Institut für Geodäsie und Geoinformation der Universität Bonn
Professur für Astronomische, Physikalische und Mathematische Geodäsie

Optimization of point grids in regional satellite gravity analysis using a Bayesian approach

Dissertation

zur

Erlangung des akademischen Grades

Doktorin der Ingenieurwissenschaften (Dr.-Ing.)

der

Landwirtschaftlichen Fakultät

der

Rheinischen Friedrich-Wilhelms-Universität Bonn

vorgelegt von

Dipl.-Ing. Judith Schall

aus

Aachen

Bonn 2020

Referent: Prof. Dr. Jürgen Kusche
Korreferent: Prof. Dr. Hans-Peter Helfrich
Korreferent: Prof. Dr. Volker Michel

Tag der mündlichen Prüfung: 27.09.2019

Optimization of point grids in regional satellite gravity analysis using a Bayesian approach

Summary

The subject of this thesis is the global and regional gravity field determination from GOCE data using the short arc approach. The focus is on the extension of the regional method regarding an adaption of the model resolution to the data by estimating an optimal nodal point configuration for the arrangement of the radial basis functions. Estimating the positions of the basis functions is a nonlinear problem, which is not easy to solve with the means of classical adjustment theory. This is especially true if the number of basis functions is to be determined from the data as well. It is for this reason that the point grid has been fixed so far, and only the linear problem, that is the determination of the scaling coefficients for a given point grid, has been solved. Here, the problem is formulated within the framework of Bayesian statistics by specifying a joint posterior density for the number of the basis functions and the rest of the parameters. For the practical solution, the reversible jump Markov chain Monte Carlo sampling algorithm is employed, which allows simulating this kind of variable dimension problem. Key points in the implementation of the approach are the marginalization of the scaling coefficients from the target density, which enables me to limit the chain to the sampling of the point grid, and the use of a proposal distribution derived from a gravity field model. The final gravity field solution is taken to be the average of the generated gravity field solutions and thus takes into account the uncertainty about the choice of the model. The method is applied to real GOCE data and compared with the global spherical harmonic model ITG-Goce02 and a regional solution that makes use of a regular distribution of basis functions. Being a part of this work, the comparison models are based on the same processing strategy. It turns out that the optimization of the point grid enormously reduces the required number of basis functions, and that the distribution of the grid points becomes adapted to the structures of the gravity field signal. The solution becomes more stable and better reflects the characteristics of the signal. This entails an improvement of up to 13% over the mentioned comparison models.

Optimierung von Punktgittern in der regionalen Schwerefeldanalyse unter Verwendung eines Bayesschen Ansatzes

Zusammenfassung

Thema dieser Arbeit ist die globale und regionale Schwerefeldbestimmung aus GOCE Daten durch die Analyse kurzer Bahnbögen. Der Schwerpunkt liegt dabei auf der Weiterentwicklung der regionalen Methode hinsichtlich der Anpassung der Modellauflösung an die Daten durch Schätzung einer optimalen Punktkonfiguration für die Anordnung der radialen Basisfunktionen. Die Schätzung der Positionen der Basisfunktionen ist ein nicht-lineares Problem und mit den Mitteln der klassischen Ausgleichungsrechnung nicht einfach zu lösen. Dies gilt insbesondere dann, wenn auch die Anzahl an Basisfunktionen aus den Daten zu bestimmen ist. Aus diesem Grund wurde das Punktgitter bislang fixiert und nur das lineare Problem, die Bestimmung der Skalierungskoeffizienten bei gegebenem Punktgitter, gelöst. Hier wird die Aufgabe im Rahmen der Bayes Statistik formuliert und eine gemeinsame a posteriori Dichte für die Zahl der Basisfunktionen und die übrigen Parameter angesetzt. Die Lösung erfolgt über den reversible jump Markov chain Monte Carlo Sampling Algorithmus, der es erlaubt, Probleme dieser Art von variabler Dimension zu simulieren. Besonderheiten bei der Umsetzung des Verfahrens sind die Marginalisierung der Skalierungskoeffizienten aus der Zieldichte, die es ermöglicht, sich auf das Sampling des Punktgitters zu beschränken, und die Verwendung einer Vorschlagsverteilung abgeleitet aus einem Schwerefeldmodell. Die finale Schwerefeldlösung wird durch Mittelbildung aus den generierten Schwerefeldlösungen abgeleitet und berücksichtigt somit die Unsicherheit über die Wahl des Modells. Die Methode wird auf GOCE Echtdaten angewendet und mit dem globalen Kugelfunktionsmodell ITG-Goce02 und einer regionalen Lösung basierend auf einer gleichmäßigen Punktverteilung verglichen. Die Vergleichsmodelle sind Teil dieser Arbeit und verwenden dieselbe Prozessierungsstrategie. Es zeigt sich, dass die Optimierung des Punktgitters die Zahl der benötigten Basisfunktionen enorm reduziert und die Verteilung der Punkte sich an die Strukturen des Schwerefeldsignals anpasst. Die Lösung ist stabiler und spiegelt die Charakteristiken des Signals besser wider. Damit einher geht eine Verbesserung von bis zu 13% gegenüber den genannten Vergleichsmodellen.

Contents

1. Introduction

1.1 Motivation

The deviation of the physical shape of the Earth from a sphere, the irregular structure of the Earth's interior and its topography appear as spatial variations in the gravity field. Mass transports, mostly water redistribution, cause further temporal variations. The gravity field thus contains valuable information about the actual mass distribution of the Earth, making gravity observations interesting for many applications from the field of Earth sciences. In solid Earth geophysics, for example, the gravity field provides a constraint in crust and lithosphere modeling from seismic and other data. Knowledge of the geoid, which is an equilibrium surface of the gravity field, allows determining the mean dynamic ocean topography from altimeter observations. Classical geodetic applications are the unification of height systems over continental boundaries and the possibility to derive physical heights by referencing GPS-based height measurements to the geoid. Most of what is known about the global gravity field today has been discovered by means of the satellite missions CHAMP, GRACE and GOCE. These satellites were equipped with innovative sensor technology realizing the concepts of Satellite to Satellite Tracking and Satellite Gravity Gradiometry. The satellite observations are usually not used directly, but they are inverted into a three-dimensional model using adapted evaluation strategies. The operational analysis centers almost exclusively use spherical harmonics for representation. Even if today many groups work on alternative representations, they are not often used for applications. One reason might be that the dissemination as a data product is more difficult for regional models, which is due to the variety of methodologies and a lack of standards. Notwithstanding this, regional models offer a number of advantages compared to spherical harmonics (see also Jekeli, 2005):

Spherical harmonics are different from zero almost everywhere on the sphere. As a consequence, any observation is related to the entire spherical harmonic series expansion. By contrast, space-localizing basis functions concentrate their energy in a small portion of the sphere. A single observation can therefore be described by only a few basis functions, and only data in the area of influence of the basis functions are required to adjust them. This reduces the computational burden and thereby enables to rigorously compute regionally high-resolution gravity field models; it facilitates the work on local data sets and helps to handle gaps in the data.

Moreover, spherical harmonics realize a globally uniform resolution. But the resolution that can actually be obtained from the data depends on down to which scale the signal is still sufficiently large compared to the noise, which is not uniform. On the one hand, the gravity signal is not equally smooth in every geographical region. This is especially true for the higher frequencies, which the GOCE mission is specifically sensitive to. On the other hand, the noise is not equally high everywhere either, because the satellite altitude varies along the eccentric orbit and with it the amount of amplification of the measurement noise in the downward continuation process. Also, because of the convergence of the satellite tracks, there are comparatively many observations in the polar regions, leading to more precise estimates, so that even smaller signals can still be resolved. The two effects (i.e. the influence of eccentricity and convergence) can also be observed in the maps of propagated geoid errors of GOCE gravity field models (see Bingham et al., 2011). To sum up, for spherical harmonics the model resolution is uniform, but the achievable resolution is not. Accordingly, when the maximum degree of the spherical harmonic expansion is chosen high enough to capture local signal in areas of rough topography (e.g. in mountain areas or in the area of a deep sea trench), this will lead to overfitting of noise in the smoother oceanic regions. To reduce the noise and to stabilize the solution, one can use regularization. In the context of global gravity field determination, mostly Kaula regularization is used. Here the coefficients are constraint towards zero

according to the global average signal-to-noise ratio. In this way, the local signal will finally get lost, while over the oceans a certain degree of noise might remain. By contrast, in regional analysis the model resolution can be chosen suitable for the local study area, and the prior information can be adapted to some extent to the local signal by estimating a variance factor from the data restricted to this region, leading to an improvement in the regularization.

The feasibility to adapt the resolution of a regional model is frequently mentioned to motivate regional modeling, as was also done here. However, most of the approaches for regional gravity field analysis, such as the one implemented in Bonn, do not (fully) take advantage of this yet. The reason is that the model resolution is closely connected to the choice of the point grid for the arrangement of the nodal points of the basis functions. Adapting the point grid to the data in terms of a formal optimization is not a trivial task because this problem is nonlinear and variable in dimension. Most of the time, the point grid is therefore simply defined in advance. For example, in Bonn we make use of a dense and uniform grid with an additional margin to avoid edge effects. Even though this choice may enable us to adapt the model resolution better to the local conditions in a particular region than would be possible with spherical harmonics, we thereby disregard possible variations within the region with all the before mentioned disadvantages. Furthermore, the huge number of basis functions, in particular in the edge region, represents an overparameterization of the problem and leads to instabilities. In the current state, we cannot even solve a problem without regularization, also when working on terrestrial data, where is no problem for downward continuation. By an adaption of the model resolution, which is certainly connected with a reduction of the number of basis functions, the stability will hopefully improve, so that the solution is less dependent on the prior information.

1.2 State of research

When estimating the scaling coefficients of the basis functions in regional gravity field analysis, the linear relationship leads to a quadratic and hence convex objective function with only one overall (global) minimum. However, the problem of estimating the locations of the basis functions jointly with the scaling coefficients is nonlinear. For a nonlinear problem, the objective function is in general not convex (Boyd and Vandenberghe, 2004); it is rather a mountainous landscape with several (local) minima. When choosing the estimation technique, one should have in mind that the widely used method of least squares from classical adjustment theory involves linearization. It is thus only advisable if sufficiently good approximate values are available, because otherwise the gradient at the point of linearization might be misleading, so that the algorithm would run in the wrong direction and into a local minimum. For the locations of the basis functions, there is no general rule on how to derive adequate approximate values. Even if the concept of point masses, which are often introduced via the discretization of Newton's law of gravitation, might suggest that one can deduce them from prior knowledge of the mass distribution, this knowledge is rather limited. In summary, finding approximate values for our problem is at least not obvious, and the use of local optimization techniques therefore not sensible. Furthermore, with a better arrangement an equally good data fit can certainly be achieved with fewer basis functions. But the right number is not known in advance. We thus find ourselves in the unusual situation that the number of things we don't know is one of the things we don't know, to use the words of Green and Hastie (2009). Although one could use e.g. hypothesis tests to decide whether an individual basis function is necessary or not, they are not suitable for the simultaneous estimation of the number of model parameters and the parameters themselves. In summary, the presented problem is difficult to solve with the tools of classical statistics, and different approaches deal with these difficulties in different ways:

The first and by far the largest group of approaches define the point grid prior to the actual analysis and solve for the linear scaling coefficients only. An early example is Balmino, who proposed in 1972 to arrange point masses according to the extreme values of a map of gravity anomalies at various depths. He regarded this as superior to a uniform distribution. The Federal Agency for Cartography and Geodesy (in German: Bundesamt für Kartographie und Geodäsie, BKG), the solution of which enters the official geoid model for Germany, employs point masses at different heights on uniform grids of different resolution (Liebsch et al., 2006). In Bonn, we make use of radial basis functions, which similar to spherical splines incorporate prior information on the smoothness of the gravity field. This representation was developed by Eicker (2008), who also compared different grids on their regularity using criteria such as the size and shape of the area elements or the distance of the grid points. She found that the Reuter grid and the triangular vertex grid do best. Notwithstanding, she decided on the triangular grid for numerical experiments, as it has nearly spherical area elements, which matches well with the radial symmetry of the basis functions. The resolution of the grid was chosen so that the global number of the grid points corresponds to at least the number of spherical harmonic coefficients that would be considered appropriate for the actual scenario. The triangular vertex grid will be used for comparison purposes also in this work. Further, Bentel (2013) and Naeimi (2013) both published a comprehensive study of radial basis functions, which among other aspects of regional modeling also includes the choice of the point grid. They came to the conclusion that the actual point grid only plays a minor role as long as it is uniform.

Wavelet functions, another sort of radial basis functions, concentrate in a specific frequency band; they are thus well suited for multiresolution techniques. Wavelets are usually set on a uniform grid. The resolution of the grid varies within the approach depending on the resolution of the wavelets. For example, Schmidt et al. (2008) made use of Blackman scaling and wavelet functions on a Reuter grid. Poisson wavelets, which can be transferred to multipoles and are therefore also known as Poisson multipole wavelets, were used by Holschneider et al. (2003), Chambodut et al. (2005) or Panet et al. (2011). Their grid choice is based on hierarchically subdividing a cube or icosahedron and then projecting the corresponding points onto the sphere. Apparently, efforts have also been made to adapt the shape and positions of the wavelets (or the tree-dimensional positions of the multipoles) to the signal using an iterative approach (c.f. Hayn et al., 2013). Least squares collocation requires no basis functions actually, but under certain conditions it can be reduced to a parametric approach with radial basis functions, the locations of which are specified by the locations of the observations (see Barthelmes, 1989, and references there). The same type of arrangement was also chosen by other approaches (e.g. Dampney, 1969).

Up to here, all the aforementioned approaches were based on radial basis functions. But also for other representations, such as mascons, one has to specify the arrangement of the basis functions. A mascon, short for mass concentration, is a layer of mass put on the surface of the sphere in a particular region. The gravitation-generating mass is determined by the thickness of the layer and stated in terms of equivalent water height. The shape of a mascon is variable; it can refer to an arbitrary geographical region, such as a drainage basin (c.f. Luthcke et al., 2006). However, more often regularly shaped mascons are applied. For example, the NASA GSFC GRACE solution is based on equal-area rectangular mascons with a uniform arrangement (Luthcke et al., 2013; Rowlands et al., 2010). The NASA Jet Propulsion Laboratory makes use of spherical cap mascons being either individual in size and position (e.g. tied to individual glaciers, see Ivins et al., 2011) or equally large and globally uniformly distributed (Watkins et al., 2015).

The second group of approaches build the point grid in a stepwise manner and in each step select the grid point that is particularly useful in the sense of reducing the objective function value. Thus, the arrangement of the basis functions is adapted to the data. However, once a grid point has been introduced, it is not changed anymore during the course of the algorithm. This can therefore not lead to the optimal solution. The majority of the approaches considers the point of the largest absolute

value of the data to be useful and iterates the procedure until either a certain number of iterations or a stop criterion, such as the desired approximation accuracy, is reached. This approach was, for example, also taken by Cordell (1992) and Antunes et al. (2003) for the positioning of point masses; the remaining model parameters were set by means of simple empirical rules. Another example is Marchenko and Abrikosov (1995), who worked with radial multipoles, which among others also incorporate the concept of point masses. After having fixed the horizontal position of a new basis function as has just been described, they estimated the multipole parameters degree and depth, which together define the shape, and the moment, which corresponds to the scaling coefficient, best adapting in comparison with a local empirical covariance function. Finally, all multipole moments were once again estimated in a total adjustment. Marchenko et al. (2001) employed the approach for geoid determination from airborne data in the Skagerrak and summarized that the multipole analysis yields a similar accuracy as collocation with only 10-20% of the basis functions.

At TU Delft the so-called data-adaptive network design (DAND) was developed (Klees and Wittwer, 2007). It can be realized by removing all points from an initial grid with too few or too little observations around. This type of grid was used for the positioning of multipole wavelets in the analysis of airborne data. As a result, the number of basis functions was approximately 30% of the observations. The equation system had a better condition, making regularization sometimes superfluous. And the solution was described as qualitatively better compared to using standard networks. Moreover, in 2008 Klees et al. proposed to further refine the result of a standard regional analysis by sequentially introducing new basis functions at the position of the highest residual. But in contrast to the two approaches mentioned before, Klees made further restrictions: the residual must be large enough, the number of observations in the neighborhood not too small and the next basis function not too far. The shape or bandwidth of the new function was adjusted by means of general cross validation, which is a tool for model comparison based on the leave-one-out principle; the scaling coefficient was estimated on the basis of the surrounding observations. After all data points had been processed, all coefficients were estimated again in a joint adjustment and on the basis of all observations. Klees reported that this kind of data-adaptive positioning is more efficient than a real optimization but at the price of a higher number of basis functions. However, he expected problems to arise if the number of data points goes to infinity, since the number of basis functions is determined by the number of observations and, apart from that, by a series of threshold values. Both approaches were used in combination by Wittwer (2009).

A fundamentally other approach, the regularized functional matching pursuit, was proposed by Fischer and Michel (2012) (see also Fischer, 2011). As part of this approach, a dictionary with global and different local basis functions is defined along with a point grid of possible positions. A basis function is then selected from the dictionary and so positioned that after the related scaling coefficient has been estimated, the residual sum of squares becomes minimal. Fischer and Michel demonstrated the advantage of this approach over wavelets and splines when applied to heterogeneous data or big amounts of data. Their results suggest that the resulting distribution of points reflects the continental boundaries and further topographic structures. A drawback is that after a new basis function has been introduced, the set of scaling coefficients is no longer optimal. According to Michel and Telschow (2016), this leads to the algorithm selecting the same basis functions several times to improve the corresponding coefficients. Therefore, they proposed an enhancement of the method including an adjustment of all coefficients in every step, which was implemented by a sophisticated orthogonal projection. This led to even fewer basis functions and thus to a sparser solution than what had been obtained without transformation. In gravity field analysis from satellite data with heterogeneous distribution, equally good results were achieved but with much fewer basis functions compared to an approach with splines under the data points.

And finally, the third and last group of approaches aim at real optimization of the point grid. An early example is the concept of free-positioned point masses proposed by Barthelmes (1986) (see

also Barthelmes, 1989; Barthelmes et al., 1991). He introduced new basis functions iteratively at the place of the highest remaining signal because this was shown to be the optimal position if the basis functions were orthogonal. Since they are not, a nonlinear least squares adjustment is performed to improve the position together with those of the other point masses, which are also not optimal anymore once the new point mass has been added. According to Barthelmes, they are good enough to be used as starting values, though. Because of this approximation, apart from some others to reduce the computing time, the method might not definitely yield the globally optimal solution. Anyway, the accuracy can be improved at any time just by adding further basis functions, as was noted by Barthelmes. Indeed, the number of functions in this approach is determined by the specification of the desired approximation accuracy. The method was used for the approximation of boundary values calculated from the long-wave part of a spherical harmonic model; in comparison to a uniform distribution, the same accuracy was reached much earlier, i.e. with fewer basis functions.

Since then, the approach of Barthelmes has been replicated several times (e.g. Lehmann, 1993; Claessens et al., 2001). Recently, it has been picked up again by M. Lin (Lin et al., 2014; Lin, 2016). In his version, a quasi-Newton method was used for the nonlinear optimization, which promises a better convergence by the use of second order derivatives and further allows imposing boundary constraints on the parameters. In addition to the specification of a stop criterion, there are thus a number of other settings to be made. Lin tested his approach intensively on synthetic and real data. He compared, for example, least squares collocation to point masses arranged according to a geographical grid and to the free-positioned point masses and found that the latter yield similarly good results as collocation and even better results than the regular distribution, while using fewer basis functions. Like already Barthelmes et al. (1991) before him, he demonstrated that the resultant distribution follows the structures of the gravity field; in other words, there are more functions where the gravity field is rough. He reported that an optimization of the horizontal position and depth yields better results than a purely radial optimization but also evokes problems in the presence of data gaps. The basis functions move to the center of the gaps or, alternatively, towards the border of the investigated area to minimize the data misfit. Like Barthelmes (1989), Lin therefore preferred the radial optimization, which also in this situation led to reasonable results.

