bench mark for the other mission designs. No polar gaps are involved, only one model error (no exact repeat), computer round-off and interpolation errors are possible error sources. The maximum degree is $L = 180$ for the synthesis as well as the analysis step. In total 499752 observations have been used which is equivalent to 28.9 days. The satellite orbit has a height of 252.8 km with respect to a spherical approximation of the Earth.

Because of the 'perfect' observations, regularisation is not necessary. The least-squares solution is

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

where the weight matrix P is the inverse of a scaled unit matrix. The scaling factor is unimportant for the solution itself since it drops out. The scaling would matter when using the error variance-covariance matrix Q_x of $\hat{x}$ but this is not pursued here.

Figure 5.5, left, displays the relative differences between OSU91A, the 'true' gravity field in the simulations, and the solved coefficients from the almost perfect observations for the initial solution. The maximum relative difference is 4.9 for the first iteration and after the second iteration the relative error in the $\bar{C}_{lm}$ coefficients is below 0.1% for almost all coefficients (not shown). Throughout the remainder of this thesis the degree versus order plots will be restricted to the$\bar{C}_{lm}$ coefficients since the $\bar{S}_{lm}$ coefficients yield similar results.

With these coefficient differences geoid errors are computed, as shown in figure 5.5, right panel. The maximum geoid errors are in the neighbourhood of $\lambda = 0, \phi = 0$ which is caused by the not exact closure of the repeat. The not exact repeat yields maxima in regions with large geoid signals as well, such as the Andes in South America and near India and Indonesia. The RMS of the geoid errors is already small, 0.24 mm, but the extremes due to the model error are too large (1 cm). The second iteration yields a maximum geoid error below 1 mm with an RMS of 0.01 mm, compare table 5.1. Again the maximum geoid error is due to the model error, but this time it is negligible. Apparently, two iterations are sufficient.

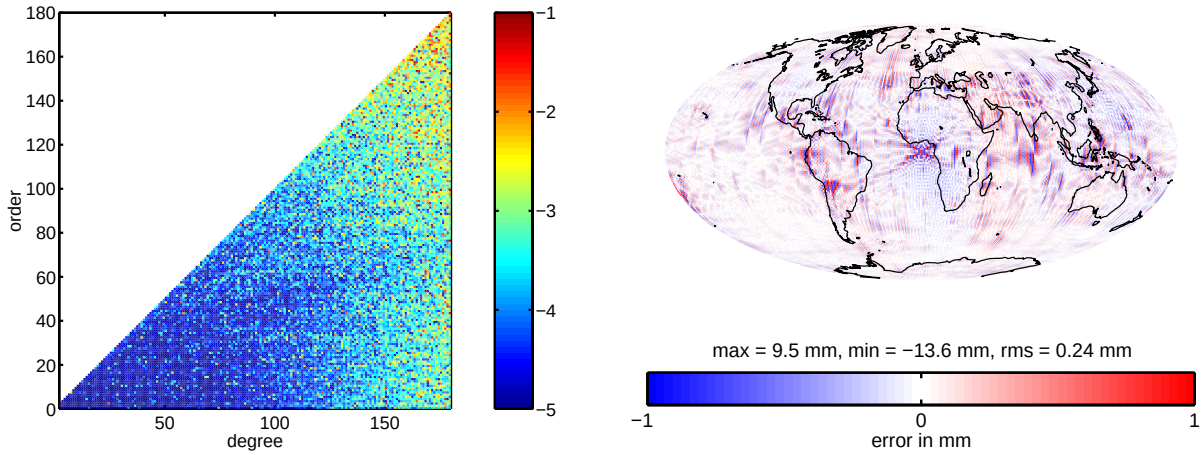

Figure 5.5: *Circular polar orbit, no noise. Relative error in the spherical harmonic coefficients, first iteration (left). The differences (*OSU91A *- solved)/*OSU91A *are shown on a logarithmic scale. The right panel displays the geoid errors due to the coefficient differences, first iteration.*

Table 5.1: *Global geoid errors due to the coefficient differences* OSU91A *- solved, final solutions. Units are in mm. Two, seven, and eight iterations are required for the three cases respectively.*

	geoid error (mm)		
mission	RMS	max	min
polar	0.01	0.3	-0.1
GOCE circular	0.72	11.9	-5.1
GOCE	0.87	14.3	-6.1

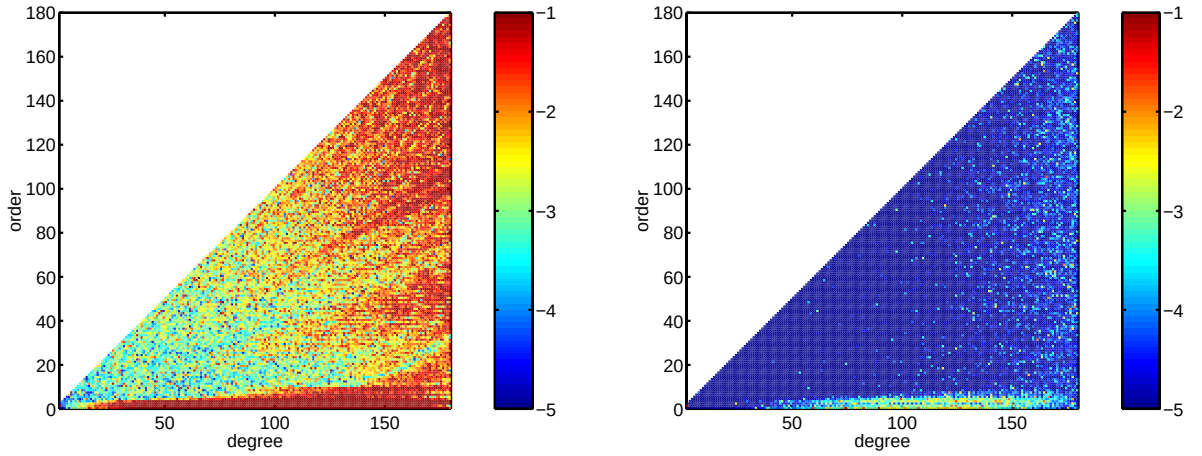

Figure 5.6: *Circular inclined orbit, no noise. Relative error in the spherical harmonic coefficients, first iteration (left) and after seven iterations (right). The differences (*OSU91A *- solved)/*OSU91A *are shown on a logarithmic scale.*

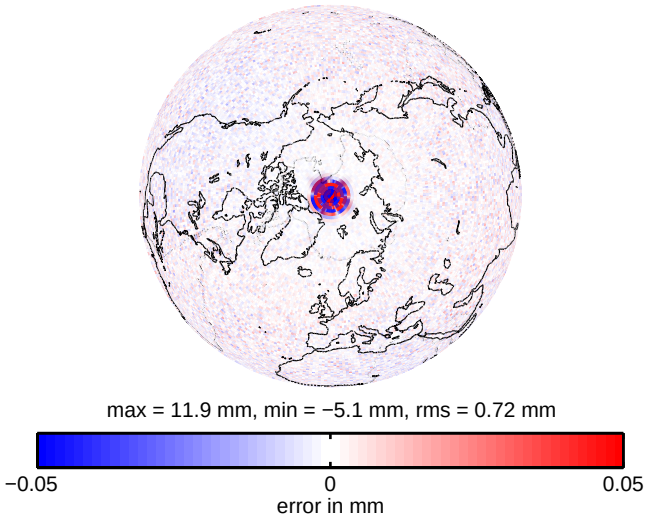

Figure 5.7: *Circular inclined orbit, no noise. Geoid errors after seven iterations due to the coefficient differences* OSU91A *- solved.*

5.3.2 Circular inclined orbit

The GOCE orbit will not be polar and because $I \neq 90^{\circ}$ two polar gaps exist. Due to the polar gaps the parameter estimation is ill-posed. Even though for GOCE the total area of the gaps is small compared to the total observation area, and even though the interpolation errors and model errors are small, the ill-posedness might cause the solution errors to be large. Therefore, an inclined circular orbit is tested. The inclination is 96.6° and the mission length and orbital height are exactly equal to those of the polar orbit case. This test case reveals the effect of small errors in the presence of polar gaps. Again regularisation appeared to be unnecessary, and with successive iterations the model error (no exact repeat) has to be overcome.

Figure 5.6 displays the relative coefficient error for the first iteration. Especially the low order coefficients are affected, which is typical for a polar gap. The maximum relative difference is of the order 10^5. The maximum corresponding geoid errors of 10^2 m are located in the polar regions (not shown). Although the interpolation errors and model errors are small, the ill-posedness of the inverse problem yields an oscillating solution with large amplitude in the polar regions. Fortunately, the solution is affected only locally, which means that also here the model error can be overcome by iteration. Consequently, the geoid error can be reduced significantly.

In total seven iterations are needed to obtain almost the exact solution, the eighth iteration did not show any improvement. The relative coefficient error is below 0.01% for the major part of the coefficients, and it is between 0.1 and 1% for part of the low order coefficients, figure 5.6, right. The RMS geoid error is 0.72 mm with a maximum of 11.9 mm and a minimum of -5.1 mm located in the polar regions, see figure 5.7. The RMS in the measurement area, that is, $-83.4^{\circ} \leq \phi \leq 83.4^{\circ}$, is 5×10^{-3} mm. These results show that even small errors, like model errors have some significance since small coefficient errors and geoid errors remain. The best-approximate solution with minimum norm is unique but, evidently, this need not be the exact solution. Compared to the effect of measurement noise hereafter, however, the errors discussed here are negligible. Figure 5.7 shows that the total area without observations is small indeed.

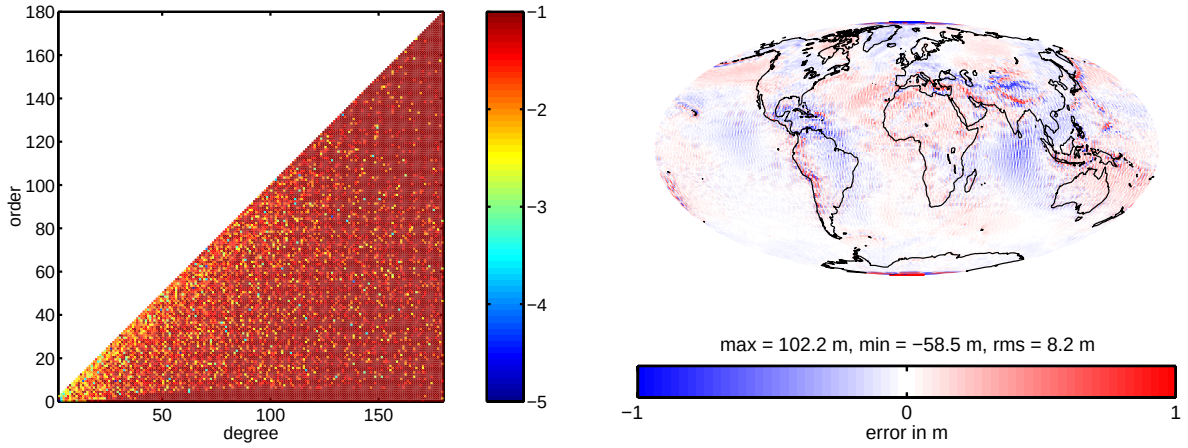

Figure 5.8: GOCE *orbit, no noise. Relative error in the spherical harmonic coefficients, first iteration (left). The differences (*OSU91A *- solved)/*OSU91A *are shown on a logarithmic scale. The right panel displays the geoid errors due to the coefficient differences, first iteration.*

5.3.3 Non-circular GOCE orbit

The GOCE orbit will in reality not be exactly circular. On the average the orbit can be described by a circular orbit but the maximum vertical deviations are plus and minus 9 km for a 1 month repeat. Orbit integration yielded a repeat after 501145 observations or 29.0 days at an average height of 244.6 km. Again the inclination is $I = 96.6^\circ$.

The relative coefficient differences after one iteration for such a GOCE-like orbit are displayed in figure 5.8. The maximum error is of the order of 10^5 with clearly the highest errors at the low orders, but also high degree and orders are affected. The coefficient errors after one iteration have an RMS of 8.2 m, the maximum is 102.2 m and the minimum -58.5 m. The corresponding geoid errors are shown in figure 5.8. Again, the largest errors are located in the polar areas, but due to the model error, large errors occur in areas where the gravity gradient is large.

The first iteration yields extremely large errors. Fortunately, the iteration converges and the coefficient errors after eight iterations are similar to those displayed in figure 5.6, right. The ninth iteration did not show any improvement. The relative coefficient errors are at the level of the circular orbit coefficient errors. The RMS geoid error after eight iterations is 0.87 mm, with a maximum of 14.3 mm, while the minimum is -6.1 mm. Again, in the measurement area the RMS geoid error is 5×10^{-3} mm. It is therefore concluded that the non-circular orbit model error can be overcome by iteration.

5.4 Noisy observations

The test with the observations without noise shows that even with a polar gap the coefficients can be recovered almost exactly. The circular approximation in case of the true GOCE orbit is unimportant, since it can be removed by iteration. From now on, therefore, only the inclined circular orbit is considered and several regularisation methods will be tested and the quality of the solutions will be studied in detail. To this end, simulated coloured noise errors are added to the observations (see also figure 5.2).