Yet another example for the use of optimization techniques in the context of the given problem is Antoni (2012), who tested both global and local optimization strategies. For instance, he applied genetic algorithms, which are a class of algorithms that utilize the concepts of the theory of evolution, such as natural selection or random mutation, to find their way to the global optimum. However, the number of basis functions has to be set before the start of the algorithm. Therefore, and also because he could not exactly reproduce the results of the random algorithm, Antoni preferred a local optimization approach. Again, the basis for the approach was the approach of Barthelmes. But in contrast to him, Antoni made use of bandlimited radial basis functions, the horizontal position and shape of which were estimated; he introduced several basis functions at the same time and applied the trust region algorithm as a powerful alternative to the least squares optimization technique. The starting values were specified as the maxima of the observational data, which were determined by procedures of image processing. The decision about the number of basis functions was driven by a number of threshold values, whereas basis functions that were poorly determined were removed at the end of the algorithm. Like in the present work, the applicability of the approach was tested in an easy simulation scenario with only few basis functions. The algorithm produced a good reconstruction of the signal but with considerably more basis functions. Also for a realistic scenario, Antoni reported good reconstruction results, without however comparing them with the results of a standard approach.

1.3 Thesis objectives and the scientific context

The aim of this thesis is to develop, implement and apply an approach to optimize the point grid in a radial basis function framework, i.e. the number and locations of the basis functions, together with the usual model parameters, namely the scaling coefficients and a regularization parameter. The approach is tailored to the determination of the gravity field from satellite observations, but in principle it can be used for the approximation of any data set with some modifications.

As said earlier, we face a nonlinear problem where no approximate values are available, which makes the use of local optimization strategies difficult. To overcome these difficulties, global optimization is applied, where the success is not dependent on the availability of good approximate values. The optimization of the point grid also involves the number of basis functions, which is inconvenient in that the number of parameters itself is actually a parameter. Therefore, I apply a method to model selection from the field of Bayesian statistics, which enables me to estimate the number of parameters in analogy to an ordinary parameter from the data. All the other approaches mentioned in the previous section either define the number of basis functions in advance or stop the algorithm when a desired approximation accuracy is reached or certain threshold values are exceeded. Apart from the choice of the stop criterion, many of the mentioned approaches have to make a number of further settings and define threshold values. In the present approach, only few specifications are required, which mostly influence the convergence of the procedure rather than the results. However, as usual in the context of Bayesian statistics, I have to define prior densities, and although I made efforts to not introduce subjective prior knowledge by the choice of non-informative densities, complete ignorance cannot be expressed with a prior density (Robert, 2007, p. 127). Moreover, the local optimization techniques of the third group all provide also an estimate for the covariance matrix of the positions of the basis functions. They all involve simplifications such that they linearize the problem, and a description of the uncertainty of nonlinear parameters only by variances and covariances is at least not complete (Barthelmes et al., 1991). I use a random sampling algorithm, which does not only yield estimates for the parameters, as it is the case e.g. for evolutionary or genetic algorithms, but theoretically the entire probability distribution and with it also a complete error description. In summary, the approach pursued here differs from existing work in that (1) it allows finding the global optimal solution including the number of basis functions, (2) it yields accuracy information for the estimated parameters in the form of the sampled posterior distribution, and (3) it does not require many specifications.

The optimization only refers to the horizontal positions of the basis functions. The vertical position or depth of a basis function can better be associated with the shape of the function, and an optimization of the shape is not part of this work. In the literature, there are different opinions to whether this is reasonable, ranging from believing that shape coefficients and positions are independent (Carlson and Foley, 1991) to recommending that the two aspects, since they are interrelated, should be considered simultaneously (Klees and Wittwer, 2007). I would intuitively assume that the shape of the basis functions on the one side and their number and positions on the other side are strongly correlated, and fixing the shape will greatly simplify the optimization of the point grid. Therefore, in this work, the parameters determining the shape are specified beforehand in such a manner that the basis functions are well adapted to the gravity observations to be approximated. In other words, just one type of kernel function is employed for the moment. An additional optimization of the shape is envisaged at a later stage.

The gravity field models generated in the course of this thesis are based on the short arc approach. This approach was developed and implemented by T. Mayer-Gürr to process the data of the satellite missions CHAMP and GRACE (Mayer-Gürr et al., 2005; Mayer-Gürr, 2006). Shortly after, A. Eicker introduced the alternative representation by means of space-localizing basis functions and embedded it into the programming system (Eicker, 2008). The present work is in some sense a continuation of

their work. Here, I computed a global GOCE model to prove that, with the same data, I can produce a model that is comparable to others. This global model also serves as comparison model for the results of the regional analysis, which basically relies on the same processing strategy. To this end, the software had to be adapted in some points, e.g. for the pre-processing of the GOCE observations and the estimation of the covariance function. My main contribution is the refinement of the regional method with respect to the optimization of the point grid, for which the methodological framework is provided in this thesis. The formulation of the approach allowed integrating the regional analysis with fixed point grid, while the new parts related to the optimization of the point grid have just been built around. However, some fundamental changes in the original software were also necessary in order to augment the efficiency of the procedure, for example in setting up the normal equations. The most important publications resulting from this work are Schall et al. (2014) for the global gravity field determination from GOCE data and Eicker et al. (2014) for the regional gravity field determination from GOCE data.

This thesis is built along the following lines. In Ch. 2, I will provide the theoretical background for global and regional gravity field determination from GOCE data by means of the short arc approach. In particular, Sec. 2.4.5 summarizes the explanation for the choice of the regional basis functions, which was the key point in the work of A. Eicker. Moreover in Sec. 2.5.2, the determination of a covariance function for the stochastic modeling of the GOCE gradients is described, which makes use of the Fourier transform to guarantee positiv definiteness.

In Ch. 3, the Bayesian framework is presented, in which the new approach for the optimization of the point grid is embedded. The first three sections introduce necessary basic terms and explain the Bayesian way of parameter estimation (the Bayes inference) for an analytically tractable problem. Sec. 3.4 presents random sampling algorithms, which can be applied when an analytical solution is not available. Particularly, Sec. 3.4.5 derives the Metropolis-Hastings-Green algorithm (also reversible jump Markov chain Monte Carlo) implemented in this thesis. The following section 3.5 defines the probability distributions that are used in the present approach as prior or proposal distributions. Finally, Sec. 3.6 discusses several points which I think are important for the understanding and the implementation of the procedure.

Ch. 4 describes the new approach for the optimization of the point grid and illustrates the main points by means of an easy simulation example. The sections 4.3 and 4.4 specify the two main ingredients for the implementation of the reversible jump algorithm, which are the posterior density and the move types. One of the key points of the present implementation, the marginalization of the scaling coefficients from the target density, is formulated in Sec. 4.3.1. Sec. 4.7 presents the estimators tested in this thesis. And finally, Sec. 4.9 shows how to extend the approach for the simulation of the variance factor, which is important for real data applications.

Ch. 5 presents the results of the global analysis (Sec. 5.1), the results of the regional analysis with the standard regular grid (Sec. 5.2), and the results of the optimization of the point grid (Sec. 5.3). In the latter section, the convergence and mixing behavior of the simulated chains is considered, and the question is answered where the chain converges. Moreover, it is tested how a change in the shape of the basis functions affects the algorithm. Sec. 5.3.8 presents the gravity field models resulting from the optimization of the point grid, and Sec. 5.3.9 demonstrates how the stability of the normal equation system changes during the run of the Markov chain. In the final discussion part, the outcomes of the different variants (global, regional with fixed grid, regional with optimized grid) are compared to each other and interpreted.

The final chapter 6 includes concluding remarks and ideas for future work.

2. Global and regional gravity field analysis from GOCE data

2.1 The gravity field and its functionals

According to Newton, the gravitational potential caused by a solid body of volume V and evaluated at point $\boldsymbol{r}$ is

$$V(\boldsymbol{r}) = G \iiint_V \frac{\rho(\boldsymbol{r}_Q)}{|\boldsymbol{r} - \boldsymbol{r}_Q|} dV \tag{2.1}$$

with G being the gravitational constant and $\boldsymbol{r}_Q$ a vector pointing to the volume element dV of density $\rho(\boldsymbol{r}_Q)$. Outside the attracting masses, the gravitational potential can be shown to satisfy the Laplace equation:

$$\Delta V = \frac{\partial^2 V}{\partial x^2} + \frac{\partial^2 V}{\partial y^2} + \frac{\partial^2 V}{\partial z^2} = 0. \tag{2.2}$$

Furthermore, it vanishes when going towards infinity and therefore fulfills the conditions imposed on an harmonic function. Every harmonic function is also analytic, which means that it is continuous and has continuous derivatives of any order (e.g. Hofmann-Wellenhof and Moritz, 2006, p. 8).

An observer on Earth is influenced not only by the gravitational potential but also by the centrifugal potential Φ, which add up to the gravity potential W:

$$W = V + \Phi. \tag{2.3}$$

The geoid is defined as the equipotential surface of the gravity potential with potential value W_0. The gravity potential is usually split up into a normal part U and a disturbing part T:

$$W = U + T. \tag{2.4}$$

As the normal potential also includes a centrifugal part, it cancels in the difference, which results in the disturbing potential also being an harmonic function. The equipotential surface of the normal field with the same potential as the geoid, i.e. $U_0 = W_0$, is used as geometric reference ellipsoid. The search for the geoid is now looking for all points which have the same potential as the level ellipsoid, i.e.

$$W(N, \lambda, \varphi) = U(0, \varphi) = U_0 \tag{2.5}$$

(Barthelmes, 2013). An approximation in terms of the disturbing potential known as Bruns' formula is

$$N = \frac{T}{\gamma}. \tag{2.6}$$

The gravity anomaly vector is defined as

$$\boldsymbol{\delta g} = \nabla W - \nabla U. \tag{2.7}$$

The gravity anomaly is derived by

$$\delta g = |\nabla W| - |\nabla U|. \tag{2.8}$$

Note that this is the classical definition of gravity anomalies. Here again, an approximation in terms of the disturbing potential is usual, bringing us to the fundamental formula of geodesy:

$$\delta g = \frac{\partial T}{\partial r} - \frac{2}{r}T. \tag{2.9}$$

The difference in direction is represented by the deflection of the vertical:

$$dN = -\varepsilon\, ds \tag{2.10}$$

$$\varepsilon = -\frac{dN}{ds}. \tag{2.11}$$

Division into a north-south and an east-west component leads to

$$\xi = -\frac{dN}{dx} = -\frac{1}{R}\frac{dN}{d\varphi} \tag{2.12}$$

$$\eta = -\frac{dN}{dy} = -\frac{1}{R\cos\varphi}\frac{dN}{d\lambda}. \tag{2.13}$$

2.2 Overview of the satellite mission GOCE

The satellite mission GOCE (Gravity field and steady-state Ocean Circulation Explorer; Rummel et al., 2011; Drinkwater et al., 2007) was launched on March 17th, 2009 and completed its mission at the end of 2013. It was primarily designed to map the time-mean part of the gravitational field of the Earth. It is thus complementary to the predecessor mission GRACE, which mainly aimed at the temporal variations. ESA realized GOCE as first Earth Explorer within their Living Planet program. The series of Earth Explorers consists of small and specialized missions developed in close cooperation with the scientific community. GOCE was expected to see short wavelengths with high accuracy or, in other words, to achieve high spatial resolution. The mission objectives were $1-2$ cm in terms of cumulative error in geoid heights and $1-2$ mGal in terms of cumulative error in gravity anomalies, both with respect to a spatial resolution of 100 km half wavelength. Today, after mission completion, global mean accuracies of 2.4 cm and 0.7 mGal could be achieved respectively for geoid and gravity anomalies (Brockmann et al., 2014). Applications for GOCE data can be found in various disciplines. For example, GOCE gravity field models are used to derive a high resolution model of the geodetic ocean topography when combined with altimetry data. An other example is the improvement in the realization of national height references surfaces as needed for using GNSS for height determination in the ordnance survey. The ability of GOCE to see the small features in the gravity field can be explained by the innovative measuring concept, which consists of measuring the second derivatives of the gravity potential, making the signal rougher and details better visible. Another reason is the orbit height of approximately 250 km, which is very low when compared with other missions, such as GRACE with 500 km. The attenuation of the gravity signal, which increases with the orbit height, is therefore less. To realize the concept of gravity gradiometry, a gradiometer was used for the first time in orbit. To prevent the satellite from sinking and to enable accurate gradients, drag had to be continuously compensated. To cover the high power requirement of especially the propulsion system, the satellite was set into a sun synchronous orbit, which contradicts a polar orbit and is the reason for the so called polar gap in GOCE data. GOCE was further equipped with a GPS system, which provided the precise positions for geolocating the gravity gradients and for determining the long-wave part of the gravity field. The star cameras provided important information for the inertial orientation of the gradients.

2.3 The short-arc approach

In the early days of satellite geodesy, there were only few—mainly optical and laser ranging—observations available. As the number of simultaneous laser ranging observations was not sufficient to determine the three-dimensional position of the satellite, the position was represented by the force function, and only few parameters were estimated to improve the force field. Thus, orbit determination and gravity field estimation were rather close. As the orbit was a-priori not known, the orbit had to be integrated, which is summarized as the differential orbit improvement method. In view of the sparse observations along the orbit, long arcs of days to weeks had to be used to improve redundancy (Ilk et al., 2008).

With the advent of the new satellite missions CHAMP, GRACE and GOCE, which were among others equipped with a GPS device, the data situation has fundamentally improved. Today, the orbit can be determined pointwise independent of the gravity field. As the positions are now comparatively well-known a-priori, they can be introduced into the force function, and the observation equations can be set up. Thus, besides the classical technique, a bundle of alternative methods became possible, which mainly differ in the number of differentiations of the satellite position vector or respectively in the number of integrations of the equation of motion (see Löcher, 2010 for a systematic overview). The integral equation approach, also referred to as *short arc approach*, has been one of these new approaches. It is based on the solution of Newton's equation of motion as boundary value problem, which takes the form of an integral equation of Fredholm type. It was Schneider who first proposed this solution for the purposes of satellite geodesy (Schneider, 1968). The original approach was substantially modified (cf. Mayer-Gürr, 2006) and has been frequently applied in the Astronomical, Physical and Mathematical Geodesy group of Bonn University: Mayer-Gürr et al. (2005) used this approach for the calculation of gravity field models from CHAMP data. In Mayer-Gürr (2006) it was adapted to be applied for GRACE data analysis. It was applied for the calculation of the gravity field model ITG-Grace2010. Finally, it was used for many simulations to plan for future missions, documented in Elsaka (2010).

Furthermore, the new missions are equipped with sensors to measure the gravity field in-situ. This is obvious for GOCE because the gradiometer measures pointwise the second derivatives of the gravity potential. Thus also for new types of observation, there is no need for long arcs, as the gravity field is measured directly (Ilk et al., 2008). In case of GOCE, the use of short arcs is one of several approaches to make the huge problem manageable in the first place: with a 1 sec data sampling, GOCE gathers 30 million observations per tensor component and year. Because of the characteristics of the measuring device, the gravity gradients are highly correlated, which would be reflected in a full covariance matrix. Knowing that GOCE enables to recover the gravity field with high spatial resolution, e.g. up to degree and order 250, that would result in 63,000 parameters. This situation leads to a design matrix of 2 TByte, a covariance matrix of 3,000 TByte and a system of normal equations with a size of 30 GByte. Even if those matrices would fit into the main storage of a computer, already from a computational point of view, it would be sensible to apply certain algorithms to reduce the problem. In the short arc approach, the observation equations are set up per arc, and the normal equations are calculated per arc and subsequently accumulated. This procedure can be nicely parallelized. When using the addition theorem of normal equations, it is implicitly assumed that the observations of different arcs are uncorrelated. As for GOCE, there are correlations with periods that exceed the length of a short arc; thus, when applying the short arc approach for the analysis of GOCE data, one obviously neglects this in the stochastic model. This is mitigated by introducing additional parameters into the deterministic model. Short arcs for the processing of gradiometer observations has been used by Eicker in several simulations. Schall et al. (2011) reported the first application to real data, and Schall et al. (2014) presented a complete GOCE model calculated using the short arc approach with the SST part following the integral equation approach. Eicker et al. (2014) compared the global model to results from regional analysis.

Generally, independent of the type of observation, the short arc approach has some properties that are worth mentioning: it allows to make use of a full covariance matrix per orbit arc and offers a nice possibility to handle outliers through arcwise weighting. Also, the short arc approach, which is obviously capable to deal with short pieces of the satellite orbit, is suitable to be applied in the frame of regional gravity field analysis. Here, one would cut the satellite observations to the region of interest, as only these data are necessary to adjust the scaling coefficients of the regional basis functions.

2.4 Representation by global or local basis functions

In the following, we work on the two-dimensional unit sphere Ω, which is the set of all points in three-dimensional Euclidean space that have a distance of one from the origin. A point on Ω is specified by its Cartesian coordinates $\boldsymbol{x}$, where $\boldsymbol{x} \in \mathbb{R}^3$ and $|\boldsymbol{x}| = 1$. One can alternatively use spherical polar coordinates, (λ, ϑ), where $\lambda \in [0, 2\pi]$, $\vartheta \in [0, \pi]$. On the contrary, an arbitrary point of the three-dimensional Euclidean space is denoted by $\boldsymbol{r}$ or (λ, ϑ, r).

2.4.1 Spherical harmonics

Any function F on the unit sphere can be expanded into a series of surface spherical harmonics:

$$F(\boldsymbol{x}) = \sum_{n=0}^{\infty} \sum_{m=-n}^{n} f_{nm} Y_{nm}(\boldsymbol{x}), \tag{2.14}$$

where the f_{nm} denote the spherical harmonic coefficients, and the Y_{nm} denote the (surface) spherical harmonics of degree n and order m. Herein,

$$Y_{nm}(\lambda, \vartheta) = \begin{cases} P_{nm}(\cos\vartheta)\cos m\lambda & \text{for } m \geq 0 \\ P_{n|m|}(\cos\vartheta)\sin|m|\lambda & \text{for } m < 0 \end{cases} \tag{2.15}$$

with the P_{nm} being fully (4π-) normalized associated Legendre functions of the first kind. Spherical harmonics form a set of orthogonal base functions. The orthogonality relations fulfilled by fully normalized spherical harmonics read

$$\int_{\Omega} Y_{nm} Y_{n'm'} d\Omega = 4\pi\, \delta_{nn'} \delta_{mm'}. \tag{2.16}$$

The addition theorem,

$$\frac{1}{\sqrt{2n+1}} \sum_{m=-n}^{n} Y_{nm}(\lambda, \vartheta) Y_{nm}(\lambda', \vartheta') = P_n(\cos\psi), \tag{2.17}$$

relates spherical harmonics and Legendre polynomials, the latter being the Legendre functions of order zero.