The noise characteristic is such that the low frequencies, that is small β_{km}, are badly determined. Consequently, the spherical harmonic coefficients with m not too large are affected by the coloured noise. For a specific m, say $m = 10$, the degree l runs from $l = 10$ to $L = 180$. Considering the PSD in figure 5.2 one might argue that the frequencies with β_{km} smaller than five really experience the coloured noise. If $m = 10$ then $-4 \leq k \leq 5$ are the corresponding frequencies. If l is small, and therefore $l - m$ small, then a relatively large number of frequencies is affected since k runs from $-l$ to l, see eq. (4.12).

Table 5.2: MSE *and bias of the* TR *solutions.*

no.	constraint	α	MSE[a]	BNR	max(BSR)[b]
1	signal	2.4	1	0.2	0.2 (54.8)
1a	1st deriv.	2.0×10^{-4}	1.1	0.2	0.3 (518.4)
1b	2nd deriv.	7.5×10^{-9}	1.4	0.3	0.3 (228.1)

[a] In proportion to mission 1.

[b] The maximum BSR based on the OSU91A degree-order variances. The maximum BSR between brackets has been computed using the solution coefficients.

However, if l is large, then the number of frequencies distorted by the coloured noise is relatively small. In addition, if m is large, then also the number of km combinations smaller than five is reduced. In summary: if m is small as well as $l - m$, then the influence of the coloured noise is large, otherwise the influence is small.

5.4.1 Tikhonov regularisation

As we know from chapter 2, Tikhonov regularisation (TR) amounts to minimising

$$\|Ax - y^\varepsilon\|_P^2 + \alpha\|Lx\|_K^2$$

with solution

$$\hat{x}_\alpha = (A^T P A + \alpha L^T K L)^{-1} A^T P y^\varepsilon.$$

The operator L is a linear differential operator and TR with signal, first and second derivative constraint is tested. Matrix K is diagonal with elements $10^{10} l^4$ which is the inverse of the well known Kaula rule. The matrix $L = I$ for the signal constraint, whereas it has diagonal elements $(l+1)$ and $(l+1)(l+2)$ for the first and second derivative constraint respectively (the radial derivative is used)[1]. The regularisation parameter is obtained by minimising the trace of the MSEM. Results are presented in table 5.2. The regularisation parameter α is obtained by minimising the trace of the MSEM, that is, minimising the MSE.

Comparing mission 1a with 1, the low degrees are constrained less and the high degrees are constrained more: $(l+1)^2\alpha = 2 \cdot 10^{-3}, \ldots, 6.6$ for $l = 2, \ldots, 180$. The same holds true for 1b with respect to 1a and 1: $(l+1)^2(l+2)^2\alpha = 1 \cdot 10^{-6}, \ldots, 8.1$ for $l = 2, \ldots, 180$. Since there are more high degree coefficients than low degree coefficients, the BNR increases from 1 to 1b (fifth column in table 5.2). In terms of the MSE, TR with signal constraint is expected to give slightly better results than TR with first or second derivative constraint.

Coefficient errors. Figure 5.9 shows the relative differences between the true coefficients (OSU91A) and the solved coefficients (TR with signal constraint). The second iteration did not give further improvement. As the coefficient differences for TR with first and second derivative constraint are almost exactly equal to those of TR with signal constraint, they are not shown.

The bias-to-signal ratio (BSR) of TR(0) and TR(2) are shown in figure 5.10. (TR(0) denotes regularisation with signal constraint, TR(1) with first derivative constraint, etc.) Instead of the true signal, that is, the OSU91A coefficients or the solution coefficients, the degree-order variances of OSU91A have been used to avoid excessive extremes caused by the very small coefficients, compare table 5.2. In accordance with the stronger regularisation for higher degrees, the higher degree coefficients of TR(2) are more biased. The BSR of TR(1) lies between that of TR(0) and TR(2) and is not shown. The major part of the bias is located in the low order coefficients due to the polar gap.

[1] Bouman and Koop (1998b) erroneously used the inverse of $(l+1)(l+2)$ for the second derivative constraint.

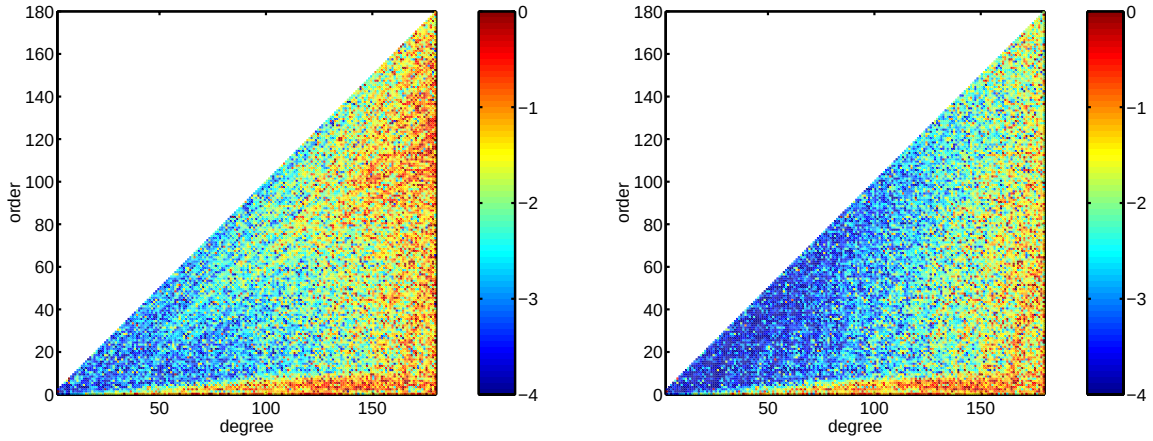

Figure 5.9: TR. *Relative error in the spherical harmonic coefficients, initial solution (left) and after one iteration (right). The differences (*OSU91A *- solved)/*OSU91A *are shown on a logarithmic scale.*