For geodetic applications, (solid) spherical harmonics are used, which are the solutions of the Laplace equation and therefore harmonic. In combination, they can represent any harmonic function, e.g. also the gravitational potential:

$$V(\lambda, \vartheta, r) = \sum_{n=0}^{\infty} \frac{1}{r^{n+1}} \sum_{m=-n}^{n} v_{nm} Y_{nm}(\lambda, \vartheta). \tag{2.18}$$

Here, the spherical harmonic coefficients have been named as v_{nm}. More often, one finds the following form:

$$V(\lambda,\vartheta,r) = \frac{GM}{R}\sum_{n=0}^{\infty}\sum_{m=-n}^{n}\left(\frac{R}{r}\right)^{n+1} v_{nm}Y_{nm}(\lambda,\vartheta), \tag{2.19}$$

which differs by the factor $GM \cdot R^n$ with GM being the Earth's gravitational constant and R the equatorial radius of the Earth, making the coefficients small and unitless. For the following calculation, a change in the way of writing is needed:

$$V(\lambda,\vartheta,r) = \frac{GM}{R}\sum_{n=0}^{\infty}\sum_{m=0}^{n}\left(\frac{R}{r}\right)^{n+1} \big(c_{nm}C_{nm}(\lambda,\vartheta) + s_{nm}S_{nm}(\lambda,\vartheta)\big). \tag{2.20}$$

Here, v_{nm} was replaced by the coefficients c_{nm} and s_{nm}, and the sum starts at order zero. As already mentioned earlier, the disturbing potential like the gravitational potential is an harmonic function and can be written in terms of spherical harmonics. The expression is equivalent to that of the gravitational potential (e.g. Eq. (2.20)) but with another set of spherical harmonic coefficients.

With this in hand, geoid heights can easily be calculated according to the equations given in Sec. 2.1. Also gravity anomalies can be computed by taking the radial derivative of the disturbing potential. For the calculation of the deflections of the vertical, Cartesian derivatives in a local north oriented reference frame have to be calculated. This is often approached by using the chain rule. In this work, we formulate the Cartesian derivatives directly as linear combination of spherical harmonics in an Earth fixed reference frame (following Ilk (1983)). To get the derivatives in the local frame, a transformation has to be done subsequently. The derivative of the potential written in terms of fully normalized quantities reads

$$\boldsymbol{g} = \nabla V(\lambda,\vartheta,r) = GM\sum_{n=0}^{\infty}\sum_{m=0}^{n} R^n\big(c_{nm}\nabla\Big(\frac{1}{r^{n+1}}C_{nm}(\lambda,\vartheta)\Big) + s_{nm}\nabla\Big(\frac{1}{r^{n+1}}S_{nm}(\lambda,\vartheta)\Big)\big), \tag{2.21}$$

where

$$\nabla\Big(\frac{1}{r^{n+1}}C_{nm}(\lambda,\vartheta)\Big) = \sqrt{\frac{2n+1}{2n+3}}\,\frac{1}{2r^{n+2}}\begin{pmatrix}\alpha_{n+1,m-1}C_{n+1,m-1} - \alpha_{n+1,m+1}C_{n+1,m+1}\\ -\alpha_{n+1,m-1}S_{n+1,m-1} - \alpha_{n+1,m+1}S_{n+1,m+1}\\ -2\alpha_{n+1,m}C_{n+1,m}\end{pmatrix} \tag{2.22}$$

$$\nabla\Big(\frac{1}{r^{n+1}}S_{nm}(\lambda,\vartheta)\Big) = \sqrt{\frac{2n+1}{2n+3}}\,\frac{1}{2r^{n+2}}\begin{pmatrix}\alpha_{n+1,m-1}S_{n+1,m-1} - \alpha_{n+1,m+1}S_{n+1,m+1}\\ \alpha_{n+1,m-1}C_{n+1,m-1} + \alpha_{n+1,m+1}C_{n+1,m+1}\\ -2\alpha_{n+1,m}S_{n+1,m}\end{pmatrix}. \tag{2.23}$$

The unknown quantities are explained later on. Setting up the GOCE observation equations requires the second Cartesian derivatives, which can be built following the same strategy:

$$\boldsymbol{M} = \nabla\nabla V(\lambda,\vartheta,r) = GM\sum_{n=0}^{\infty}\sum_{m=0}^{n} R^n\big(c_{nm}\nabla\nabla\Big(\frac{1}{r^{n+1}}C_{nm}(\lambda,\vartheta)\Big) + s_{nm}\nabla\nabla\Big(\frac{1}{r^{n+1}}S_{nm}(\lambda,\vartheta)\Big)\big), \tag{2.24}$$

where

$$\nabla\nabla\Big(\frac{1}{r^{n+1}}C_{nm}(\lambda,\vartheta)\Big) = \sqrt{\frac{2n+1}{2n+5}}\,\frac{1}{4r^{n+3}}\begin{pmatrix}A^c_{xx} & A^c_{xy} & A^c_{xz}\\ A^c_{yx} & A^c_{yy} & A^c_{yz}\\ A^c_{zx} & A^c_{zy} & A^c_{zz}\end{pmatrix} \tag{2.25}$$

$$\nabla\nabla\Big(\frac{1}{r^{n+1}}S_{nm}(\lambda,\vartheta)\Big) = \sqrt{\frac{2n+1}{2n+5}}\,\frac{1}{4r^{n+3}}\begin{pmatrix}A^s_{xx} & A^s_{xy} & A^s_{xz}\\ A^s_{yx} & A^s_{yy} & A^s_{yz}\\ A^s_{zx} & A^s_{zy} & A^s_{zz}\end{pmatrix} \tag{2.26}$$

and

$$\begin{aligned}
A^c_{xx} &= \alpha_{n+2,m-2}C_{n+2,m-2} - (2+\delta_{1m})\alpha_{n+2,m}C_{n+2,m} + \alpha_{n+2,m+2}C_{n+2,m+2}\\
A^c_{xy} &= A^c_{yx} = -\alpha_{n+2,m-2}S_{n+2,m-2} - \delta_{1m}\alpha_{n+2,m}S_{n+2,m} + \alpha_{n+2,m+2}S_{n+2,m+2}\\
A^c_{xz} &= A^c_{zx} = -2\alpha_{n+2,m-1}C_{n+2,m-1} + 2\alpha_{n+2,m+1}C_{n+2,m+1}\\
A^c_{yy} &= -\alpha_{n+2,m-2}C_{n+2,m-2} - (2-\delta_{1m})\alpha_{n+2,m}C_{n+2,m} - \alpha_{n+2,m+2}C_{n+2,m+2}\\
A^c_{yz} &= A^c_{zy} = 2\alpha_{n+2,m-1}S_{n+2,m-1} + 2\alpha_{n+2,m+1}S_{n+2,m+1}\\
A^c_{zz} &= 4\alpha_{n+2,m}C_{n+2,m}\\
A^s_{xx} &= \alpha_{n+2,m-2}S_{n+2,m-2} - (2-\delta_{1m})\alpha_{n+2,m}S_{n+2,m} + \alpha_{n+2,m+2}S_{n+2,m+2}\\
A^s_{xy} &= A^s_{yx} = \alpha_{n+2,m-2}C_{n+2,m-2} - \delta_{1m}\alpha_{n+2,m}C_{n+2,m} - \alpha_{n+2,m+2}C_{n+2,m+2}\\
A^s_{xz} &= A^s_{zx} = -2\alpha_{n+2,m-1}S_{n+2,m-1} + 2\alpha_{n+2,m+1}S_{n+2,m+1}\\
A^s_{yy} &= -\alpha_{n+2,m-2}S_{n+2,m-2} - (2+\delta_{1m})\alpha_{n+2,m}S_{n+2,m} - \alpha_{n+2,m+2}S_{n+2,m+2}\\
A^s_{yz} &= A^s_{zy} = -2\alpha_{n+2,m-1}C_{n+2,m-1} - 2\alpha_{n+2,m+1}C_{n+2,m+1}\\
A^s_{zz} &= 4\alpha_{n+2,m}S_{n+2,m}
\end{aligned} \tag{2.27}$$

and

$$\begin{aligned}
\alpha_{n+1,m-1} &= \sqrt{(n-m+2)(n-m+1)(1+\delta_{1m})}\\
\alpha_{n+1,m} &= \sqrt{(n+m+1)(n-m+1)}\\
\alpha_{n+1,m+1} &= \sqrt{(n+m+2)(n+m+1)(1+\delta_{0m})}
\end{aligned}$$

$$\begin{aligned}
\alpha_{n+2,m-2} &= \sqrt{(n-m+4)(n-m+3)(n-m+2)(n-m+1)(1+\delta_{2m})}\\
\alpha_{n+2,m-1} &= \sqrt{(n+m+1)(n-m+3)(n-m+2)(n-m+1)(1+\delta_{1m})}\\
\alpha_{n+2,m} &= \sqrt{(n+m+2)(n+m+1)(n-m+2)(n-m+1)}\\
\alpha_{n+2,m+1} &= \sqrt{(n-m+1)(n+m+3)(n+m+2)(n+m+1)(1+\delta_{0m})}\\
\alpha_{n+2,m+2} &= \sqrt{(n+m+4)(n+m+3)(n+m+2)(n+m+1)(1+\delta_{0m})}
\end{aligned} \tag{2.28}$$

Here, it was agreed on $C_{nm} = 0$ if $m < 0$, and $S_{nm} = 0$ if $m \leq 0$. And $s_{nm} = 0$ if $m = 0$ as usual.

For the evaluation of gravity field models, i.e. different sets of spherical harmonic coefficients, it is interesting to have a look onto the degree variances. The signal degree amplitude

$$\sigma(n) = \sqrt{\sum_{m=0}^{n} c_{nm}^2 + s_{nm}^2} \tag{2.29}$$

is the square root of the signal degree variance and indicates how much energy is contained in the specific degree. The given degree amplitude is in terms of unitless coefficients, but other definitions in terms of physical units are also possible, for example

$$\sigma_N(n) = R \cdot \sigma(n) \tag{2.30}$$

for the degree amplitude in geoid heights. Kaula's rule of thumb,

$$\sigma(n) \approx \sqrt{(2n+1) \cdot \frac{10^{-10}}{n^4}}\,, \tag{2.31}$$

is an approximation for the signal content in the different degrees. Furthermore, there are the difference degree amplitudes, which give a hint onto the strength of the difference signal. If one of

the gravity models dominates the error, the difference degree amplitudes can be thought of as error of this specific solution. If the formal errors of a model are known from the estimation procedure, error degree amplitudes can be calculated as

$$\sigma_\sigma(n) = \sqrt{\sum_{m=0}^{n} \sigma_{c_{nm}}^2 + \sigma_{s_{nm}}^2}\,. \tag{2.32}$$

If the comparison to the difference degree variances gives a good result, the stochastic model that was used in the estimation procedure can be assumed to work well. As a consequence of the polar gap, the near zonal spherical harmonic coefficients of GOCE-only models are highly correlated and, taken individually, not very meaningful. They are thus often left out in the calculation of degree variances. A rule of thumb derived by van Gelderen and Kopp (1997) (see also Sneeuw and van Gelderen (1997)) indicates up to which order the coefficients are affected by the polar gap:

$$m_{\max} = \left|\frac{\pi}{2} - i\right| n \tag{2.33}$$

with i being the inclination of the satellite orbit, i.e. $i = 96.7°$ for GOCE, to be inserted in radians. As an alternative, one can use the median of the coefficients per degree, which is not affected by the polar gap as well. Note that the degree amplitude calculated by leaving out specific coefficients has to be scaled in order to make the signal content comparable to other models.

2.4.2 Radial basis functions

In gravity field analysis, one often has to cope with heterogeneous data. A regional representation is then in many cases more suitable than a representation in terms of global base functions, as stated earlier in Sec. 1.1. A variety of different space localizing base functions has been proposed for representation. Schmidt et al. (2007) gives an overview of splines and wavelets, which are both derived from radial base functions or, as he calls them, spherical base functions. Wavelets are used e.g. by Chambodut et al. (2005), Panet et al. (2011), Schmidt et al. (2008), Klees et al. (2008). Here the signal is decomposed in a smoothed part and additional detail signals, which are represented by so called scaling and wavelet functions. By doing so, one gets a representation of multiple resolution. Another type of base functions, the Slepians, are as concentrated as possible in both the space and frequency domain. They can, for example, be obtained by maximizing the power of a bandlimited function within a certain spatial region, see e.g. Wieczorek and Simons (2005). A mascon is a regional surface mass. In gravity field recovery, the gravity effect of individual mascons is superimposed, and the mass, or in other words, the height of water with equivalent mass is estimated (cf. Rowlands et al., 2010). In this thesis, radial basis functions are used; see Freeden et al. (1998) for a comprehensive review.

As an alternative to spherical harmonics, the function F on the unit sphere can also be represented as a linear combination of radial basis functions:

$$F(\boldsymbol{x}) = \sum_{k=0}^{\infty} a_k \Phi_k(\boldsymbol{x}). \tag{2.34}$$

Here, Φ_k denotes the kth basis function being located at the nodal point $\boldsymbol{x}_k$, and a_k is the respective scaling coefficient. As a consequence of the symmetry, the basis function can be expressed as series expansion in terms of Legendre polynomials:

$$\Phi_k(\boldsymbol{x}) = \sum_{n=0}^{\infty} \sqrt{2n+1}\,\varphi_n P_n(\boldsymbol{x}\cdot\boldsymbol{x}_k) \tag{2.35}$$

$$= \sum_{n=0}^{\infty} \sum_{m=-n}^{n} \varphi_n Y_{nm}(\boldsymbol{x}) Y_{nm}(\boldsymbol{x}_k), \tag{2.36}$$

where the shape coefficients φ_n determine the appearance of the function, and Eqs. (2.35) and (2.36) are related by the addition theorem Eq. (2.17).

The gravitational potential in terms of radial basis functions reads

$$V(\boldsymbol{r}) = \sum_{k=0}^{\infty} a_k \Phi_k(\boldsymbol{r}). \tag{2.37}$$

Note that the basis function, though indicated by using the same notation, is not the same as the basis function in Eq. (2.34) but contains the upwards continuation operator as already known from the spherical harmonics:

$$\Phi_k(\boldsymbol{r}) = \frac{GM}{R} \sum_{n=0}^{\infty} \left(\frac{R}{r}\right)^{n+1} \sqrt{2n+1}\, \varphi_n P_n(\cos \sphericalangle(\boldsymbol{r}, \boldsymbol{r}_k)), \tag{2.38}$$

where $\cos \sphericalangle(\boldsymbol{r}, \boldsymbol{r}_k) = \frac{\boldsymbol{r}}{r} \cdot \frac{\boldsymbol{r}_k}{R}$. Here, the basis functions are arranged on a sphere with radius R at nodal points $\boldsymbol{r}_k$.

The second derivatives of the gravity potential, which are available from the observations of the gravity mission GOCE, can be related to the RBFs by

$$\boldsymbol{M} = \nabla\nabla V(\boldsymbol{r}) = \sum_{k=0}^{\infty} a_k \nabla\nabla \Phi_k(\boldsymbol{r}) \tag{2.39}$$

with

$$\nabla\nabla \Phi_k(\boldsymbol{r}) = \begin{pmatrix} \frac{\partial^2\Phi}{\partial x^2} & \frac{\partial^2\Phi}{\partial x \partial y} & \frac{\partial^2\Phi}{\partial x \partial z} \\ \frac{\partial^2\Phi}{\partial y \partial x} & \frac{\partial^2\Phi}{\partial y^2} & \frac{\partial^2\Phi}{\partial y \partial z} \\ \frac{\partial^2\Phi}{\partial z \partial x} & \frac{\partial^2\Phi}{\partial z \partial y} & \frac{\partial^2\Phi}{\partial z^2} \end{pmatrix} \tag{2.40}$$

(Eicker, 2008). The basis functions can be interpreted as functions of the radial distance $r = \sqrt{x^2+y^2+z^3}$ and the quantity $t = \frac{xx_k+yy_k+zz_k}{rR}$, which is the cosine of the opening angle. Applying the chain rule twice leads to the second derivatives in the global system:

$$\begin{aligned} \frac{\partial^2\Phi}{\partial\alpha\partial\beta} &= \frac{\partial}{\partial\beta}\left(\frac{\partial\Phi}{\partial r}\frac{\partial r}{\partial\alpha} + \frac{\partial\Phi}{\partial t}\frac{\partial t}{\partial\alpha}\right) \\ &= \frac{\partial^2\Phi}{\partial r^2}\frac{\partial r}{\partial\beta}\frac{\partial r}{\partial\alpha} + \frac{\partial^2\Phi}{\partial r\partial t}\frac{\partial t}{\partial\beta}\frac{\partial r}{\partial\alpha} + \frac{\partial\Phi}{\partial r}\frac{\partial^2 r}{\partial\alpha\partial\beta} \\ &\quad + \frac{\partial^2\Phi}{\partial t^2}\frac{\partial t}{\partial\beta}\frac{\partial t}{\partial\alpha} + \frac{\partial^2\Phi}{\partial t\partial r}\frac{\partial r}{\partial\beta}\frac{\partial t}{\partial\alpha} + \frac{\partial\Phi}{\partial t}\frac{\partial^2 t}{\partial\alpha\partial\beta} \end{aligned} \tag{2.41}$$

with

$$\begin{aligned} \frac{\partial\Phi}{\partial r} &= \frac{GM}{R}\sum_{n=0}^{\infty} -\frac{(n+1)}{r}\left(\frac{R}{r}\right)^{n+1}\sqrt{2n+1}\,\varphi_n P_n(t) \\ \frac{\partial\Phi}{\partial t} &= \frac{GM}{R}\sum_{n=0}^{\infty} \left(\frac{R}{r}\right)^{n+1}\sqrt{2n+1}\,\varphi_n \frac{\partial P_n(t)}{\partial t} \\ \frac{\partial^2\Phi}{\partial r^2} &= \frac{GM}{R}\sum_{n=0}^{\infty} \frac{(n+1)(n+2)}{r^2}\left(\frac{R}{r}\right)^{n+1}\sqrt{2n+1}\,\varphi_n P_n(t) \\ \frac{\partial^2\Phi}{\partial t^2} &= \frac{GM}{R}\sum_{n=0}^{\infty} \left(\frac{R}{r}\right)^{n+1}\sqrt{2n+1}\,\varphi_n \frac{\partial^2 P_n(t)}{\partial t^2} \\ \frac{\partial^2\Phi}{\partial r\partial t} &= \frac{\partial^2\Phi}{\partial t\partial r}\frac{GM}{R}\sum_{n=0}^{\infty} -\frac{(n+1)}{r}\left(\frac{R}{r}\right)^{n+1}\sqrt{2n+1}\,\varphi_n \frac{\partial P_n(t)}{\partial t} \end{aligned} \tag{2.42}$$

and

$$\begin{aligned} \frac{\partial r}{\partial \alpha} &= \frac{\alpha}{r} & \frac{\partial^2 r}{\partial \alpha \partial \beta} &= \frac{1}{r}\delta_{\alpha\beta} - \frac{\alpha\beta}{r^3} \\ \frac{\partial t}{\partial \alpha} &= \frac{\alpha_k}{rR} - \frac{\alpha t}{r^2} & \frac{\partial^2 t}{\partial \alpha \partial \beta} &= -\frac{t}{r^2}\delta_{\alpha\beta} - \frac{\alpha_k\beta + \alpha\beta_k}{r^3 R} + \frac{3\alpha\beta t}{r^4} \end{aligned} \tag{2.43}$$

2.4.3 Predefined point grids

To set up the RBFs as defined in Eq. (2.38), one has to define the shape coefficients and the nodal point grid. Here, we will start with the latter. This section presents different point grids, which are all defined prior to the gravity field estimation step and independent of the observations. In contrast to the optimization of point grids being described later on, this approach is easy to implement and rather flexible, i.e. suitable for any application.

The geographical grid

The geographical grid is defined by the points of intersection of the meridians of longitude ($\lambda = \text{const}$, $\vartheta = \text{var}$) and parallels of latitude ($\lambda = \text{var}$, $\vartheta = \text{const}$) with constant angular distances $\Delta\lambda$, $\Delta\vartheta$.

The triangular vertex grid

This grid is created by the subdivision of an icosahedron. The icosahedron is one of the five Platonic solids. These are a special type of polyhedron, whose faces are regular (equilateral, equiangular), they all look the same (congruent) and they are evenly arranged. Thus, the Platonic solids are very symmetric, and since their nodes have equal distances and lie on the surface of a sphere, they provide a good starting point for a uniform segmentation of the sphere. Depending on the polyhedron chosen and the technique applied, and depending on whether the grid points are identified with the nodes or with the centers of the triangles, various kinds of point grids are created. An overview with great illustrations is given by Popko (2012). The icosahedron is the most important polyhedron for the subdivision of the sphere. An advantage is that an icosahedral grid can be easily subdivided for multiresolution or multigrid approaches. In the variant of Eicker (2008) (see also Kusche et al., 2001), the sides of the spherical triangles are evenly covered with points. The points are connected by lines running parallel to the sides of the triangles. Since these lines do not intersect in a single point, the mean of the intersection points is used to define the nodes of the new triangles. The nodes of all triangles taken together form the basis for the definition of the point grid. This implementation of the triangular grid is what Popko calls the class 1 version 3 type (Popko, 2012, pp. 199–219). The level of densification is controlled by the level parameter n. Depending on n, the global number of grid points I is given by $I = 10(n+1)^2 + 2$.

The quasi-random grid

Quasi-random numbers are not truly random, but they have a better coverage than real random numbers and are therefore sometimes used instead of them. For the quasi-random grid, the longitude of the grid points is divided into equal angular intervals. The z-coordinate is defined by a quasi-random sequence. For this, the interval $(-1, 1)$ is cut into halves several times, and the resulting numbers are set one after the other in a prescribed order. For details see Eicker (2008).

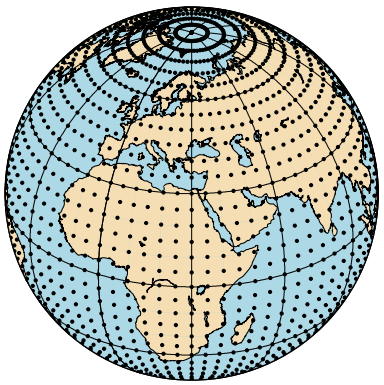


Figure 2.1: Geographical grid of $\Delta\lambda = \Delta\vartheta = 5^\circ$ spacing. In addition to the grid points, grid net lines are plotted with a spacing of 30°. This type of grid is obviously not uniform at all, which is due to the convergence of meridians.

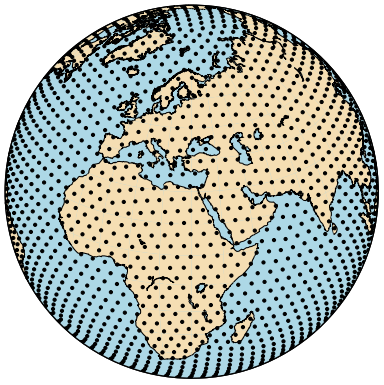


Figure 2.2: Triangular grid of level 16, which corresponds to 2892 grid points globally.

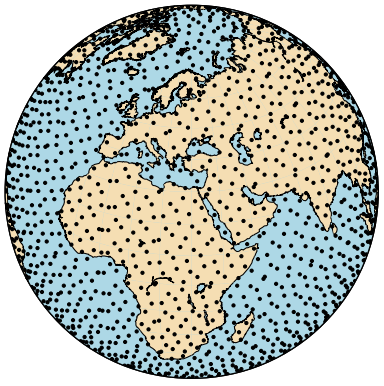


Figure 2.3: Quasi random grid

Figs. 2.1 to 2.3 show the three point grids at approximately the same resolution. These are only a few examples for possible point grids. Another example is the Reuter grid, which is frequently used in gravity field analysis (Reuter, 1982). Less common are the hexagonal grid and the Fibonacci grid (used by Bentel (2013) and Naeimi (2013), respectively).

In regional gravity field analysis, it is often claimed for a point grid that is as uniform as possible. Since the choice of the point grid is closely connected to the model resolution, one does probably not want to treat some regions differently without a reason. A uniform point distribution generates a uniform model resolution just like when using spherical harmonics. Moreover, the equal overlap of the radially symmetric basis functions is good for the stability of the normal equation system. As for the geographical grid, the grid net lines defining the locations of the grid points converge at the poles. For this reason, it is not used for the arrangement of the basis functions. Thus the geographical grid, although being equiangular, is obviously not what we mean when asking for a uniform distribution. But what exactly does it mean? The term is not clearly defined. For example, it is used for a grid with equal area elements as well as for one with equal distances. Accordingly, there are different criteria for the uniformity of a point grid and measures to assess how well they are fulfilled. Eicker (2008) compared different point grids by means of different criteria considering

the area and shape of the area elements and the distance of the points. She found that the Reuter grid and the triangular vertex grid do best. In her opinion, the test concerning the maximum of the distances of all points of the sphere to the nearest grid point is of particular importance. The shorter this distance, the closer the area element is to a spherical cap. The other way around, the distance can be used to measure the deviation from the ideal of a spherical cap. Such a circular segmentation would be perfectly suitable for being used with radial basis functions, which are isotropic and thus also have a circular shape in some sense. Following this line of reasoning, Eicker finally decided to use the triangular vertex grid, which in the present work is used again for comparison with the new approach.

2.4.4 A glance at reproducing kernel Hilbert spaces

We will begin with some information about functional spaces on the sphere to an extent that is needed to understand the choice of basis functions; further details about (general) Hilbert and Reproduction Kernel Hilbert Spaces can be found e.g. in Meschkowski (1962).

Hilbert spaces are vector spaces equipped with a norm that is derived from an inner product via $\|\cdot\| = \langle\cdot,\cdot\rangle^{1/2}$. Furthermore, a Hilbert space H is required to be complete, i.e. every Cauchy sequence of elements of H has to converge to an element of H. The Euclidean space, which is basically the $\mathbb{R}^n$ supplemented by the inner product $\langle \boldsymbol{v}, \boldsymbol{w}\rangle = \sum_k v_k w_k$, is a familiar example of a Hilbert space. Another example is the space of square integrable functions, $L_2(\mathbb{R})$, which can be thought of as the generalization of the Euclidean space in the sense that the elements are no longer finite dimensional vectors but functions on the real line. The inner product here turns into $\langle f, g\rangle_{L_2(\mathbb{R})} = \int f(x)g(x)dx$ inducing the norm $\|f\|^2_{L_2(\mathbb{R})} = \int f(x)^2 dx$. To become an element of $L_2(\mathbb{R})$, the function f has to possess finite norm, a requirement that is obviously met by functions whose square is integrable.

As in this thesis we are concerned with the approximation of spherical data, it is sensible to introduce the space of square integrable functions on the sphere, here denoted by $L_2(\Omega)$. For two functions $f(\boldsymbol{x})$ and $g(\boldsymbol{x})$ on the unit sphere, i.e. $\|\boldsymbol{x}\| = 1$ using the Euclidean norm, the inner product and norm are defined as

$$\langle f, g\rangle_{L_2(\Omega)} = \frac{1}{4\pi}\int f(\boldsymbol{x})g(\boldsymbol{x})d\Omega \tag{2.44}$$

$$\|f\|^2_{L_2(\Omega)} = \langle f, f\rangle_{L_2(\Omega)}, \tag{2.45}$$

respectively. When we express f and g by spherical harmonics according to Eq. (2.14) with coefficients f_{nm} and g_{nm}, (2.44) and (2.45) can be rewritten as

$$\langle f, g\rangle_{L_2(\Omega)} = \sum_{n=0}^{\infty}\sum_{m=-n}^{n} f_{nm}g_{nm} \tag{2.46}$$

$$\|f\|^2_{L_2(\Omega)} = \sum_{n=0}^{\infty}\sum_{m=-n}^{n} f^2_{nm}, \tag{2.47}$$

which can be shown in consideration of the orthogonality relations, Eq. (2.16). It is also because we decided on using 4π-orthogonal spherical harmonics that the additional coefficient $\frac{1}{4\pi}$ was introduced in Eq. (2.44). Note that RBFs, in contrast to spherical harmonics, are not orthogonal with respect to the L_2-norm, which is clear from the inner product

$$\langle \Phi_i, \Phi_j\rangle_{L_2(\Omega)} = \sum_{n=0}^{\infty}\varphi_n^2\sum_{m=-n}^{n} Y_{nm}(\boldsymbol{x}_i)Y_{nm}(\boldsymbol{x}_j). \tag{2.48}$$

The concept of reproducing kernel Hilbert spaces (RKHS) enables us to define a space that only includes functions with a certain smoothness. The RKHS is a special Hilbert space equipped with a kernel, which by multiplying with a function reproduces this function. The inner product is defined slightly different as

$$\langle f, g \rangle^2_{H(\Omega)} = \sum_{n=0}^{\infty} \sum_{m=-n}^{n} \frac{f_{nm} g_{nm}}{\lambda_n} \tag{2.49}$$

$$\|f\|^2_{H(\Omega)} = \sum_{n=0}^{\infty} \sum_{m=-n}^{n} \frac{f_{nm}^2}{\lambda_n}, \tag{2.50}$$

where λ_n denote the eigenvalues of the kernel. As already stated earlier, for the function f to be an element of $H(\Omega)$, the norm has to be finite. This criterion is only met if the spherical harmonic coefficients, f_{nm}, approach zero sufficiently fast, which can be associated with f showing a certain degree of smoothness. Comparing Eq. (2.50) to Eq. (2.47), we see that the coefficients, for they are additionally divided by λ_n, should decrease even faster than it is necessary in L_2. Generally, the functions have to be smoother than those belonging to the L_2, while the degree of smoothness depends on the kernel. The more restrictive the kernel, the smoother the functions. When we build the inner product of the radial basis functions, this time with respect to the norm induced by the kernel, we find

$$\langle \Phi_i, \Phi_j \rangle_{H(\Omega)} = \sum_{n=0}^{\infty} \frac{\varphi_n^2}{\lambda_n} \sum_{m=-n}^{n} Y_{nm}(\boldsymbol{x}_i) Y_{nm}(\boldsymbol{x}_j). \tag{2.51}$$

2.4.5 The choice of basis functions in Eicker (2008)

In the frame of data modeling, one usually searches for a comparatively simple model that is still adequate to explain the data. This is also what is postulated by Occam's razor, a principle that will become important later on in this thesis. To restrict the solution space to those simple models using the concepts that has been introduced in the last section, the problem breaks down to just choosing the appropriate norm. If the task is to approximate a function with abrupt changes and jumps, the L_2-norm, which also allows for non-continuous functions, might be a good choice. However, the gravity potential is continuous, so that the L_2 might not be suitable.

Within this thesis, I generally use the regional approach as was introduced by Eicker (2008) and described concisely in Eicker et al. (2014). Eicker formulates the problem of regional gravity analysis within the frame of the theory of reproducing kernel Hilbert spaces, where the norm is induced by the choice of the kernel, as explained earlier. Particularly, Eicker uses the covariance function of the gravity potential,

$$\mathcal{C}(\boldsymbol{e}_i, \boldsymbol{e}_k) = \sum_{n=0}^{\infty} \frac{\sigma_n^2}{2n+1} \sum_{m=-n}^{n} Y_{nm}(\boldsymbol{e}_i) Y_{nm}(\boldsymbol{e}_k), \tag{2.52}$$

as kernel function with the eigenvalues $\lambda_n = \frac{\sigma_n^2}{2n+1}$, which is quite common in geodesy (Tscherning, 1977). The covariance function describes the similarity of neighboring points; it is thus appropriate to make assumptions on the smoothness of the field. However, if the covariance function of the gravity potential is introduced as reproducing kernel, the gravity potential itself will not be a part of the space that is associated with the kernel (Tscherning, 1977). According to Moritz (1980, Ch. 25), this problem is primarily a theoretical one because on the one hand side the degree variances are not perfectly known and on the other side they could be modified to make the kernel slightly rougher. Eicker (2008) tested this approach by adapting the regularization matrix accordingly but did not notice an effect in the practical application.

As already described above in Sec. 2.4.4, the functions to be part of a RKHS are forced in that their Legendre coefficients go towards zero faster than the eigenvalues of the kernel; the functions have to be at least as smooth as the kernel. In fact, the kernel function is the roughest function that still belongs to the space and is therefore a straightforward choice as basis function because this way the space can fully be exploited. If the kernel function is used as basis function, the basis functions are referred to as spherical splines. Eicker follows a different way: with a look onto the inner product of the RBFs, she chooses the coefficients in such a way that the basis functions become decorrelated with respect to the norm induced by the kernel, which will be exploited later in the regularization step. Thus, she chooses the shape coefficients according to

$$\varphi_n = \frac{\sigma_n}{\sqrt{2n+1}}, \tag{2.53}$$

which causes the inner product of the basis functions, Eq. (2.51), to become the Dirac impulse:

$$\sum_{n=0}^{\infty} \frac{\varphi_n^2}{\lambda_n} \sum_{m=-n}^{n} Y_{nm}(\boldsymbol{x}_i) Y_{nm}(\boldsymbol{x}_j) = \delta(\boldsymbol{x}_i, \boldsymbol{x}_j) \tag{2.54}$$

(c.f. Eicker, 2008, Eq. (5.3.1)). In the above equation, $\sigma_n^2 = \sum_{m=-n}^{n} f_{nm}^2$ are the degree variances. The appearance of the basis functions is thus optimally adapted to the spectral characteristics of the gravity field, the representation of which they are constructed for, e.g. the RBFs reflect the decrease of the gravity field towards the higher degrees.

2.5 Gravity field adjustment in the Gauss-Markov model

In the current section, it is shown how to derive the spherical harmonic coefficients or the scaling coefficients of the radial basis functions from the analysis of GOCE data. The processing of Satellite to Satellite Tracking (SST) data was extensively described by Mayer-Gürr (2006), and the same approach was applied here without modification. In this section, we are therefore mainly occupied with the Satellite Gravity Gradiometry (SGG) part. However, in the last part of the section, which is about the combined solution, the SST part is assumed to be available in the form of normal equations. For further processing details, it is referred to the results, Ch. 5.

2.5.1 Setting up the observation equations

The gravity gradients provided by GOCE serve as observations. What we are looking for are the coefficients of the expansion of the potential into spherical harmonics or radial basis functions. The gravity gradients correspond to the second Cartesian derivatives of the gravity potential. The expansion of the potential has thus only to be differentiated twice:

$$\boldsymbol{M} = GM \sum_{n=0}^{\infty} \sum_{m=0}^{n} R^n \big(c_{nm} \nabla\nabla \Big(\frac{1}{r^{n+1}} C_{nm}(\lambda, \vartheta) \Big) + s_{nm} \nabla\nabla \Big(\frac{1}{r^{n+1}} S_{nm}(\lambda, \vartheta) \Big) \big). \tag{2.55}$$

Even if the signal content of the gravity gradients is unlimited in theory, in practice, to do numerical calculations, the problem has to be chosen smaller, and the number of basis functions has to be limited. Moreover, the sensor is only sensitive within a limited frequency range, i.e. the actual signal content of the data is limited. Because of the differential measuring principle, the coefficient of spherical harmonic degree $n = 0$ cannot be determined accurately. As common in gravity field adjustment, the coefficient is set to a constant value instead of being estimated. The same is true

for the coefficients of degree $n = 1$, where it is assumed that it becomes zero because of a particular choice of the reference system. With these limitations, the functional relationship is better written as

$$\boldsymbol{M} - \boldsymbol{M}_0 = GM \sum_{n=2}^{N} \sum_{m=0}^{n} R^n \Big(c_{nm} \nabla\nabla\Big(\frac{1}{r^{n+1}} C_{nm}(\lambda, \vartheta) \Big) + s_{nm} \nabla\nabla\Big(\frac{1}{r^{n+1}} S_{nm}(\lambda, \vartheta) \Big) \Big). \tag{2.56}$$

Equivalently, the functional relationship in terms of RBFs reads

$$\boldsymbol{M} - \boldsymbol{M}_0 = \sum_{k=1}^{K} a_k \nabla\nabla\Bigg(\underbrace{\frac{GM}{R} \sum_{n=2}^{N} \left(\frac{R}{r} \right)^{n+1} \sqrt{2n+1}\, \varphi_n P_n(\cos \sphericalangle(\boldsymbol{r}, \boldsymbol{r}_k))}_{=\Phi_k(\boldsymbol{r})} \Bigg). \tag{2.57}$$

As above, the expansion starts from degree $n = 2$, and the model resolution, i.e. the number of basis functions and the expansion degree of the kernel, has been limited. Vice versa, any signal that is contained in the data but not in the model has to be removed from the data. Besides the central term, this also includes time variations, which however are not very important for GOCE. Although gravity field determination from GOCE data in the above form is a linear problem, one frequently reduces not only the central term but an entire reference field. When working with regional basis functions on a regionally restricted area, this is particularly important, as structures that are large compared to the size of the area cannot be captured adequately.

At this point, only the scaling coefficients are searched for. The remaining quantities, i.e. the point grid and the shape coefficients, have to be specified prior to the gravity field adjustment. There are some criteria that help to decide about the proper resolution: strength and characteristic of the measurement noise, signal strength which varies because of the sensor used, the orbit height and the region. As stated above, the RBFs are always applied to model residual fields. The shape coefficients should be chosen in such a way that they adequately represent the signal content to be modeled. For the lower degrees, the formal errors of the reference solution may be chosen, as they represent how much energy is left after the reference field was substracted. For the higher degrees, Kaula's rule can be used. For the sake of completeness, it should be added that at this stage of the work the grid is chosen uniformly. Having chosen all these parameters, we get a linear relationship between observations and parameters, and the observation equations can be directly derived according to

$$\boldsymbol{y} + \boldsymbol{v} = \boldsymbol{A}\boldsymbol{\beta} \tag{2.58}$$

with $\boldsymbol{y}$ including the gravity gradients arranged in vector format. Because of their higher quality, only 4 of 6 gradients are worth being considered in gravity field adjustment. The design matrix $\boldsymbol{A}$ includes the partial derivatives of the parameters being included in the parameter vector $\boldsymbol{\beta}$. As the measurements are taken within the gradiometer reference system, the observation equations have yet to be transformed from the Earth fixed frame to this frame. This is realized by using the star camera data to set up the corresponding rotation matrix. The observations itselves should not be transformed because, in this way, gradients of different accuracy would be mixed. To make the problem of GOCE gravity field recovery manageable in the first place, the observation equations are set up separately for every arc. We divide the observations in arcs $\boldsymbol{y}_i$ and set up the corresponding design matrices. When accumulating the normal equations from the individual arcs later on, it is assumed that individual arcs do not have correlations. This is certainly not true when dealing with GOCE data. To hold the error as small as possible, further empirical parameters are introduced. To give an example, if an offset per arc and tensor element is chosen, the design matrix for that particular arc has to be expanded by as many columns as different tensor elements are used, every column containing a one for the particular gradient and zero everywhere else.

2.5.2 Stochastic modeling for the gravity gradients

In the following, the stochastic modeling for the gravity gradients is described in detail, as it is not explained in Mayer-Gürr (2006) or Eicker (2008) but very important for real data analysis. The GOCE accelerometers possess a complex error behavior and so do the gravity gradients, which are derived from the accelerometer measurements. To take this adequately into account, a covariance function was estimated directly from the observations.

To do so, the observation time series, which can also be interpreted as the realization of a random process, has to be stationary and ergodic. Then one single realization of finite length would be enough to derive statistical values. Stationarity requires that the probability distribution does not change over the observation period, meaning that the moments of the distribution remain the same. To ensure that the expectation value—the first statistical moment—remains constant over time, any deterministic trend inherent in the time series must be removed. To this end, residuals were calculated from the gravity gradients by subtracting a model for the time averaged gravity field and further background models for time variable effects. Also the linear trend present in the gradiometer data was eliminated.