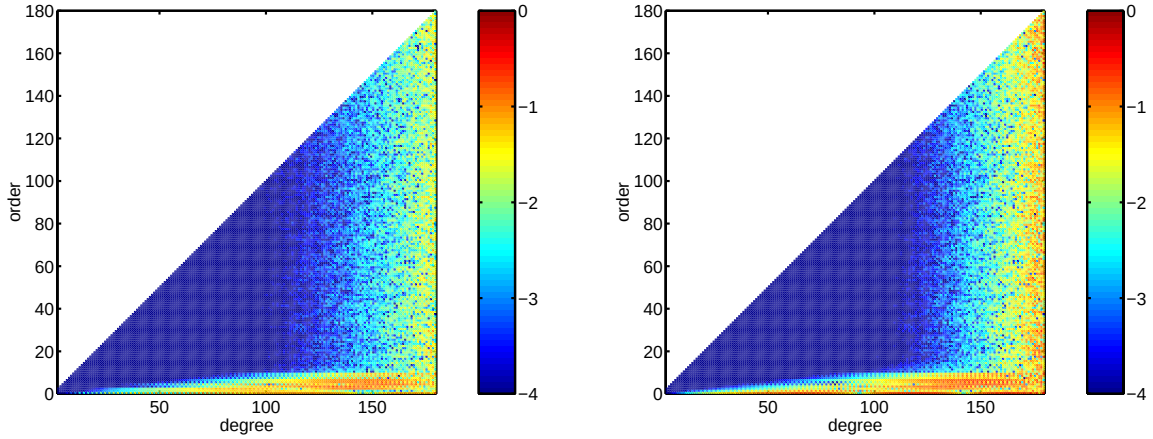

Figure 5.10: BSR *for* TR *with signal constraint (left) and second derivative constraint (right), the scale is logarithmic. The* BSR *was computed using the* OSU91A *degree-order variances.*

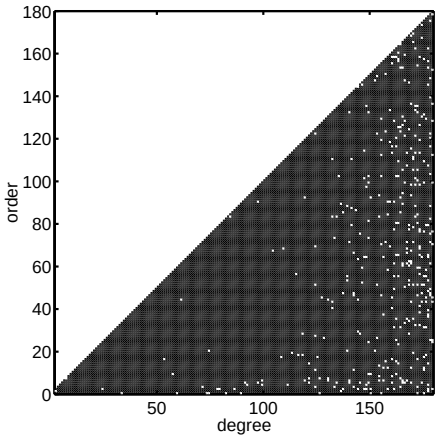

Figure 5.11: TR *with signal constraint.* SNR, *maximum is* 10^4. *The coefficients with an* SNR *larger than or equal to one are shown in black.*

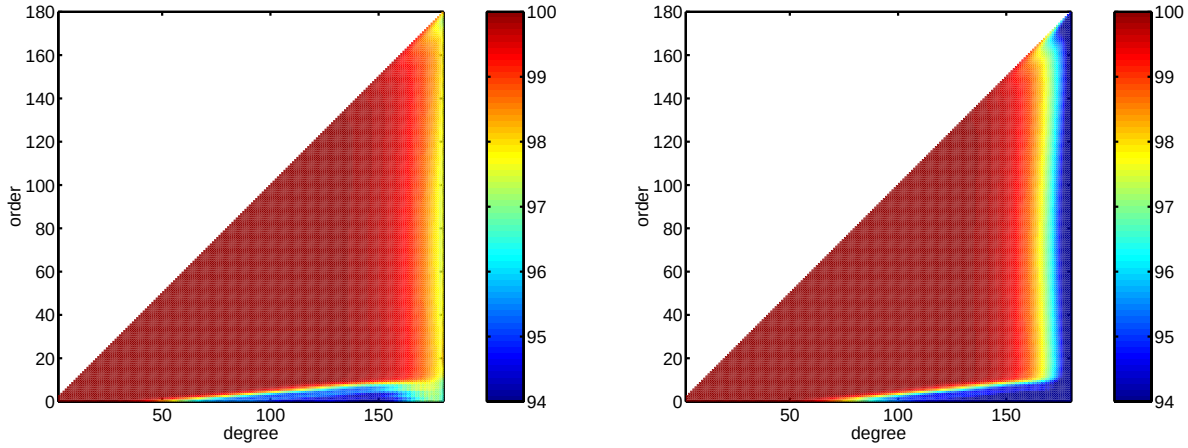

Figure 5.12: *Contribution measure for unbiased estimators.* TR *with signal constraint (left) and with second derivative constraint (right).*

The signal-to-noise ratio (SNR) of the TR with signal constraint solution is shown in figure 5.11. The SNR of TR(1) and TR(2) are similar. The maximum SNR is 10^4, that is, 4 significant digits are determined. The SNR decreases for increasing degree and, naturally, the SNR of the low order coefficients is low.

The contribution measures, for the unbiased solution (eq. (3.11)), for the TR solutions are displayed in figure 5.12. The contribution measure for the first derivative constraint is not shown since it is almost exactly equal to the second derivative constraint contribution measure. The contribution is given as a number between 0 and 100%. For all three solutions the contribution is above 89%, with minimum values of 94, 91 and 89% for TR(0), TR(1) and TR(2), respectively. Clearly, the contribution measure decreases for increasing degree, and again the low order coefficients have a relatively low contribution. Although a percentage of 90% or higher does not seem to be bad at all, the coefficient differences between the solution and OSU91A can be large, up to the size of the signal as shown earlier. Altogether, the contribution measure seems to be of little absolute value, but it is a tool to discriminate between the well determined and the not so well determined coefficients.

Bias. Recall that the MSEM is

$$MSEM = Q_x + \Delta A x x^T \Delta A^T,$$

see chapter 2. The bias term yields a full matrix while Q_x is block diagonal. Because it is difficult to handle large matrices, it was decided to stick to the block-diagonal approach, even and odd degrees, however, are no longer separated. In order to validate the block-diagonal approach for the bias part of the MSEM, the bias in the coefficients is translated into geoid heights and compared with bias propagation to geoid heights. That is, the bias of the TR(0) solution

$$\Delta x = -(A^T P A + \alpha K)^{-1} \alpha K x$$

is computed and with these coefficients geoid heights are computed. Moreover, the block-diagonal bias part of the MSEM is propagated to geoid heights. Figure 5.13 shows the geoid errors due to the bias averaged in λ-direction. Furthermore, the bias is compared with the propagated noise. From these figures one may conclude that the block-diagonal bias approach is valuable, that is, the approximation closely resembles the true errors (left panel). Furthermore, the bias is negligible with respect to the noise in the measurement area (right panel).

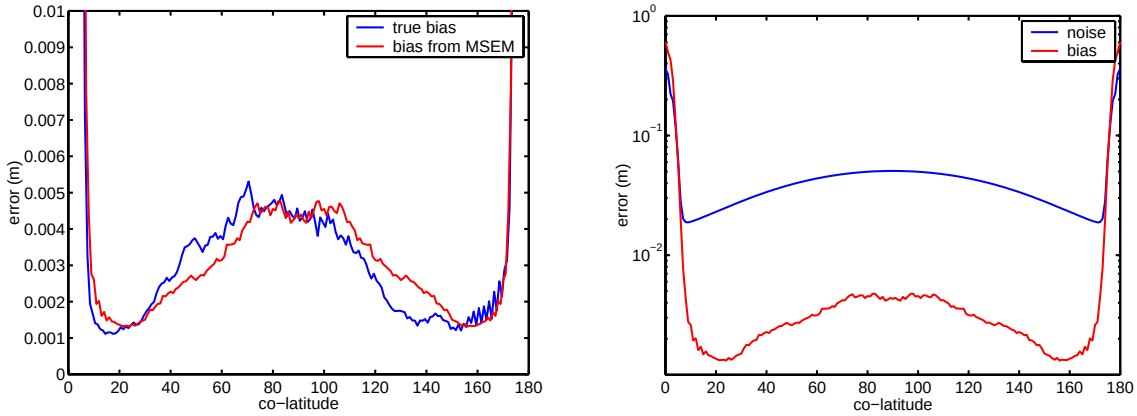

Figure 5.13: TR *with signal constraint: true geoid bias and block-diagonal bias part of the* MSEM *propagated to geoid heights (left) and propagated bias and noise on a logarithmic scale (right).*

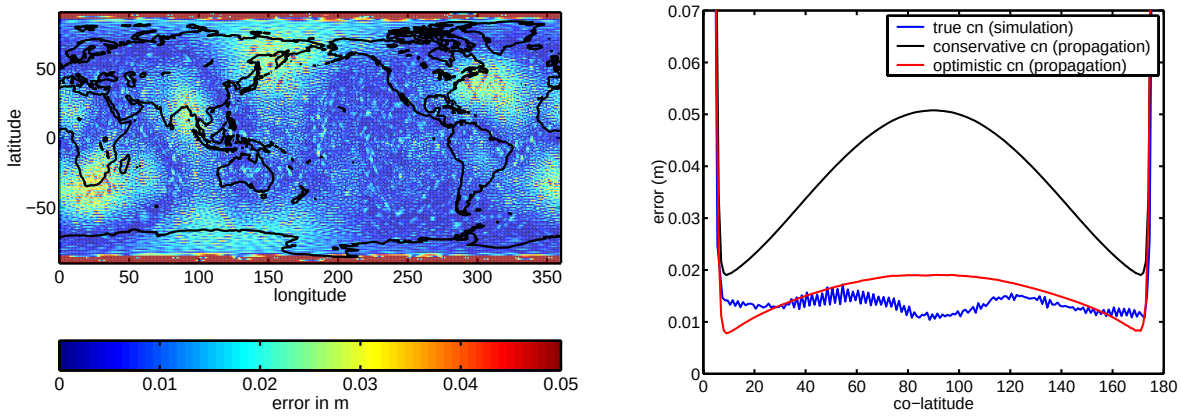

Figure 5.14: TR *with signal constraint: geographical plot of the true simulated geoid errors (left) and the simulated and propagated geoid errors averaged over* λ *(right).*

Propagated and true geoid errors. The left panel of figure 5.14 shows the true errors for the solution with Tikhonov regularisation with signal constraint. The solutions with first and second derivative constraint yield similar geoid errors. Table 5.3 shows the statistics on the geoid errors, the maximum and minimum values are located in the polar regions. Evidently, the geoid errors show geographical correlation which is caused by the coloured noise. The right panel of figure 5.14 shows the geoid errors averaged over λ as well as geoid errors from error propagation. Compared to the propagated geoid error based on the conservative coloured noise model (figure 5.2), the simulated error is too small (figure 5.14, right). This can largely be explained by the more optimistic coloured noise actually put on the data. A test has been conducted with an error PSD equal to that of figure 5.2 but with the error at the white noise level below 2 cpr. The optimistic coloured noise model is more in agreement with the actual simulated coloured noise (Klees *et al.*, 1999). The propagated error for the optimistic coloured noise model has at least the right order of magnitude. In any case, this example demonstrates that: (i) a reliable noise model is needed; (ii) if such a noise model is available the expected quality is very well described by error propagation.

Table 5.3: *Global geoid errors of the* TR *solutions due to the coefficient differences* OSU91A *- solved, final solutions. Units are in m.*

		geoid error (m)		
no.	constraint	RMS	max	min
1	signal	0.10	1.1	-1.2
1a	1st deriv.	0.09	1.1	-1.3
1b	2nd deriv.	0.12	1.0	-2.0

Table 5.4: MSE *and bias of the* BE *solution.*

no.	constraint	α	MSE[a]	BNR	max(BSR)[b]
1	TR, signal constraint	2.4	1	0.2	0.2 (54.8)
1c	BE, first iteration	1.5×10^{15}	63.7	0.4	0.6 (23.1)
1c	BE, second iteration	1.5×10^{15}	58.4	0.3	4.2 (70.3)

[a] In proportion to mission 1.

[b] The maximum BSR based on the OSU91A degree-order variances. The maximum BSR between brackets has been computed using the solution coefficients.

5.4.2 Biased estimation

Ordinary biased estimation

The biased estimator (BE)

$$\hat{x}_\alpha = (A^T P A + \alpha I)^{-1} A^T P y^\varepsilon$$

differs from Tikhonov regularisation in that the regularisation matrix simply is a scaled unit matrix. The minimisation of the MSE gives the results presented in table 5.4. Biased estimation (BE) is a rather crude regularisation method. Compared to TR, BE over-regularises the low degrees, which gives large biases there (see below). The bias with respect to the noise is larger compared to TR, the MSE is larger as well.

Coefficient errors. The relative errors in the spherical harmonic coefficients are depicted in figure 5.15 for the first and second iteration steps. The high degree coefficients as well as the low order coefficients improve from the first to the second iteration. Compared to TR(0) the BE errors are larger. The BSR for the first and second BE iteration is shown in figure 5.16. The bias in the high degree low orders increases, which may be somewhat surprising. However, a more detailed picture of the low degrees and orders shows that the bias in those coefficients is redistributed over the high degrees, see figure 5.17. The bias will be discussed in more detail hereafter. The SNR of the low orders is low, many coefficients have a size below the error, see figure 5.18, left panel. The minimum contribution measure is 97% (figure 5.18, right), but, as explained in section 5.4.1, this fact has little additional value.

Bias. The true geoid height bias and the bias part of the MSEM propagated to geoid heights for the initial BE iteration are displayed in figure 5.19, left. Shown are the geoid errors averaged over λ. Again, the block-diagonal approximation provides reasonable bias estimates since the true and propagated errors are the same. Figure 5.19, right, compares the propagated bias and noise geoid errors. The bias is much larger than the noise in the area covered with observations, while it is smaller in the polar regions. Figure 5.20 shows the propagated bias and noise for the second BE iteration. The bias can be neglected compared to the noise in the measurement area, whereas both errors remain large in the polar regions.