The computation of variances and co-variances in the classical manner, i.e. as convolution in the time domain, may yield an empirical covariance function that is not positive definite. An analytical function which is positive definite by definition is often adjusted to the empirical function to remedy this problem. I opted for a different approach and calculated the covariance function as the inverse Fourier transform of the power spectral density of the residuals. In this way, positive definiteness of the empirical covariance function is guaranteed. The approach is thus much closer to the original observations.

The discrete Fourier transform,

$$H_k = \sum_{n=0}^{N-1} h_n \exp\left(i\, 2\pi^{nk}/_N\right), \tag{2.59}$$

maps the complex functional values h_n onto the complex Fourier coefficients H_k, both being of length N (see e.g. Press et al., 2007, p. 607). When N is even, which is assumed throughout this derivation, then $k = -({}^N/_2 - 1), .., {}^N/_2$. The term ${}^k/_N$, which is part of the argument of the exponential function, denotes the frequency in units of cycles per number of samples. It is linked to an ordinary frequency by

$$f_k = \frac{k}{N\Delta t} \tag{2.60}$$

with Δt being the sampling interval. The power spectrum (or power spectral density, PSD) can now be computed by taking the absolute square of the Fourier coefficients (Press et al., 2007, pp. 602–603):

$$\mathrm{PSD}(f_k) = |H_k|^2. \tag{2.61}$$

Frequently, one is only interested in how much energy is concentrated in a specific frequency range. Then one does no longer distinguish between positive and negative frequencies but folds the PSD in half and sums up the two sides. This leads to the so-called one-sided power spectrum. If the input signal is real, the usual case, both sides are equal, so we get

$$\mathrm{PSD}(f_k) = \begin{cases} |H_k|^2 & \text{for } k = 0 \text{ or } k = {}^N/_2 \\ 2\,|H_k|^2 & \text{for } 0 < k < {}^N/_2. \end{cases} \tag{2.62}$$

So far, nothing has been said about the normalization constant, which still has to be applied to the PSD. There are plenty of different conventions to normalize a PSD, and the Parseval theorem,

$$\sum_n |h_n|^2 = \frac{1}{N} \sum_k |H_k|^2, \tag{2.63}$$

here given in its discrete form, might be helpful for the interpretation (Press et al., 2007, p. 608). When for example the PSD is normalized by $\frac{1}{N^2}$, then adding the values of the PSD together yields the variance of the signal. However, these values might be difficult to understand because they depend on the length of the input signal. But when it is additionally divided by $\Delta f = \frac{1}{N\Delta t}$, which corresponds to a normalization factor of $\frac{\Delta t}{N}$, then one can directly read the variance per frequency interval of unit length. Finally, the PSD is transformed back to the time domain. By taking the absolute square in the formula for the PSD, the imaginary part of the Fourier coefficients vanishes, and only cosine coefficients remain. As the real part of the inverse Fourier transform is identical to the cosine transform, I directly use the latter (Press et al., 2007, p. 624):

$$\mathrm{Cov} = \sum_{k=0}^{N/2} \mathrm{PSD}(f_k) \cos\left(2\pi {}^{kn}/_{N}\right). \tag{2.64}$$

The variance is specified by the first value of the covariance function. The first value is calculated by adding up the PSD, as is clear from Eq. (2.64). In order that this value becomes the variance, the proper normalization constant has to be applied, as already discussed earlier. The cross covariance function can be determined in a very similar way (cf. Press et al., 2007, pp. 648–649).

If $f_1, .., f_n$ are positive definite functions, and $c_i \leq 0$, then $f(x) = \sum_i c_i f_i(x)$ is again positive definite (Stewart, 1976). The covariance function that was obtained in Eq. (2.64) is a linear combination of cosine functions with the coefficients being the entries of the PSD. According to Stewart (1976), the cosine is a positive definite function, and the PSD is not negative by definition; see Eq. (2.61). Or in other words: if a function results from the inverse Fourier transform, and the Fourier coefficients are at least not negative, then f is positive definite (Stewart, 1976). The covariance function that is calculated using the PSD as intermediate step is thus not less than positive definite. A real-valued function f is positive definite if it fulfills

$$\sum_{i=1}^{n} \sum_{j=1}^{n} f(x_i - x_j)\xi_i\xi_j \geq 0 \tag{2.65}$$

for every choice of ξ_i. From this, some elementary properties can be derived:

$$0 \leq f(0) \tag{2.66}$$

$$|f(x)| \leq f(0) \tag{2.67}$$

(Koch et al., 2010). Apart from a few pathological cases, the covariance function calculated by our approach is even strictly positive definite, i.e. it fulfills the above formulas with the greater-than sign only. We now set up a matrix $\boldsymbol{A}$ with the elements $a_{ij} = f(x_i - x_j)$ and $i, j = 1, .., n$. We further define $\boldsymbol{\xi} = [\xi_i]$. The matrix is positive definite when

$$\boldsymbol{\xi}'\boldsymbol{A}\boldsymbol{\xi} > 0 \tag{2.68}$$

and zero only if $\xi_1, .., \xi_n = 0$. If the above formula, Eq. (2.68), is fulfilled for all vectors except the zero vector, then it is also fulfilled for the subspace of vectors with some $\xi_i = 0$, which means that also the principle submatrix of the corresponding indices is positive definite. Thus, when using only a part of our strictly positive definite covariance function to set up the covariance matrix, we still get the desired positive definite matrix.

In frequency domain, large outliers behave very similar to the Dirac delta function, i.e. they spread over the entire range of frequencies and must therefore be eliminated beforehand. I used a simple highpass filter in order to eliminate the long-wavelength noise and then removed large outliers using a threshold value. Moreover, the finite input sequence is in fact the product of an infinite sequence and a square window function. In order to mitigate leakage resulting from any act of windowing, the choice of an alternative window function is suitable although it is not particularly important which window is actually used. However, when choosing another window function, also the normalization of the PSD changes. It is important that the input signal contains neither a mean nor a trend. This is because during the process of windowing, the edges of the signal are pulled towards zero, which would result in spurious signal.

The disadvantage of this approach compared to the classical approach is that the input data have to be equidistant in time and without any gaps. Therefore not the entire observation time series can be used. To estimate a covariance function with the length of one orbit arc, strictly speaking only data of twice the length are needed. Nevertheless, the longest continuous piece of data was used, as it gives a PSD that is as representative as possible and highly resolved. Alternatively, one could average over individual bins or equivalently split up the data time series into several segments, transform them individually and average over the resulting PSDs. Using either approach would yield a smoother type of PSD. The influence of minor outliers would thereby probably decrease. Connected with this, one could use segments with a certain amount of overlap, which would counteract the signal loss as a result of windowing.

The procedure must be repeated in an iterative manner because the residuals are not sufficiently well known from the beginning. Convergence will however be fast, as deviations in the covariance matrix mainly affect the stochastic model of the parameters.

In practice, several covariance functions are estimated to account for possible changes in the stochastic behavior, e.g. after the calibration shaking of the satellite.

2.5.3 Least-squares solution & regularization

Having set up the observation equations and the stochastic model in the previous sections, we still have to decide on an objective function to be minimized. Typically, the sum of squared residuals is chosen for this purpose:

$$\begin{aligned} J(\boldsymbol{\beta}) &= \boldsymbol{v}(\boldsymbol{\beta})^T \boldsymbol{Q}_y^{-1} \boldsymbol{v}(\boldsymbol{\beta}) && (2.69)\\ &= (\boldsymbol{A}\boldsymbol{\beta} - \boldsymbol{y})^T \boldsymbol{Q}_y^{-1} (\boldsymbol{A}\boldsymbol{\beta} - \boldsymbol{y}) && (2.70)\\ &= \boldsymbol{\beta}^T \boldsymbol{A}^T \boldsymbol{Q}_y^{-1} \boldsymbol{A}\boldsymbol{\beta} - 2\boldsymbol{\beta}^T \boldsymbol{A}^T \boldsymbol{Q}_y^{-1} \boldsymbol{y} + \boldsymbol{y}^T \boldsymbol{Q}_y^{-1} \boldsymbol{y} && (2.71)\\ &= \boldsymbol{\beta}^T \boldsymbol{N}\boldsymbol{\beta} - 2\boldsymbol{\beta}^T \boldsymbol{n} + \boldsymbol{y}^T \boldsymbol{Q}_y^{-1} \boldsymbol{y} && (2.72) \end{aligned}$$

with the substitutes

$$\begin{aligned} \boldsymbol{N} &= \boldsymbol{A}^T \boldsymbol{Q}_y^{-1} \boldsymbol{A}, && (2.73)\\ \boldsymbol{n} &= \boldsymbol{A}^T \boldsymbol{Q}_y^{-1} \boldsymbol{y}. && (2.74) \end{aligned}$$

To find the minimum of the objective function, the gradient of the function is set equal to zero:

$$\nabla_{\boldsymbol{\beta}} J(\boldsymbol{\beta}) = 2\boldsymbol{N}\boldsymbol{\beta} - 2\boldsymbol{n} = 0. \quad (2.75)$$

This leads to the normal equations

$$\boldsymbol{N}\boldsymbol{\beta} = \boldsymbol{n}, \quad (2.76)$$

and finally, after inversion, to the well-known least-squares estimator in the Gauss-Markov model:

$$\hat{\boldsymbol{\beta}} = \boldsymbol{N}^{-1}\boldsymbol{n} = (\boldsymbol{A}^T\boldsymbol{Q}_y^{-1}\boldsymbol{A})^{-1}\boldsymbol{A}^T\boldsymbol{Q}_y^{-1}\boldsymbol{y}. \tag{2.77}$$

The corresponding covariance matrix of the parameters comes out to

$$\boldsymbol{Q}_\beta = \boldsymbol{N}^{-1} = (\boldsymbol{A}^T\boldsymbol{Q}_y^{-1}\boldsymbol{A})^{-1}. \tag{2.78}$$

As mentioned earlier in Sec. 2.3, there are huge matrices involved in GOCE gravity field analysis, and the normal equations can certainly not be calculated as indicated by Eqs. (2.73) and (2.74). In the frame of the short arc approach, the satellite orbit is divided up into arcs, whose observations are assumed to be uncorrelated. This simplifies the problem considerably in that it allows to accumulate the normal equations from the individual arcs according to

$$\boldsymbol{N} = \sum_i \boldsymbol{N}_i = \sum_i \boldsymbol{A}_i^T\boldsymbol{Q}_i^{-1}\boldsymbol{A}_i \tag{2.79}$$

$$\boldsymbol{n} = \sum_i \boldsymbol{n}_i = \sum_i \boldsymbol{A}_i^T\boldsymbol{Q}_i^{-1}\boldsymbol{y}_i. \tag{2.80}$$

As a characteristic of the measurement device, GOCE gradiometry is strong in the high frequencies only. Therefore, one would not use a SGG-only solution but always combine with SST. When dealing with spherical harmonics, SST and SGG parts are combined on the level of normal equations under exactly the same principle as mentioned above. For the regional analysis, the SST part might rather be (part of) the reference model, against which the solution is regularized.

Determining the coefficients of a gravity field model from measurements taken at satellite altitude represents an inverse problem, and as most of the inverse problems, it is ill-posed. This means that the solution reacts sensitively, in the form of strongly oscillating base functions, to errors in the data. More mathematically spoken, although having a unique minimal value, the objective function will be rather flat, so that many different parameter combinations might get similar low values. The parameters are thus weakly determined and strongly correlated. In case of GOCE, besides the attenuation of the gravity field signal with orbit height, ill-posedness is also caused by the polar gap. To deal with the ill-posedness, the problem is regularized. The Tychonov regularization is widely used for this purpose. Here, the objective function is manipulated in that a penalty term is added to it:

$$J(\boldsymbol{\beta}) = (\boldsymbol{A}\boldsymbol{\beta} - \boldsymbol{y})^T\boldsymbol{Q}_y^{-1}(\boldsymbol{A}\boldsymbol{\beta} - \boldsymbol{y}) + \boldsymbol{\beta}^T\boldsymbol{R}\,\boldsymbol{\beta}. \tag{2.81}$$

By doing so, the objective function is sharpened so that the variance of the estimate decreases. $\boldsymbol{R}$ is the so-called regularization matrix, which is often diagonal. The new objective function, Eq. (2.81), leads to the estimator

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{A}^T\boldsymbol{Q}_y^{-1}\boldsymbol{A} + \boldsymbol{R})^{-1}\boldsymbol{A}^T\boldsymbol{Q}_y^{-1}\boldsymbol{y} \tag{2.82}$$

and the corresponding covariance matrix

$$\boldsymbol{Q}_\beta = (\boldsymbol{A}^T\boldsymbol{Q}_y^{-1}\boldsymbol{A} + \boldsymbol{R})^{-1}. \tag{2.83}$$

One might ask if there is also some kind of physical interpretation for the regularization term. To answer that, let us have a look onto the norm of the gravity potential built in the RKHS H. When dealing with spherical harmonics, it becomes

$$||V||^2_{H(\Omega)} = \sum_{n=0}^{\infty}\sum_{m=-n}^{n}\frac{v_{nm}^2}{\lambda_n}. \tag{2.84}$$

The discrete approximations is

$$\sum_{n=0}^{N}\sum_{m=-n}^{n}\frac{v_{nm}^2}{\lambda_n}=\boldsymbol{\beta}^T\boldsymbol{R}\boldsymbol{\beta} \tag{2.85}$$

with $\boldsymbol{\beta}$ including the unknown parameters and $\boldsymbol{R}$ the reciprocal eigenvalues of the kernel. Thus, the regularization term can be interpreted as to be the norm of the potential built in a certain RKHS, which should be minimized together with the residual squared sum of the observations. If the kernel is chosen to be the covariance function of the gravity potential, a choice that has been motivated earlier, and the degree variances are approximated by Kaula's rule of thumb, this leads to the well-known Kaula regularization. Here, the coefficients are forced towards zero the more the higher the degree. A similar type of spectral weighting can also be incorporated into the regularization process of the RBFs. When using the same approach that led to the Kaula regularization of spherical harmonics onto the RBFs, we get

$$\begin{aligned}
||V||^2_{H(\Omega)} &= \langle V,V\rangle_{H(\Omega)} && (2.86)\\
&= \left\langle \sum_i a_i\Phi_i, \sum_j a_j\Phi_j \right\rangle_{H(\Omega)} && (2.87)\\
&= \sum_i\sum_j a_i a_j \langle \Phi_i,\Phi_j\rangle_{H(\Omega)} && (2.88)\\
&= \boldsymbol{\beta}^T\boldsymbol{R}\boldsymbol{\beta} && (2.89)
\end{aligned}$$

with $\boldsymbol{\beta}$ being the RBF parameters and $\boldsymbol{R}$ including the inner products of the base functions. As shown earlier, the inner product corresponds to the Dirac impulse. $\boldsymbol{R}$ is thus approximated by the unitary matrix, which is all the more an approximation because in the discrete case we do not sum over all frequencies with equal weight so that we do not strictly get a diagonal matrix. Eicker (2008) showed, however, that this fact does not pose any problem in practical calculations. The objective function is often written including an additional weighting factor α in front of the regularization term, which is subject to further optimization. From my point of view, the great strength of the regional approach is the ability to regularize optimally for the study area. With the above choice of the regularization matrix being the unit matrix, this can even be amplified, as we can split up the regularization matrix and thus define different regularization areas within the same regional patch. There exists different strategies to determine the regularization factor. One of them is the L-curve method, for which both terms of the objective function are calculated for different values of the regularization parameter. Another possibility would be to interprete the regularization factor as variance component and to determine it in an iterative manner known as variance component estimation.

3. Topics of Bayesian statistics

Estimating the point grid for the arrangement of the basis functions in regional gravity field analysis is a nonlinear problem and variable in dimension. The Bayesian statistic provides a practical solution for this problem. Therefore, in the following chapter, the theoretical background is introduced, in which the new approach is embedded later on in Ch. 4.

3.1 Fundamentals of probability theory

In frequentist statistics, a random variable is understood as wildcard for the outcomes of a random experiment, and the associated probability distribution has to be interpreted as to reflect the frequency of outcomes. In Bayesian statistics, on the contrary, probability is a measure for certainty, and probability distributions are specified for any variable. The term *random variable* is nevertheless retained in this work, as is done in Koch (2007). In the following, a random variable is denoted by a capital letter. A specific value, which is referred to as *realization*, is denoted by a small letter. For the sake of clarity, I deviate from this standard convention when using Greek letters as variable names or when talking about multivariate distributions.

A discrete-time stochastic process is a sequence of random variables $X_1, X_2, \ldots, X_n$ on a fixed set called state space. It is said to be stationary if the joint distribution of the subset $X_n, X_{n+1}, \ldots, X_{n+k}$ does not depend on n for each fixed k.

One can consider probability theory as a special case of the so-called measure theory. As the Metropolis-Hastings-Green algorithm, which is applied in this thesis, involves measure theory, a few measure-theoretic notions shall be introduced here. Suppose S is a set, and $\mathcal{B}$ is a family of subsets of S or, to be precise, a σ-field for S, then $(S, \mathcal{B})$ is said to form a measurable space. Now a measure on this space is a function mapping $\mathcal{B}$ onto the set of real numbers and satisfying certain axioms, which will not be further discussed at this point. If μ and ν are measures on the same measurable space, and μ is absolutely continuous with respect to ν, which means that both have the same null set, then there exists a function f such that

$$\mu(B) = \int_B f(x)\nu(dx), \quad B \in \mathcal{B}. \tag{3.1}$$

The function f is called density or, alternatively, Radon-Nikodym derivative of μ with respect to ν. Another notation for Eq. (3.1) would be

$$\mu(B) = \int_B \mu(dx), \tag{3.2}$$

which avoids using densities. If μ has integral one, it is called probability measure and f the corresponding probability density function. Furthermore, if the probability distribution is defined on the set of real numbers, then ν is just the familiar Lebesgue measure dx, and Eq. (3.1) becomes

$$P(a < X < b) = \int_a^b p(x)dx. \tag{3.3}$$

In this formula, the standard notation in terms of the random variable X was used, and P, p and $[a, b]$ replace μ, f and B, respectively. P can be understood as to specify the probability for X taking on a value in the range between a and b. For a distribution living on the unit sphere, ν is

spherical measure. Some examples of this type will be given at the end of this chapter, Sec. 3.5. In the discrete case, the counting measure applies, and equivalent to Eq. (3.3) one defines

$$P(X \in B) = \sum_B p(x). \tag{3.4}$$

In contrast to the continuous case, where density must be comprehended as probability per unit interval, here density can be interpreted directly as probability; a fact that is often emphasized by using the term *(probability) mass function* instead of density function. A probability distribution can also be related to a mixed measure, for example one that is composed of a discrete and continuous measure. Then the probability is

$$P(X \in A, Y \in B) = \sum_A \int_B p(x, y)dy. \tag{3.5}$$

Such a kind of mixed density will be defined later on to specify the joint density of the (discrete) number, the locations (being continuously distributed over the sphere) and the scaling coefficients of the base functions; the latter being just continuous on the real line. For the sake of brevity and because most of the following explanations can easily be transferred to the other types of distribution, I will subsequently restrict myself to the continuous setting.

A proper probability density function has to fulfill the conditions

$$p(x) \geq 0 \tag{3.6}$$

and

$$\int p(x)dx = 1, \tag{3.7}$$

which partly correspond to the above mentioned axioms. These conditions can likewise be formulated with respect to the distribution function, which has to be increasing and to satisfy $0 \leq F(x) \leq 1$.