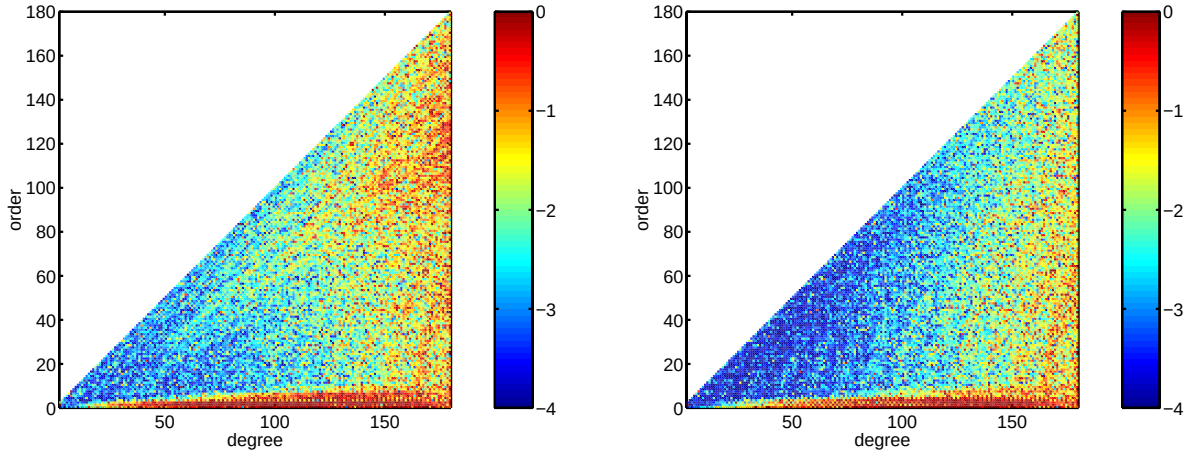

Figure 5.15: BE. *Relative error in the spherical harmonic coefficients, first iteration (left) and after two iterations (right). The differences (*OSU91A *- solved)/*OSU91A *are shown on a logarithmic scale.*

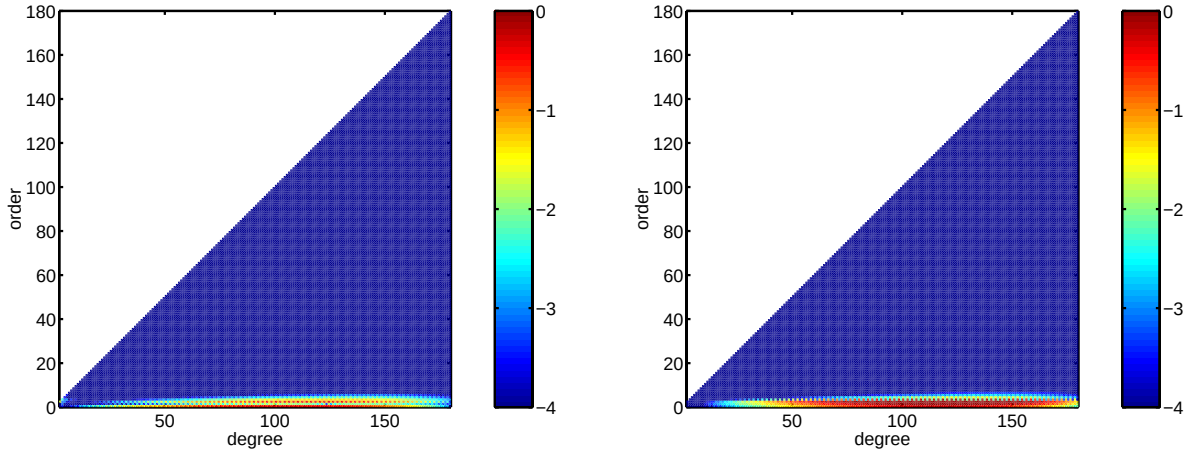

Figure 5.16: BSR *for* BE *first iteration (left) and second iteration (right), the scale is logarithmic. All coefficients are shown. The* BSR *was computed using the* OSU91A *degree-order variances.*

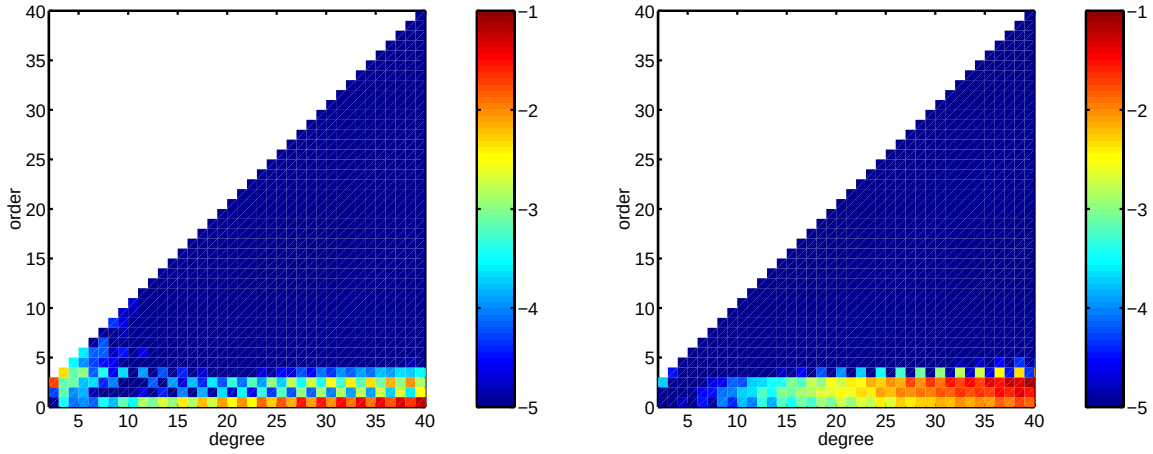

Figure 5.17: BSR *for* BE *first iteration (left) and second iteration (right), the scale is logarithmic. The coefficients up to* $l = 40$ *are shown. The* BSR *was computed using the* OSU91A *degree-order variances.*

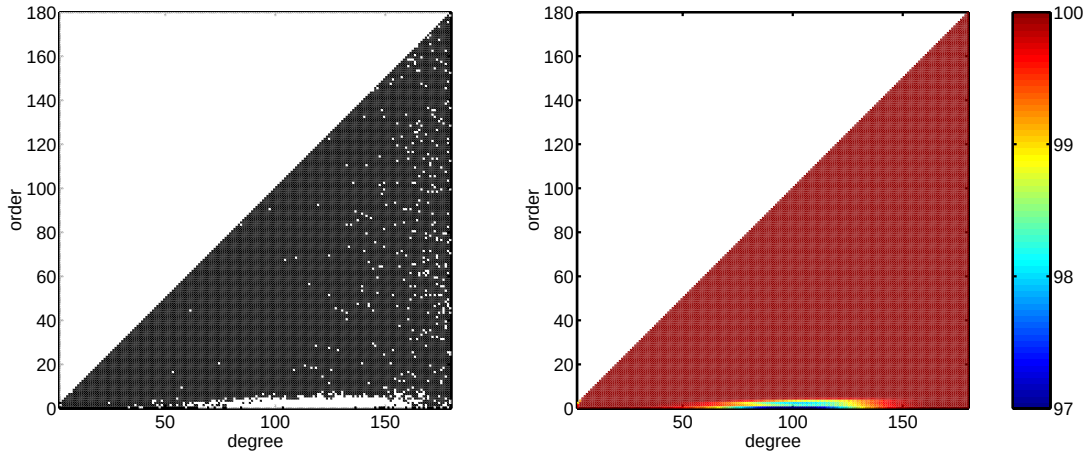

Figure 5.18: SNR *for the second* BE *iteration (left) and contribution measure for unbiased estimators,* BE *(right).*

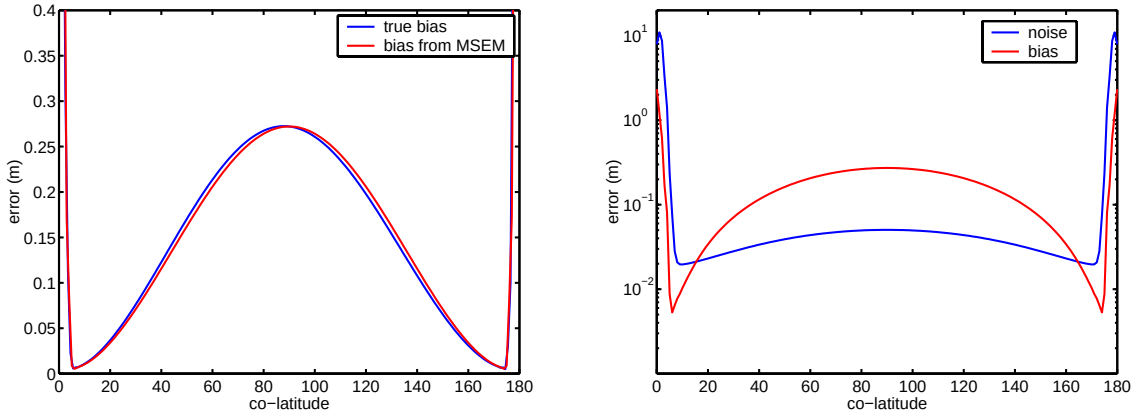

Figure 5.19: BE *first iteration: true geoid bias and block-diagonal part of the* MSEM *propagated to geoid heights (left) and propagated bias and noise on a logarithmic scale (right).*

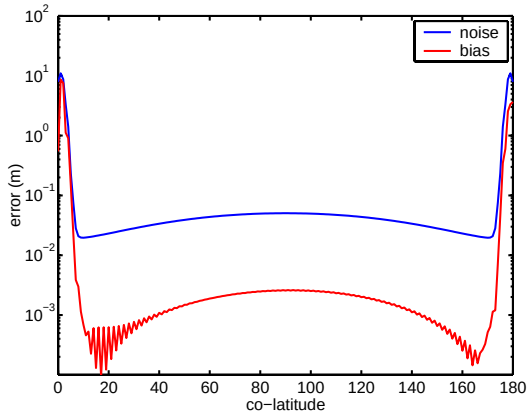

Figure 5.20: BE *second iteration: propagated bias and noise on a logarithmic scale.*

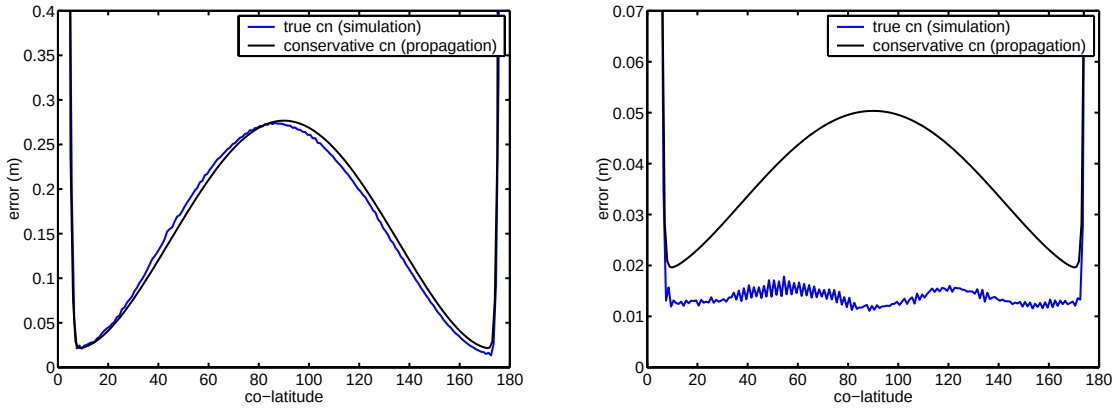

Figure 5.21: *The simulated and propagated geoid errors averaged over λ,* BE *first iteration (left) and* BE *second iteration (right).*

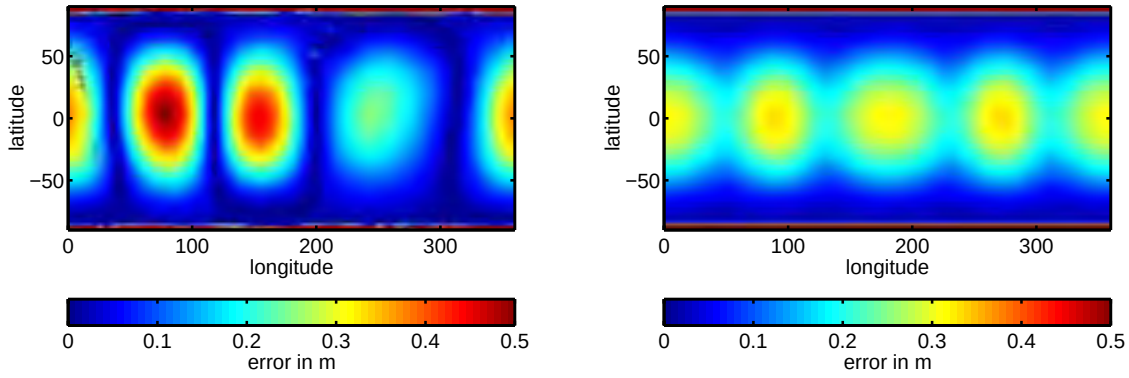

Figure 5.22: BE *first iteration: geographical plot of the absolute value of the true simulated geoid errors (left) and the propagated geoid errors (right).*

Propagated and true geoid errors. The total averaged true and propagated geoid errors are very much alike in the measurement area for the first BE iteration, which is in correspondence with the bias which is the main contributor there, figure 5.21. Comparing, however, a geographical plot of the true and propagated error there is certainly some resemblance, but discrepancies remain (figure 5.22). The geographical correlated error is due to the errors in the low degree tesseral coefficients and coefficients with $l - m$ small, which are affected most by the coloured noise, see also figure 5.17. The simulated geoid error is summarised in table 5.5.