Transforming a random variable does not change its probability distribution, but it changes the appearance of the corresponding density function. If the random variable X with distribution function $F(x)$ is transformed according to $y = g(x)$, in which g is strictly increasing and thus invertible, then the resulting random variable Y possesses the distribution function $F(h(y))$, wherein $h(y) = g^{-1}(y)$. Differentiation yields

$$\frac{dF(h(y))}{dy} = \frac{dF}{dh}\frac{dh}{dy} = p(h(y))\frac{dh}{dy} = p(y), \tag{3.8}$$

in which $p(h(y))$ and $p(y)$ are, respectively, the continuous density functions of X and Y, and h is continuously differentiable (Koch, 1999; Devroye, 1986). It should be noted that absolute value bars have to be added around the derivative when the transformation function is not strictly increasing but only injective. Eq. (3.8) can be generalized to multiple dimensions:

$$p(y_1, \ldots, y_n) = p(h_1(y_1, \ldots, y_n), \ldots, h_n(y_1, \ldots, y_n)) \; |\det \boldsymbol{J}\,| \tag{3.9}$$

with the Jacobian matrix

$$\boldsymbol{J} = \begin{bmatrix} \partial h_1/\partial y_1 & \partial h_1/\partial y_2 & \ldots & \partial h_1/\partial y_n \\ \partial h_2/\partial y_1 & \partial h_2/\partial y_2 & \ldots & \partial h_2/\partial y_n \\ \vdots & \vdots & \ddots & \vdots \\ \partial h_n/\partial y_1 & \partial h_n/\partial y_2 & \ldots & \partial h_n/\partial y_n \end{bmatrix}. \tag{3.10}$$

Eq. (3.3) can be generalized to

$$P(X_1 < x_1, \ldots, X_n < x_n) = F(x_1, \ldots, x_n) = \int_{-\infty}^{x_1} \ldots \int_{-\infty}^{x_n} p(x_1, \ldots, x_n) dx_1 \ldots dx_n, \tag{3.11}$$

which applies to a multidimensional continuous random variable summarized as *random vector* $\boldsymbol{x} = (X_1, \ldots, X_n)^T$ with values $\boldsymbol{x} = (x_1, \ldots, x_n)^T$ defined on the domain of real numbers, i.e. $\boldsymbol{x} \in \mathbb{R}^n$.

Suppose that one is not interested in the random vector $(\boldsymbol{x}_1, \boldsymbol{x}_2)$ with the joint density function $p(\boldsymbol{x}_1, \boldsymbol{x}_2)$ but only in the subset $\boldsymbol{x}_1$. The corresponding density for $\boldsymbol{x}_1$, the so-called *marginal density*, then follows by integration over $\boldsymbol{x}_2$:

$$p(\boldsymbol{x}_1) = \int p(\boldsymbol{x}_1, \boldsymbol{x}_2) d\boldsymbol{x}_2. \tag{3.12}$$

As is the case for the marginal density, the *conditional density* is a function of $\boldsymbol{x}_1$ only:

$$p(\boldsymbol{x}_1 | \boldsymbol{x}_2) = \frac{p(\boldsymbol{x}_1, \boldsymbol{x}_2)}{p(\boldsymbol{x}_2)}. \tag{3.13}$$

The conditional density is however subject to the condition that $\boldsymbol{x}_2$ takes on a particular value. From Eq. (3.13) also follows that if $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ are independent, their joint density can be written as product of their marginal densities:

$$p(\boldsymbol{x}_1, \boldsymbol{x}_2) = p(\boldsymbol{x}_1) p(\boldsymbol{x}_2). \tag{3.14}$$

3.2 Bayesian inference, point estimates & credible regions

In Bayesian statistics, any piece of information, e.g. the uncertainty about the outcome of a measurement or the value of a parameter, is formulated as probability density function. Bayes' theorem,

$$p(\boldsymbol{\beta} | \boldsymbol{y}) = \frac{p(\boldsymbol{y} | \boldsymbol{\beta}) p(\boldsymbol{\beta})}{p(\boldsymbol{y})}, \tag{3.15}$$

forms the basis for Bayesian inference. It is frequently written as proportionality relation:

$$p(\boldsymbol{\beta} | \boldsymbol{y}) \propto p(\boldsymbol{y} | \boldsymbol{\beta}) p(\boldsymbol{\beta}). \tag{3.16}$$

The prior probability density function $p(\boldsymbol{\beta})$ (in the following just prior) formalizes knowledge about the parameters available before measurements are taken. It is then modified by the sampling density $p(\boldsymbol{y}|\boldsymbol{\beta})$, which quantifies how well a parameter set can predict the given observations. The observations are fixed by measurement, whereas the unknown parameters are subject to adjustment. For this reason, the sampling density is commonly taken as a function of the unknowns and, in this form, denoted as likelihood function. The denominator of (3.15), $p(\boldsymbol{y}) = \int p(\boldsymbol{y}|\boldsymbol{\beta}) p(\boldsymbol{\beta}) d\boldsymbol{\beta}$, is referred to as the marginal likelihood or evidence. It simply acts as normalization constant and, since it does not depend on $\boldsymbol{\beta}$, is often ignored in studies where one is only interested in determining parameter values. However, it is the essential quantity when dealing with model comparison and model selection, as we will see later on in Sec. 4.7.1. The posterior $p(\boldsymbol{\beta}|\boldsymbol{y})$ is the target of Bayesian inference. Knowing the posterior makes it possible to derive all kinds of characteristic values, i.e. estimates for the unknown parameters or accuracy information.

In Bayesian statistics, parameter estimation is a problem of decision theory. Decisions are made on the numerical values of the parameters. Of course, these decisions are not arbitrary, but they

are evaluated by means of the resulting loss. Obviously, it is sensible to select values with minimal expected loss. If a quadratic loss function is applied, the decision is made on the Bayes estimate

$$\hat{\boldsymbol{\beta}}_B = \int \boldsymbol{\beta} p(\boldsymbol{\beta}|\boldsymbol{y}) d\boldsymbol{\beta}, \tag{3.17}$$

which is the expected value of the posterior. Correspondingly, the Bayes estimate for a function of the parameters reads

$$\hat{f}_B(\boldsymbol{\beta}) = \int f(\boldsymbol{\beta}) p(\boldsymbol{\beta}|\boldsymbol{y}) d\boldsymbol{\beta}. \tag{3.18}$$

A more robust loss function gives the MAP estimate

$$\hat{\boldsymbol{\beta}}_M = \underset{\boldsymbol{\beta}}{\operatorname{argmax}}\, p(\boldsymbol{\beta}|\boldsymbol{y}), \tag{3.19}$$

which is the mode, i.e. the value that maximizes the posterior. An estimate without accuracy information is only little meaningful. The posterior density provides full information and should be incorporated in further applications. Where this is not possible, e.g. for graphical display, credible or highest posterior density (HPD) regions are used. A HPD region is a subspace of the parameter space, $\mathcal{B}_\mathcal{S} \subset \mathcal{B}$, which contains the parameters with a given probability:

$$P(\boldsymbol{\beta} \in \mathcal{B}_\mathcal{S}|\boldsymbol{y}) = \int_{\mathcal{B}_\mathcal{S}} p(\boldsymbol{\beta}|\boldsymbol{y}) d\boldsymbol{\beta} = 1 - \alpha, \tag{3.20}$$

where the probability density of an inner point is higher or equal than that of an outer point:

$$p(\boldsymbol{\beta}_1|\boldsymbol{y}) > p(\boldsymbol{\beta}_2|\boldsymbol{y}) \text{ for } \boldsymbol{\beta}_1 \in \mathcal{B}_\mathcal{S}, \boldsymbol{\beta}_2 \notin \mathcal{B}_\mathcal{S}. \tag{3.21}$$

The calculation of characteristic values involves mathematical operations which are most of the time not analytically feasible, e.g. integration for the Bayes estimate or the search for extremal values in case of the MAP estimate. A special case is the linear problem with data and prior knowledge being normally distributed, for which the analytic solution is possible and presented in the following section 3.3. The general approach, however, is to sample from the posterior, i.e. to create random values from its density function, and to approximate the estimates numerically. For example, the estimates (3.17) to (3.19) become

$$\hat{\boldsymbol{\beta}}_B = \frac{1}{N} \sum_i \boldsymbol{\beta}_i \tag{3.22}$$

$$\hat{f}_B(\boldsymbol{\beta}) = \frac{1}{N} \sum_i f(\boldsymbol{\beta}_i) \tag{3.23}$$

$$\hat{\boldsymbol{\beta}}_M = \underset{\boldsymbol{\beta} \in \boldsymbol{\beta}^{(i)}}{\operatorname{argmax}}\, p(\boldsymbol{\beta}|\boldsymbol{y}). \tag{3.24}$$

3.3 The linear problem with Gaussian likelihood and prior

Let the observations $\boldsymbol{y}$ be distributed according to a normal distribution with unknown expected value $\boldsymbol{A\beta}$ and known covariance matrix $\boldsymbol{Q}_y$, i.e. $\boldsymbol{y}|\boldsymbol{\beta} \sim N(\boldsymbol{A\beta}, \boldsymbol{Q}_y)$. In this thesis, with a view on analyzing gradiometric data for an unknown gravity field (source) configuration, we will follow the custom in geodesy to consider the observations to be Gaussian. This is usually done because the total measuring error is a superposition of many elementary errors and thus, according to the central limit theorem, to a good approximation Gaussian distributed. Further suppose that also the

prior knowledge can be summarized by a normal distribution with expected value $\boldsymbol{\mu}_0$ and covariance matrix $\boldsymbol{Q}_0$, i.e. $\boldsymbol{\beta} \sim N(\boldsymbol{\mu}_0, \boldsymbol{Q}_0)$. For the above defined likelihood, this is a conjugate prior, which means that it leads to a posterior of the same kind, as will be demonstrated in the following.

Applying the Bayes theorem, Eq. (3.15), to the density functions of likelihood and prior,

$$p(\boldsymbol{y}|\boldsymbol{\beta}) \propto \exp\{-1/2\,[(\boldsymbol{y}-\boldsymbol{A\beta})^T\boldsymbol{Q}_y^{-1}(\boldsymbol{y}-\boldsymbol{A\beta})]\} \tag{3.25}$$

$$p(\boldsymbol{\beta}) \propto \exp\{-1/2\,[(\boldsymbol{\beta}-\boldsymbol{\mu}_0)^T\boldsymbol{Q}_0^{-1}(\boldsymbol{\beta}-\boldsymbol{\mu}_0)]\}, \tag{3.26}$$

we find the following expression for the posterior:

$$p(\boldsymbol{\beta}|\boldsymbol{y}) \propto \exp\{-1/2\,[(\boldsymbol{y}-\boldsymbol{A\beta})^T\boldsymbol{Q}_y^{-1}(\boldsymbol{y}-\boldsymbol{A\beta}) + (\boldsymbol{\beta}-\boldsymbol{\mu}_0)^T\boldsymbol{Q}_0^{-1}(\boldsymbol{\beta}-\boldsymbol{\mu}_0)]\}. \tag{3.27}$$

The inner part of the exponent can be rewritten as

$$\begin{aligned}
&(\boldsymbol{y}-\boldsymbol{A\beta})^T\boldsymbol{Q}_y^{-1}(\boldsymbol{y}-\boldsymbol{A\beta}) + (\boldsymbol{\beta}-\boldsymbol{\mu}_0)^T\boldsymbol{Q}_0^{-1}(\boldsymbol{\beta}-\boldsymbol{\mu}_0) \\
&\quad= \boldsymbol{y}^T\boldsymbol{Q}_y^{-1}\boldsymbol{y} + \boldsymbol{\mu}_0^T\boldsymbol{Q}_0^{-1}\boldsymbol{\mu}_0 + \boldsymbol{\beta}^T(\boldsymbol{A}^T\boldsymbol{Q}_y^{-1}\boldsymbol{A} + \boldsymbol{Q}_0^{-1})\boldsymbol{\beta} - 2\boldsymbol{\beta}^T(\underbrace{\boldsymbol{A}^T\boldsymbol{Q}_y^{-1}\boldsymbol{y} + \boldsymbol{Q}_0^{-1}\boldsymbol{\mu}_0}) \\
&\qquad\qquad\qquad\qquad\qquad\qquad = (\boldsymbol{A}^T\boldsymbol{Q}_y^{-1}\boldsymbol{A} + \boldsymbol{Q}_0^{-1})\boldsymbol{\mu} = \boldsymbol{Q}^{-1}\boldsymbol{\mu} \\
&\quad= \boldsymbol{y}^T\boldsymbol{Q}_y^{-1}\boldsymbol{y} + \boldsymbol{\mu}_0^T\boldsymbol{Q}_0^{-1}\boldsymbol{\mu}_0 + (\boldsymbol{\beta}-\boldsymbol{\mu})^T\boldsymbol{Q}^{-1}(\boldsymbol{\beta}-\boldsymbol{\mu}) - \boldsymbol{\mu}^T\boldsymbol{Q}^{-1}\boldsymbol{\mu},
\end{aligned} \tag{3.28}$$

where the quantities $\boldsymbol{\mu}$ and $\boldsymbol{Q}$ were introduced as abbreviations. As the posterior density is a function of $\boldsymbol{\beta}$ only, any term being independent of $\boldsymbol{\beta}$ is constant and can be neglected. The resulting posterior,

$$p(\boldsymbol{\beta}|\boldsymbol{y}) \propto \exp\{-1/2\,[(\boldsymbol{\beta}-\boldsymbol{\mu})^T\boldsymbol{Q}^{-1}(\boldsymbol{\beta}-\boldsymbol{\mu})]\}, \tag{3.29}$$

is obviously Gaussian, i.e. $\boldsymbol{\beta}|\boldsymbol{y} \sim N(\boldsymbol{\mu}, \boldsymbol{Q})$ with expected value and covariance matrix according to

$$\boldsymbol{\mu} = (\boldsymbol{A}^T\boldsymbol{Q}_y^{-1}\boldsymbol{A} + \boldsymbol{Q}_0^{-1})^{-1}(\boldsymbol{A}^T\boldsymbol{Q}_y^{-1}\boldsymbol{y} + \boldsymbol{Q}_0^{-1}\boldsymbol{\mu}_0) \tag{3.30}$$

$$\boldsymbol{Q} = (\boldsymbol{A}^T\boldsymbol{Q}_y^{-1}\boldsymbol{A} + \boldsymbol{Q}_0^{-1})^{-1}. \tag{3.31}$$

In accordance with the definition, the Bayes estimator is given by the expected value. For the specific problem considered in this section, the MAP estimator appears to be equal to the Bayes estimator, which is due to the symmetry of the normal distribution (Koch, 2007, p. 104). When associating $\boldsymbol{Q}_0^{-1}$ with the matrix $\boldsymbol{R}$, which has been introduced in Sec. 2.5.3 as initially arbitrary regularization matrix, and $\boldsymbol{\mu}_0$ with the zero vector, it becomes obvious that also the method of least-squares from traditional statistics comes to exactly the same result. Yet, the meaning differs on a philosophical basis.

3.4 Random sampling algorithms

3.4.1 (Inverse) transform sampling

Transform sampling (Press et al., 2007, pp. 362–363; Koch, 2007, pp. 194–196) is based on the fact that the density of a probability distribution is not invariant against a transformation. If realizations of a distribution are required, for which a random generator is not already implemented, one can instead draw samples from an easier distribution and transform the outcomes so that they become members of the desired distribution. The crucial point is, however, to find an appropriate transformation formula.

Consider the case that $p(x)$ and $p(y)$ are respectively the density functions of the sampling distribution and the target distribution, and the former is chosen to be uniform, i.e. $p(x) = 1$, then the transformation rule (3.9) simplifies to

$$p(y) = \frac{dx}{dy}. \tag{3.32}$$

Integration leads to

$$F(y) = x, \tag{3.33}$$

and inversion finally gives

$$y = F^{-1}(x). \tag{3.34}$$

The functional values of the distribution function $F(y)$ are obviously uniformly distributed. The other way round, applying the inverse distribution function to the uniform random numbers gives samples of the desired distribution. This particular case of transform sampling is referred to as *inverse transform sampling*. The prerequisite for using this technique is the ability to compute the inverted distribution function either analytically or at least numerically. This does not pose any problem in case of a discrete distribution. In general, however, integration of the density function is not feasible, which is particularly true for multivariate distributions. Then rejection sampling might be an alternative.

3.4.2 Rejection sampling

The *rejection sampling* algorithm (Press et al., 2007, pp. 365–367; Koch, 2007, p. 196) distributes points uniformly over the area under the graph of a density function and takes their x-values as realizations of the corresponding distribution. As a result, the number of samples that fall into a specific interval is proportional to the area as it should be because the area is equal to the probability by definition.

To get uniform random points from the area under the target density $p(x)$, random values x and u are generated from the proposal distribution with density $f(x)$ and from the uniform distribution $U(0, 1)$, respectively. The combined samples $(x, uf(x))$ represent random points from under the proposal. From those, every sample is discarded that does not also lie under the target density. In other words, a potential new sample is accepted if

$$uf(x) < p(x) \tag{3.35}$$

and otherwise rejected. It should be pointed out that neither $f(x)$ nor $p(x)$ has to be a density function in the strict sense. It is sufficient if they are only known up to a normalization constant. Yet, the proposal must be at least as high as the target density over the entire domain; otherwise, some regions would not be sufficiently covered. See Fig. 3.1 for an illustration of the procedure.

Rejection sampling is more generally applicable than inverse transform sampling, as it applies irrespective of whether or not the cumulative distribution function is known. It might, however, need a lot of rejections before one sample is accepted. The rate of acceptance is related to the area ratio between target density and proposal. If the target density has a complicated form, one might have difficulties to find a proposal that is a good envelope, and the algorithm becomes rather inefficient. The situation gets even worse when rejection sampling is applied to a multivariate distribution. This is because the acceptance probability decreases exponentially with the dimension of the problem. In those cases, the simulation of a Markov chain might be more advisable.

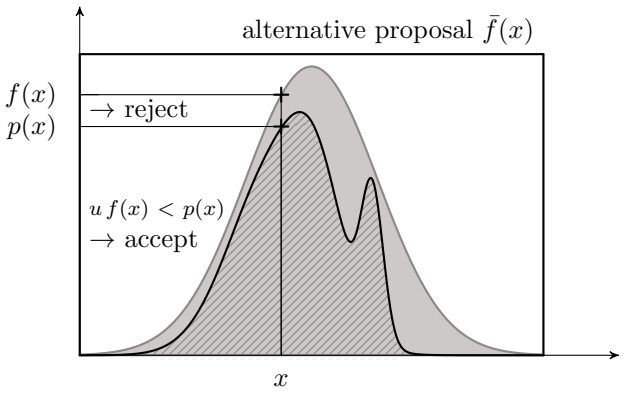


Figure 3.1: Illustration of rejection sampling: points are generated uniformly distributed under the proposal (grayish colored area). They are accepted or rejected according to a simple decision rule so that the remainder of the points is uniformly distributed under the target density (hatched area). The proposal must lie over the target density and should ideally build a close envelope. In the picture, this is obviously better fulfilled by the proposal $f(x)$ than by the rectangular density $\bar{f}(x)$.

3.4.3 Sampling by the simulation of a Markov chain

Markov chain Monte Carlo (MCMC) denotes a class of algorithms that gain information about a probability distribution from the simulation of a Markov chain. Making only few demands on the target distribution, these algorithms are universally applicable even to high-dimensional distributions. Broadly speaking, MCMC consists of randomly changing the actual state of the chain, thereby defining a new state, which then again is modified and so on. In this, it is similar to global optimization algorithms like the evolutionary or genetic algorithms. The difference is, however, that the individual samples do represent not only the way to the maximum but the entire probability distribution. In this way, we do not only get a point estimate but also stochastic information. In contrast to the elementary sampling algorithms such as inverse-transform or rejection sampling, the samples resulting from a Markov chain are correlated. Depending on the problem, it might therefore be necessary to reduce the correlations.