Iteration. Because the bias exceeds the noise in the area covered by observations, one may expect that subsequent iterations reduce the bias in this area. The bias yields residuals $\hat{y} - A\hat{x}_\alpha$ which are larger than the noise. This erroneous signal may be overcome by iteration as is the case for model errors. Let the linear model be $y^\varepsilon = Ax + \epsilon$ with the first solution

$$\begin{aligned} \hat{x}_1 &= (A^T P A + \alpha K)^{-1} A^T P y^\varepsilon \\ &= (A^T P A + \alpha K)^{-1} A^T P (Ax + \epsilon). \end{aligned}$$

The difference between the 'true' solution x and $\hat{x}_1$ is

$$\begin{aligned} \Delta\hat{x}_1 := \hat{x}_1 - x &= (A^T P A + \alpha K)^{-1} A^T P A x - x + (A^T P A + \alpha K)^{-1} A^T P \epsilon \\ &= \Delta x + \eta_1 \end{aligned}$$

Table 5.5: *Global geoid errors of the* BE *solutions due to the coefficient differences* OSU91A *- solved. Units are in m.*

	geoid error (m)		
BE	RMS	max	min
1st iteration	1.92	43.2	-29.0
2nd iteration	1.48	18.6	-21.5

with the bias term

$$\Delta x = -(A^T P A + \alpha K)^{-1} \alpha K x$$

and the error term

$$\eta_1 = (A^T P A + \alpha K)^{-1} A^T P \epsilon.$$

If the error term η_1 is small with respect to the bias term (which is true in the measurement area for BE), then $\Delta\hat{x}_1 \approx \Delta x$. Putting $\hat{y}_1 = A\hat{x}_1$, the second solution is

$$\begin{aligned}
\hat{x}_2 &= \hat{x}_1 + (A^T P A + \alpha K)^{-1} A^T P (y^\varepsilon - \hat{y}_1) \\
&= x + \Delta\hat{x}_1 + (A^T P A + \alpha K)^{-1} A^T P (Ax + \epsilon - A(x + \Delta\hat{x}_1)) \\
&= x + \Delta\hat{x}_1 - (A^T P A + \alpha K)^{-1} A^T P A \Delta\hat{x}_1 + \eta_1 \\
&= x - (A^T P A + \alpha K)^{-1} \alpha K \Delta\hat{x}_1 + \eta_1.
\end{aligned}$$

If $\Delta\hat{x}_1 \approx \Delta x$ then the bias term is reduced in $\hat{x}_2$ compared to $\hat{x}_1$:

$$\Delta x_2 := E\{\hat{x}_2 - x\} = -(A^T P A + \alpha K)^{-1} \alpha K \Delta x$$

and the corresponding MSEM is

$$MSEM(\hat{x}_2) = Q_x + \Delta x_2 \Delta x_2^T. \tag{5.1}$$

The averaged geoid error is shown in figure 5.21 on the right. In the measurement area the result is equal to the TR results, whereas the error in the polar regions remains large.

Instead of keeping α fixed in the iteration, one could decide to determine α anew, say α_2, minimising the trace of (5.1). Then, the second solution is

$$\hat{x}_2 = x - (A^T P A + \alpha_2 K)^{-1} \alpha_2 K \Delta\hat{x}_1 + \eta_2$$

with $\eta_2 = (A^T P A + \alpha_2 K)^{-1} A^T P \epsilon$. However, to be consistent with TR α is kept fixed.

Generalised biased estimation

The simulated geoid errors for generalised biased estimation (GBE) are not shown, in the area with observations they are as shown in figure 5.21, right. The RMS geoid error is 0.42 m, with extremes of 4.2 m and -4.8 m. Although GBE is designed to give the minimum MSE, it does not in case of ill-posedness due to long wavelengths as is the case here. Bouman and Koop (1998a) argue that GBE is affected more by the polar gap compared to TR. To the small singular values

$$d_i = \langle x, v_i \rangle^{-2}$$

is added if $\sigma^2 = 1$, cf. section 2.3.3. However, the small singular values correspond for a polar gap also to low orders. Consequently, the product $\langle x, v_i \rangle$ becomes large since these low frequencies have high energy. The squared inverse therefore becomes small yielding hardly any stabilisation for the small singular values involved with the low orders. Indeed, excluding the low orders gave better results for GBE than for TR (Bouman and Koop, 1998a).

5.5 Summary

The polar gap affects the low orders, but in case of 'faultless' observations the exact solution is almost obtained. The circular approximation of the orbit can be overcome by iteration if the repeat is nearly exact and there are no data gaps. Of the tested regularisation methods, Tikhonov regularisation performs best in the presence of a polar gap. Ordinary and generalised biased estimation yield larger errors in the polar regions than TR. Of the TR solutions, the one with the signal constraint gives the smallest predicted MSE, however TR with first derivative or second derivative constraint yield comparable results. If the polar gaps are not taken into account then all regularisation methods perform equally well in terms of RMS geoid height errors.

A complete understanding of the quality of a satellite based gravity field solution is obtained by studying the potential coefficients themselves as well as their implications for the geographical distribution of e.g. geoid errors. In order not to draw too optimistic or pessimistic conclusions about the quality, the noise model must be reliable. Of the quality measures, the contribution measure is of little additional value compared to BSR and SNR. The bias in the geoid heights can be neglected in the measurement area and becomes large in the polar areas. Almost all low order coefficients, however, suffer from the bias and it cannot be neglected: the bias is up to 20-30% of the signal using TR, whereas it is up to 420% of the signal using BE. It is therefore concluded that the quality of the individual low order coefficients may be poor in presence of a polar gap, but that the lumped effect of the poor quality is limited to the unsurveyed areas. When the block-diagonal approach of the bias part in the MSEM is compared with the true bias then it can be concluded that the former approximation is good enough.