In order to really understand how MCMC works, some technical terms have to be clarified first. A comprehensive reference for the technical details is Geyer (2005). First of all, a Markov chain denotes a discrete-time stochastic process, $X_1, \ldots, X_n$, defined on an arbitrary state space that has the Markov property and stationary transition probabilities. Thus, the conditional probability of going to X_{n+1} having visited $X_n, \ldots, X_1$ first does only depend on X_n and second is constant. Next, it should be specified what a stationary Markov chain is. For a stochastic process to be stationary, remember from Sec. 3.1 that the joint distribution of $X_n, X_{n+1}, \ldots, X_{n+k}$ must not depend on n. The joint distribution can be written in terms of the marginal times the conditionals, and for a Markov chain the latter is constant by definition. So for a Markov chain to be stationary, it suffices when the distribution of X_n does not depend on n. In other words, the underlying distribution does not change, which gives it the name invariant, equilibrium or also simply stationary distribution. If we now simulate a stationary Markov chain, then the individual states of the chain can be taken as realizations of the invariant distribution. And if we achieve to simulate a Markov chain whose invariant distribution is equal to the target distribution, we finally get the desired samples. But how can we simulate a Markov chain with a specific invariant distribution? The general procedure is to repeatedly apply an update mechanism, which changes the state of the chain according to a certain

transition probability. So the problem boils down to finding an update with transition probability kernel P that does not change the target distribution π when being applied to it:

$$\int \pi(dx)P(x,B) = \pi(B). \tag{3.36}$$

This is the so-called equilibrium equation. Satisfying Eq. (3.36), P is said to preserve π. In the above equation, the general notation in terms of probability measures was used in the way it had been introduced in Sec. 3.1. Of course, in the case that P was simply a discrete probability distribution, one could just use the corresponding probability mass function. But as we will see in the following two sections, the transition kernel might become very complex, and it is not always possible to express it in terms of a density function. So the more general notation seems to be appropriate. Further note that, although the transition kernel being a conditional probability, we wrote $P(x,B)$ instead of $P(B|x)$, which is in accordance with the literature from the field of Markov chain theory.

To simplify the search for an adequate transition kernel, one often defines the Markov chain to be reversible. This means that the chain could just as well be run in opposite direction without changing the underlying distribution. Mathematically, this is expressed by

$$\int_A \pi(dx)P(x,B) = \int_B \pi(dx')P(x',A), \tag{3.37}$$

which is the so-called detailed balance condition in a continuous version. If P is reversible with respect to π, then P also preserves π. This becomes immediately clear when the integral is built over the whole state space, which again yields Eq. (3.36). So one can equivalently look for a kernel that fulfills detailed balance, Eq. (3.37), which in general is easier to show.

Fortunately, there are already adequate update mechanisms available, e.g. the Gibbs or the Metropolis-Hastings update, so that one does not have to seek for it oneself. In the following section, we will take a look at the Metropolis-Hastings algorithm because it has recently been extended to distributions of variable dimension.

3.4.4 The Metropolis-Hastings update

In the course of the Metropolis-Hastings update, samples are generated from a proposal distribution. Some samples will thus appear too often with respect to the target distribution, while others appear not often enough. To regulate this imbalance, a proposed sample is not readily accepted. Instead, an acceptance probability is introduced according to which the proposals are decided on. So the whole transition probability kernel of the Metropolis-Hastings update takes this form:

$$P(x,B) = r(x)I(x,B) + \int_{B^*} q(x,x'^*)\alpha(x,x')\mu(dx'^*). \tag{3.38}$$

P is the probability of going from x to a state inside the region B. q is the (normalized) proposal density; it denotes the probability to propose the step and α to accept it. Furthermore, we must also consider the probability that the sample is rejected independent of what has been proposed,

$$r(x) = 1 - \int q(x,x'^*)\alpha(x,x')\mu(dx'^*), \tag{3.39}$$

while x is already an element of B. I is the identity kernel, which becomes one in this case. The notation x'^* was used to indicate that not necessarily the entire new state has to be generated. x' could be a completely new vector x'^*, but it could also be the vector x with just one new component x'^* (one-at-a-time-Metropolis-Hastings) or something in between. Note that depending on the dimension of the proposal, also the measure $\mu(dx'^*)$ is always an other one. The transition

kernel looks rather complicated, and now it is obvious why the notation in terms of probability measures is reasonable. When we insert the transition kernel, Eq. (3.38), into the integrated detailed balance condition, Eq. (3.37), and assume that the target distribution π has density p, which is not necessarily normalized, we get

$$\int_A p(x) \int_{B^*} q(x, x'^*)\alpha(x, x')\mu(dx'^*)\mu(dx) + \int_{A\cap B} p(x)r(x)\mu(dx) \\ = \int_B p(x') \int_{A^*} q(x', x^*)\alpha(x', x)\mu(dx^*)\mu(dx') + \int_{A\cap B} p(x')r(x')\mu(dx'). \quad (3.40)$$

The last term on both sides is equal, so that Eq. (3.40) reduces to

$$\int_A \int_{B^*} p(x)q(x, x'^*)\alpha(x, x')\mu(dx'^*)\mu(dx) = \int_B \int_{A^*} p(x')q(x', x^*)\alpha(x', x)\mu(dx^*)\mu(dx'). \quad (3.41)$$

At this point, the target distribution in the forward and backward step is defined on the same domain and has a density with respect to the same underlying measure. The same is true for the proposal distribution. This means that $\mu(dx) = \mu(dx')$, and $\mu(dx^*) = \mu(dx'^*)$. Under these conditions, the acceptance probability that maintains detailed balance can easily be deduced from Eq. (3.41):

$$\alpha(x, x') = \min(1, R) \quad (3.42)$$

with the so-called odds ratio

$$R = \frac{p(x')q(x', x^*)}{p(x)q(x, x'^*)}. \quad (3.43)$$

In summary, the whole algorithm works as follows: a potentially new state is generated from the proposal distribution. The proposed state is then compared to the last state of the chain on the basis of the odds ratio. If this is larger than one, the proposal is adopted as the new state of the Markov chain. If it is less than one, the proposal is accepted with the probability given by the odds ratio or otherwise rejected. In the case of rejection, the new state of the chain is set equal to the old state.

The algorithm goes back to Metropolis et al. (1953). This early version made use of a symmetric proposal distribution, so that the proposal density values of the forward and backward step cancel each other out in the odds ratio. The odds ratio then simply becomes the ratio of the target density values. The algorithm in this form is also designated as random walk Metropolis. Later on in 1970, Hastings generalized the approach to apply for arbitrary proposal distributions. Even proposals being independent of the current state were possible, which earned it the name independent walk. The most recent version was published in Green (1995). It is a very general formulation, which also includes the two aforementioned earlier versions. The so-called Metropolis-Hastings-Green algorithm will be subject of the next section, Sec. 3.4.5.

3.4.5 The Metropolis-Hastings-Green update

In the odds ratio in the form it was presented in the last section, Eq. (3.43), subsequent states of the chain are compared by means of their density values. But this fails when the states have different dimensions. Then the measures in Eq. (3.41) are all different and do not vanish in the ratio anymore. In summary, comparing probability distributions of different dimension on the basis of probability densities does not make any sense. Instead, one has to make use of probability measures. In a

nutshell, "Metropolis-Hastings-Green is just like Metropolis-Hastings except that measures replace densities", to use the words of Geyer (2005). Consequently, the odds ratio must be written as

$$R = \frac{\pi(dx')Q(x', dx)}{\pi(dx)Q(x, dx')}. \tag{3.44}$$

It is, however, not immediately clear how to evaluate the probability measures in practical calculations. To get around this, Green assumed in his seminal work from 1995 that the product measure of the individual measures of the target and the proposal distribution is symmetric for the forward and backward step. This symmetric product measure does cancel when building the ratio, so we end up with the odds ratio

$$R = \frac{p(x')q(x', x)}{p(x)q(x, x')} \tag{3.45}$$

again, which only contains ordinary density functions. But to apply this formula, the prerequisite is to ensure that the product measure is symmetric—more on this later on in this section.

Bayes inference is most of the time not analytically feasible. The way out is to draw samples and to determine the parameters numerically. Frequently, MCMC techniques are employed for this purpose. Then the posterior density is associated with the target density of the Markov chain, and the sought-for parameters are the states of the chain. Simulating a Markov chain yields a sample of parameters, from which the best in some sense is selected as final estimate. In this context, the Metropolis-Hastings-Green algorithm presented in this section is of particular interest, because it allows for model determination, where different models in general have different dimensions. The model is simply treated as further unknown parameter. The chain wanders around the parameter space visiting different parameter combinations and, at the same time, jumps between different dimensional models. These jumps are reversible by definition, which earned the algorithm the name reversible jump Markov chain Monte Carlo (RJMCMC). Using the same notation as Green, which is M_k for the model with k being the model identifier and θ_k for the corresponding parameter vector, the odds ratio can be written in more detail:

$$R = \frac{p(k', \theta'_{k'}|y)q(k', k)q(\theta'_{k'}, \theta_k)}{p(k, \theta_k|y)q(k, k')q(\theta_k, \theta'_{k'})}. \tag{3.46}$$

$q(k, k')$ is the probability to propose a move from M_k to $M_{k'}$[1]. If there is more than just one move type available to realize this transition, the probability to choose the specific move type must be multiplied into this probability. $q(\theta_k, \theta'_{k'})$ is the proposal for the parameter vector. The selection of the proposal distribution is crucial to establish the above-mentioned symmetry between the steps. For example, when in the forward step the proposal is made to drop one of the model parameters, then in the backward step it must be added again. This would fulfill what Green also calls the dimension matching criterion. Often dimension matching can easily be realized, which is also true for the move types presented in the course of this thesis. But there are other moves that cannot easily be put into formulas. If for example the dimension of a model shall be augmented by one, and thus one random number is generated, by means of which more than just one variable is modified as is e.g. the case for the split-and-merge move, then this cannot any longer be expressed by the proposal. Instead, Green suggested an alternative way of writing. He directly introduces the density $p(u)$ (respectively $p(u')$ for the backward step), from which the random values are actually generated.

[1] Although it might be intuitive to think of $p(k, k')$ as the proposal for the model, it actually is a so-called mixing probability. As one of his innovations, Green showed that the mixing probability is permitted to depend on the actual state of the chain, i.e. on (k, θ_k). In this thesis, a possible dependence on the model parameters θ_k is not taken into consideration and thus neglected.

As a result, the corresponding measures would not cancel any more and must be taken into account in the odds ratio:

$$R = \frac{p(k', \theta'_{k'}|y)q(k', k)q(u')}{p(k, \theta_k|y)q(k, k')q(u)} \left| \frac{\partial(\theta'_{k'}, u')}{\partial(\theta_k, u)} \right|. \tag{3.47}$$

The last term is the Jacobi determinant for the transformation from (θ_k, u) to $(\theta'_{k'}, u')$. In this way of writing, the dimension matching criterion is as easy as $\dim(\theta_k) + \dim(u) = \dim(\theta'_{k'}) + \dim(u')$.

3.5 Probability distributions on the line and unit sphere

3.5.1 Discrete and continuous uniform distribution

The *discrete uniform distribution* assigns equal probabilities to all possible realizations x of a random variable:

$$p(x) = \begin{cases} \dfrac{1}{b-a+1} & \text{for } x \in \{a, a+1, \ldots, b\} \\ 0 & \text{otherwise} \end{cases} \tag{3.48}$$

with a and b being integers and $a < b$. The domain is commonly chosen to be the set of integers, as was also done here, though there might have been other options.

The *continuous uniform distribution* is characterized by a density function that is constant over a specific range of the real line:

$$p(x) = \begin{cases} \dfrac{1}{b-a} & \text{for } x \in [a, b] \\ 0 & \text{otherwise} \end{cases} \tag{3.49}$$

with the values x and the parameters a, b being real numbers and $a < b$ (Koch, 2007, p. 20). In the following, the continuous uniform distribution is shortly denoted as $U(a, b)$.

Uniform random number generators are available in any programming language. For details concerning their functionality, see Press et al. (2007, pp. 341–358).

3.5.2 Normal distribution

The random vector $\boldsymbol{x}$ with values $\boldsymbol{x} \in \mathbb{R}^n$ has *multivariate normal distribution* if its density function is given by

$$p(\boldsymbol{x}) = \frac{1}{(2\pi)^{n/2}(\det \boldsymbol{Q})^{1/2}} \exp\{-\frac{1}{2}[(\boldsymbol{x} - \boldsymbol{\mu})' \boldsymbol{Q}^{-1} (\boldsymbol{x} - \boldsymbol{\mu})]\} \tag{3.50}$$

with the vector of expected values $\boldsymbol{\mu}$ and the positive definite covariance matrix $\boldsymbol{Q}$ (Koch, 2007, p. 51). In what follows, the normal distribution is shortly denoted by $N(\boldsymbol{\mu}, \boldsymbol{Q})$. For the *univariate normal distribution*, the density function simplifies to

$$p(x) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\{-\frac{1}{2}[(x - \mu)^2/\sigma^2]\} \tag{3.51}$$

with the positive standard deviation σ.

An easy approach to generate samples from the normal distribution is the Box-Muller algorithm (Box and Muller, 1958; Press et al., 2007, p. 364). The algorithm relies on the strategy of transform sampling. Uniform random numbers x_1, x_2 are transformed to quantities y_1, y_2 by

$$y_1 = \sqrt{-2 \ln x_1} \cos 2\pi x_2 \tag{3.52}$$

$$y_2 = \sqrt{-2 \ln x_1} \sin 2\pi x_2. \tag{3.53}$$

Applying the transformation law for probability distributions, Eq. (3.9), yields

$$p(y_1, y_2) = \left[\frac{1}{\sqrt{2\pi}} \exp\left(-y_1^2/2\right) \right] \left[\frac{1}{\sqrt{2\pi}} \exp\left(-y_2^2/2\right) \right]. \tag{3.54}$$

The quantities y_1, y_2 can thus be taken as independent realizations of the standard normal distribution. Note that there are faster alternatives to the Box-Muller algorithm, which e.g. avoid to evaluate the trigonometric expressions.

3.5.3 Cauchy distribution

The random variable X with values $x \in \mathbb{R}$ has Cauchy distribution if its pdf is given by

$$p(x) = \frac{\gamma}{\pi} \left(\frac{1}{(x - x_0)^2 + \gamma^2} \right) \tag{3.55}$$

with the location parameter x_0 and the positive scale parameter γ specifying the half width at half maximum.

A possible way to sample from the Cauchy distribution is based on the observation that the ratio of two independent normally distributed random variables is standard Cauchy distributed. Considering the Box-Muller transform, Eqs. (3.52) and (3.53), we get

$$x = \frac{y_2}{y_1} = \tan 2\pi\, x_2. \tag{3.56}$$

So x can be determined from the uniform random number x_2 by transformation.

The half-Cauchy distribution corresponds to the positive half of the Cauchy distribution centered at zero. It is frequently used as prior for the variance parameter in hierarchical models, as was also done in the present work.

3.5.4 Geometric distribution

The discrete random variable X with values x from the set of integers has geometric distribution if its pdf is given by

$$p(x) = \begin{cases} (1-p)^x p & \text{for } x \in \{0, 1, 2, \ldots\} \\ 0 & \text{otherwise} \end{cases} \tag{3.57}$$

(Devroye, 1986, p. 498); see Fig. 3.2 for an illustration. The tuning parameter has thereby to satisfy $0 < p \leq 1$. The density specifies the number x of failures that may occur before a success in a random experiment with probability p. This leads directly to the following sampling scheme: generate random numbers between 0 and 1, stop when a number occurs that is less than p, i.e. a success, and count the preceding unsuccessful trials. The resulting count can be taken as realization of the geometric distribution. For $p > 1/3$ the method is probably unbeaten in terms of efficiency (Devroye, 1986, p. 498). However, if p is small, many random numbers might be necessary for one

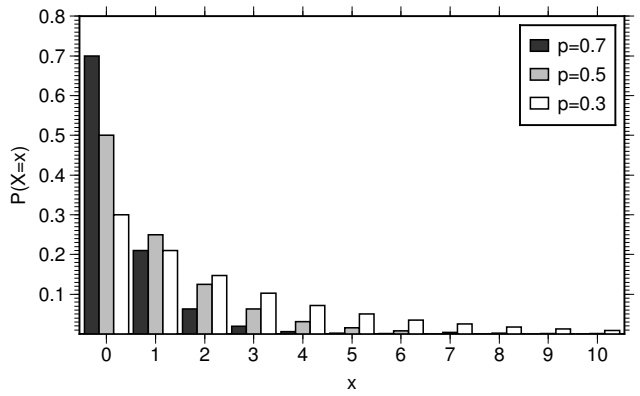


Figure 3.2: Probability mass function of the geometric distribution for three different values of the tuning parameter.

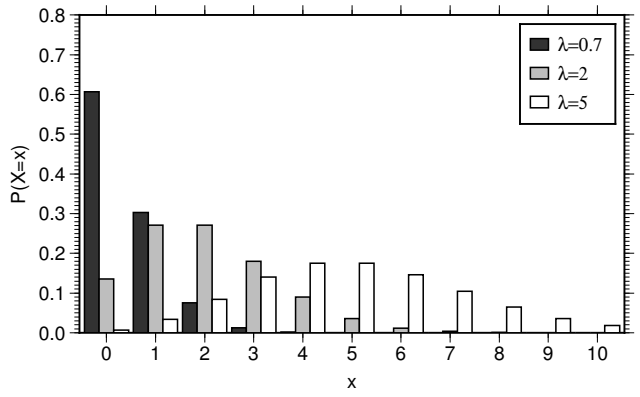


Figure 3.3: Probability mass function of the Poisson distribution for three different values of the tuning parameter.

single realization, and the method will thus decelerate considerably. Alternatively, inverse transform sampling can be applied, which requires one random number per realization only. Note that the geometric distribution is the discrete counterpart of the continuous exponential distribution. This becomes clear when inserting $p = 1-e^{-\lambda}$ into (3.57). Instead of applying inverse transform sampling to the geometric distribution, it would be more elegant to use it for the exponential distribution. Samples of the geometric distribution can afterwards be achieved by truncation to the integer part (Devroye, 1986, pp. 499–500).

3.5.5 Poisson distribution

The discrete random variable X with values x from the set of integers has Poisson distribution if its pdf is given by

$$p(x) = \begin{cases} \frac{\lambda^x}{x!} \exp\{-\lambda\} & \text{for } x \in \{0, 1, 2, \ldots\} \\ 0 & \text{otherwise} \end{cases} \tag{3.58}$$

(Devroye, 1986, p. 501); see Fig. 3.3 for an illustration. The parameter λ has to fulfill $0 < \lambda$. Values x can be obtained by standard inverse transform sampling (Devroye, 1986, p. 505). However, if λ is large, there might be more efficient ways of proceeding.

3.5.6 Spherical uniform distribution

The following distributions are defined on the two-dimensional unit sphere, S^2, which is the set of all points in three-dimensional Euclidean space that have a distance of one from the origin. A point of S^2 is specified by its Cartesian coordinates $\boldsymbol{x}$, where $\boldsymbol{x} \in \mathbb{R}^3$ and $|\boldsymbol{x}| = 1$. One can alternatively use spherical polar coordinates, (λ, ϑ, r), where $\lambda \in [0, 2\pi]$, $\vartheta \in [0, \pi]$ and $r = 1$. As the radius

is obviously always equal to one, the notation (λ, ϑ) is also possible. The two representations are related via

$$\boldsymbol{x} = \begin{pmatrix} \cos\lambda \sin\vartheta \\ \sin\lambda \sin\vartheta \\ \cos\vartheta \end{pmatrix}. \tag{3.59}$$

The above concept generalizes straightforwardly to arbitrary dimension. If needed, the sphere embedded in $\mathbb{R}^n$ is called S^m-sphere, where $m = n - 1$. Nevertheless, whenever the term 'sphere' appears in the following, it refers to the ordinary S^2-sphere.

The *spherical uniform distribution* is characterized by the fact that equal probabilities are assigned to equally sized area elements. Alternatively, one can also ask for a constant density with respect to the spherical area element. As stated earlier in Eq. (3.7), a probability density function has to have integral one. If the density function is initially set to an arbitrary constant, e.g. $p(\boldsymbol{x}) = 1$, then integration over the unit sphere,

$$\iint 1 \sin\vartheta d\lambda d\vartheta = 4\pi, \tag{3.60}$$

yields the normalization constant that is needed to derive the proper density function:

$$p(\boldsymbol{x}) = \frac{1}{4\pi}. \tag{3.61}$$

If in Eq. (3.60) the sin-term is assigned to the density function instead of being assigned to the area element, the density function is obtained in its most usual form (Mardia and Jupp, 1999, p. 160):

$$p(\lambda, \vartheta) = \frac{\sin\vartheta}{4\pi}. \tag{3.62}$$

Note that this step is actually a change of variables, which requires the application of the transformation law for probability distributions, Eq. (3.9). In fact, the additional sin-term represents the Jacobian determinant of the transformation from spherical to Cartesian coordinates, Eq. (3.59). Finally, there is a special version of the uniform distribution that is different from zero only within a limited part of the sphere. If this particular region is denoted by S and its area by A, then the density function for the area limited uniform distribution reads

$$p(\lambda, \vartheta) = \begin{cases} \dfrac{\sin\vartheta}{A} & \text{for } (\lambda, \vartheta) \in S \\ 0 & \text{otherwise} \end{cases} \tag{3.63}$$

(Fraiture, 2012). For a spherical rectangular and a spherical cap, A can be calculated analytically according to

$$A = (\lambda_{\max} - \lambda_{\min})(\cos\vartheta_{\min} - \cos\vartheta_{\max}) \tag{3.64}$$

and

$$A = 2\pi(1 - \cos\psi), \tag{3.65}$$

respectively. Here, λ_{min}, λ_{max}, ϑ_{min} and ϑ_{max} are used for the boundaries of the rectangle, and ψ is the opening angle of the spherical cap. For an arbitrarily shaped region, the approximate normalization constant can be achieved by numerical integration.

There are several approaches to sample from the spherical uniform distribution (e.g. von Neumann, 1951, Cook, 1957, Muller, 1959, Knop, 1970). The most intuitive approach might be that of von Neumann. He proposed to pick samples from inside the unit square and then to throw away those

that do not fall into the unit circle. The rest of the samples after normalization is distributed uniformly on the circle line. This idea also applies to any other dimension. Each of the approaches listed above is suitable if random points with global coverage are needed. Random points limited to a specific area can then be achieved by an additional rejection step. An algorithm that avoids rejection is described with some more detail in the following. The only source that was identified on that subject is Fraiture (2012). Two random variables are independent if their joint density can be written as product of the individual marginal densities, as was seen before in Eq. (3.14). As this is possible in the present case,

$$p(\lambda, \vartheta) = \frac{\sin \vartheta}{4\pi} = \frac{1}{2\pi} \frac{\sin \vartheta}{2} = p(\lambda)p(\vartheta), \tag{3.66}$$

λ and ϑ can be simulated individually. While λ can easily be generated from the uniform distribution $U(0, 2\pi)$, inverse transform sampling is applied for ϑ. This results in the transformation from the uniform random number u to the angle ϑ according to

$$\vartheta = \arccos(1 - 2u). \tag{3.67}$$

As the algorithm works with spherical coordinates, it can easily be adapted to apply for limited areas. For this purpose, λ is limited to the range $[\lambda_{min}, \lambda_{max}]$, and the term $(1 - 2u)$ in Eq. (3.67), which corresponds to a random number on the interval $[-1, 1]$ in the global case, is limited to $[\cos \vartheta_{max}, \cos \vartheta_{min}]$. If random points are required in an area that is not bounded by the great and small circles of longitude and latitude, the points can be generated in a bounding box surrounding the region of interest, and unwanted samples can just be thrown away. Alternatively, a complex algorithm that allows to sample directly from an arbitrarily shaped region can be found in Fraiture (2012).

3.5.7 Fisher distribution and approximations

The Fisher distribution on the sphere is the analogue to the isotropic, bivariate normal distribution in the plane (Kent, 1982). It is defined on the S^2-sphere and belongs to the general von Mises-Fisher distribution, which is valid for any dimension. The special about the von Mises-Fisher distribution is the linear argument of the exponential distribution, as will be seen later. Distributions with a more complex argument are summarized in the general Fisher-Bingham family. Besides the von Mises-Fisher distribution, this also includes the so-called Kent distribution, which is a distribution with elliptic isodensity lines.

The Fisher distribution belongs to the field of directional statistics, where it is used to model the uncertainties of directional data. In the geodetic community, the Fisher distribution has been of little interest so far. This might be surprising, as geodesists often have to deal with directional data. The reason is probably that the spherical normal distribution can be approximated by a planar normal distribution in case of small errors, which can normally be expected in geodetic applications. In fact, if the concentration parameter of the von Mises distribution goes towards infinity, then the von Mises distribution goes towards the normal distribution (Mardia and Jupp, 1999, p. 38). As there exists no geodetic literature on that subject, with the only exception being the publications of the group of Prof. Graferend from the University in Stuttgart (Grafarend and Awange, 2012; Cai and Grafarend, 2007), distributions of Fisher-type will be treated with some detail in this section.

The pdf of the Fisher distribution with respect to Cartesian coordinates reads

$$p(\boldsymbol{x}) = \frac{\kappa}{4\pi \sinh \kappa} \exp\{\kappa \boldsymbol{\mu}^T \boldsymbol{x}\}, \tag{3.68}$$

where $\boldsymbol{x}$ is a point of S^2, and $\boldsymbol{\mu}$ and κ are called mean direction and concentration parameter, respectively. Both $\boldsymbol{x}$ and $\boldsymbol{\mu}$ are unit vectors, and κ has to fulfill $\kappa \geq 0$. Introducing the spherical polar coordinates (λ, ϑ) for $\boldsymbol{x}$ and (λ_0, ϑ_0) for $\boldsymbol{\mu}$, we get

$$p(\lambda, \vartheta) = \frac{\kappa \sin \vartheta}{4\pi \sinh \kappa} \exp\{\kappa[\sin \vartheta \sin \vartheta_0 \cos(\lambda - \lambda_0) + \cos \vartheta \cos \vartheta_0]\}. \tag{3.69}$$

Ulrich (1984) published an approach, which was updated by Wood (1994) later on, to sample from a general class of distributions on the S^m-sphere. These distributions, which are not necessarily exponential but possess a linear argument, include the von Mises-Fisher distribution as a special case. Wood (1987) also published an approach for the Fisher-Bingham family restricted to the S^2-sphere. Both approaches made use of a decomposition in normal and tangential components. The same technique was also described by Mardia and Jupp (1999, p. 169) for the von Mises-Fisher distribution. As a result of the mentioned decomposition, the two components become independent, and the latter, i.e. the tangential component, is found to be uniformly distributed on the $S^{(m-1)}$-sphere. Simulation is thus straightforward: draw the normal component from the marginal density and the tangential component from the uniform distribution, and then combine both. In three dimensions, i.e. in case of the Fisher distribution, with the z-axis being oriented in direction to the mode, the angular coordinates λ and ϑ itselves can be interpreted as tangential and normal components. Further, the density function (3.69) simplifies to

$$p(\lambda, \vartheta) = \frac{\kappa \sin \vartheta}{4\pi \sinh \kappa} \exp\{\kappa \cos \vartheta\}, \tag{3.70}$$

where the above notation has been retained for simplicity. Integration over λ gives the marginal for ϑ,

$$p(\vartheta) = \frac{\kappa \sin \vartheta}{2 \sinh \kappa} \exp\{\kappa \cos \vartheta\}, \tag{3.71}$$

and dividing the joint density by this term leads to the conditional density for λ given ϑ,

$$p(\lambda) = \frac{1}{2\pi}, \tag{3.72}$$

which is actually the marginal for λ. Both quantities are obviously independent, and λ follows the uniform distribution, as was already announced earlier.

Samples of ϑ can be achieved by inverse transform sampling. This results in the following transformation:

$$\cos \vartheta = \frac{\log(-u(\exp\{\kappa\} - \exp\{-\kappa\} + \exp\{\kappa\}))}{\kappa} \tag{3.73}$$

with the uniform random number u. At this moment, the created random sample (λ, ϑ) is given in a local coordinate system. To get samples from the general Fisher distribution with an arbitrary oriented mean direction, one can rotate the local system back to the global one. Note that this transformation has Jacobian determinant one, so that the density function, Eq. (3.70), is not modified (see also Mardia and Jupp, 1999, p. 169). Further, expressing the old coordinates by the new coordinates yields the density of the general Fisher distribution, either Eq. (3.68) or (3.69).

Wenzel (2012) proposed modifications for both the density function and the sampling procedure. The reason is that in the original versions stability problems arise already for moderate values of the concentration parameter. Using the relation

$$\sinh \kappa = \frac{\exp\{\kappa\} - \exp\{-\kappa\}}{2}, \tag{3.74}$$

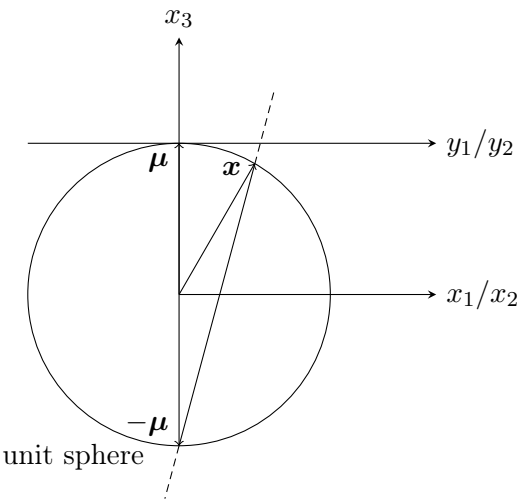


Figure 3.4: Illustration of the stereographic projection

one finds after some simple calculations an improved expression for the density function:

$$p(\lambda, \vartheta) = \frac{\kappa \sin \vartheta}{2\pi(1 - \exp\{-2\kappa\})} \exp\{\kappa[\sin \vartheta \sin \vartheta_0 \cos(\lambda - \lambda_0) + \cos \vartheta \cos \vartheta_0 - 1]\}. \tag{3.75}$$

Regarding the sampling process, Wenzel proposed to use

$$\cos \vartheta = 1 + \frac{1}{\kappa}\left(\ln u + \ln\left(1 - \exp\{-2\kappa\}\frac{u-1}{u}\right)\right) \tag{3.76}$$

instead of Eq. (3.73).

There is yet another possibility to obtain samples from the Fisher distribution. Suppose that $\boldsymbol{x}$ belongs to the Fisher distribution, and λ, ϑ are the spherical polar coordinates of $\boldsymbol{x}$. Then $\boldsymbol{y}$ with

$$y_1 = \cos \lambda \vartheta \tag{3.77}$$
$$y_2 = \sin \lambda \vartheta \tag{3.78}$$

is approximately normally distributed (Mardia and Jupp, 1999, pp. 172–173). Vice versa, if $\boldsymbol{y}$ is sampled from the bivariate normal distribution on the plane and transformed back by

$$\lambda = \operatorname{atan2}(y_1, y_2) \tag{3.79}$$
$$\vartheta = \sqrt{y_1^2 + y_2^2} \tag{3.80}$$

with

$$\operatorname{atan2}(y_1, y_2) = \begin{cases} \arctan \frac{y_2}{y_1} & \text{if } y_1 > 0 \\ \arctan \frac{y_2}{y_1} + \pi & \text{if } y_1 < 0, y_2 > 0 \\ \arctan \frac{y_2}{y_1} - \pi & \text{if } y_1 < 0, y_2 < 0, \end{cases} \tag{3.81}$$

then the point defined by the above angular coordinates is Fisher distributed. This holds true for highly concentrated distributions and small angles. The Eqs. (3.79) and (3.80) can be interpreted as wrapping of the tangential plane around the sphere.

Various other useful distributions arise by projection. Dortet-Bernadet and Wicker (2008), for example, use inverse stereographic projections of general multivariate normal distributions to model

clusters on the sphere. For reasons that will become clear later, I look at the isotropic bivariate normal distribution only.

The stereographic projection is basically a mapping from the sphere onto the plane. For a selected direction $\boldsymbol{\mu}$ and a point of the sphere $\boldsymbol{x}$, the stereographic projection is defined by the intersection of a straight line through $\boldsymbol{x}$ and the opposite pole $-\boldsymbol{\mu}$ with a plane orthogonal to $\boldsymbol{\mu}$. To simplify matters, I introduce a local coordinate system, whose z-axis is aligned with $\boldsymbol{\mu}$; for a sketch of the situation, see Fig. 3.4. I also use the tangential plane to the sphere at $\boldsymbol{\mu}$, i.e. at the 'north pole', as image plane. The projection is thus slightly different from that defined by Dortet-Bernadet and Wicker, who make use of the 'equatorial plane'. Applying the theorem of intersecting lines on the situation in Fig. 3.4, we get the formulas for the stereographic projection:

$$y_1 = \frac{2x_1}{1+x_3} = \frac{2\sin\vartheta\cos\lambda}{1+\cos\vartheta} \tag{3.82}$$

$$y_2 = \frac{2x_2}{1+x_3} = \frac{2\sin\vartheta\sin\lambda}{1+\cos\vartheta} \tag{3.83}$$

with $\boldsymbol{x}$ and $\boldsymbol{y}$ being corresponding points on the sphere and the plane respectively and λ, ϑ being the angular coordinates of $\boldsymbol{x}$. Inversion of the Eqs. (3.82) and (3.83) yields the inverse stereographic projection,

$$\lambda = \operatorname{atan2}(y_1, y_2) \tag{3.84}$$

$$\vartheta = 2\arctan\frac{\sqrt{y_1^2+y_2^2}}{2}, \tag{3.85}$$

which maps the plane onto the sphere. To obtain samples of the 'spherical normal distribution', one might come up with the idea of taking y_1 and y_2 from the planar normal distribution and projecting them back to the sphere by (3.84) and (3.85). The thus created distribution will obviously be slightly different from the Fisher distribution. Its density function can be derived by analysing the projection. The corresponding Jacobian matrix is

$$J = \begin{pmatrix} \frac{\partial y_1}{\partial\vartheta} & \frac{\partial y_1}{\partial\lambda} \\ \frac{\partial y_2}{\partial\vartheta} & \frac{\partial y_2}{\partial\lambda} \end{pmatrix} = \begin{pmatrix} \frac{2\cos\lambda}{1+\cos\vartheta} & \frac{-2\sin\vartheta\sin\lambda}{1+\cos\vartheta} \\ \frac{2\sin\lambda}{1+\cos\vartheta} & \frac{2\sin\vartheta\cos\lambda}{1+\cos\vartheta} \end{pmatrix} \tag{3.86}$$

with the Jacobian determinant

$$|J| = \frac{4\sin\vartheta}{(1+\cos\vartheta)^2}. \tag{3.87}$$

The final density is composed of the original density written in terms of the new coordinates and the Jacobian. With respect to Cartesian coordinates, it becomes

$$p(\boldsymbol{x}) = \frac{1}{2\pi\sigma^2}\exp\left\{-\frac{1}{2}[(2x_1/1+x_3)^2 + (2x_2/1+x_3)^2]/\sigma^2\right\}\frac{4}{(1+x_3)^2}. \tag{3.88}$$

Note that up to this point we are still working in a local coordinate system. To overcome this limitation, i.e. to allow the mean direction to be directed in any direction, the samples can be simply rotated, as was already described earlier in this chapter. As before, the density function is not affected by the rotation.

The central projection is very similar to the stereographic projection. The only difference is the projection center, which is no longer the point opposite to the mode but the center of the sphere. In formulas, the central projection reads

$$y_1 = \frac{x_1}{x_3} = \frac{\sin\vartheta\cos\lambda}{\cos\vartheta} \tag{3.89}$$

$$y_2 = \frac{x_2}{x_3} = \frac{\sin\vartheta\sin\lambda}{\cos\vartheta}. \tag{3.90}$$

The inverse central projection then follows to

$$\lambda = \operatorname{atan2}(y_1, y_2) \tag{3.91}$$

$$\vartheta = \arctan \sqrt{y_1^2 + y_2^2}. \tag{3.92}$$

Just as in the above case, random points of a distribution which is close to the Fisher distribution can be achieved by inverse central projection of y_1 and y_2, which belong to the bivariate normal distribution on the plane. Note that due to the choice of the projection center, the resulting points will cover the northern part of the sphere only. To derive the corresponding density function, again the Jacobian matrix is calculated:

$$J = \begin{pmatrix} \frac{\partial y_1}{\partial \vartheta} & \frac{\partial y_1}{\partial \lambda} \\ \frac{\partial y_2}{\partial \vartheta} & \frac{\partial y_2}{\partial \lambda} \end{pmatrix} = \begin{pmatrix} \frac{\cos \lambda}{\cos \vartheta^2} & \frac{-\sin \vartheta \sin \lambda}{\cos \vartheta} \\ \frac{\sin \lambda}{\cos \vartheta^2} & \frac{\sin \vartheta \cos \lambda}{\cos \vartheta} \end{pmatrix} \tag{3.93}$$

with the Jacobian determinant

$$|J| = \frac{\sin \vartheta}{\cos \vartheta^3}. \tag{3.94}$$

Written in terms of Cartesian coordinates, the resulting density function reads

$$p(\boldsymbol{x}) = \frac{1}{2\pi\sigma^2} \exp\{-\frac{1}{2}[(x_1/x_3)^2 + (x_2/x_3)^2]/\sigma^2\} \frac{1}{x_3^3}. \tag{3.95}$$

3.6 Discussion

3.6.1 Remark on random number generators

The today available random number generators differ considerably in terms of quality (Press et al., 2007, pp. 341–342). Some of them are actually outdated; others are over-engineered at least for typical applications. It is particularly warned against using the generators of the C++ Standard Library, as they are implementation-dependent and might have a relatively short return period. In my software, I decided to integrate the Boost Random Number Library[2], which offers a bundle of different random generators. I chose the solid generator by Matsumoto and Nishimura (1998), which is said to be both fast and sufficient in quality.

A random number generator requires an initial seed before use. Often simply the system time is utilized for this purpose. Note that in my work an iterative procedure was employed, which was implemented by running the same program repeatedly. This also means, however, that the random generator must be initialized more than just once, namely in each iteration. The system time seems to be no longer adequate, particularly for small problems with short computing time. One possible solution would be to increment the random generator seed in each iteration. Alternatively, one could remember the actual state of the generator at the end of one iteration and reuse it for the next. This is how the problem was tackled in my implementation. By avoiding the incrementation, one does not need to think about whether the sequences of random numbers initiated by the seeds i, $i + 1$, ... are truly independent. Note that the possibility to set a seed that is different from the system time is the only way to come to reproducible results, which I think are important for further analyses or for debugging.

[2] http://www.boost.org/

3.6.2 Remark on the sin-term

As written earlier in the Preliminaries, Sec. 3.1, changing the parametrization of a probability distribution only affects the appearance of the associated density function. The same is true for the transition from Cartesian to spherical polar coordinates or vice versa. The density function with respect to the differential in angular coordinates, $d\lambda d\vartheta$, looks different in that it has an additional $\sin\vartheta$-term. This term corresponds to the Jacobian determinant of the transformation from angular to Cartesian coordinates, as can be verified easily by means of the transformation law for probability distributions, Eq. (3.9). Remember that density can be understood as probability per unit interval. The deviation of the density function from unity can be explained by the fact that the area of a surface element that is caused by a uniform change in the angular coordinates varies over the globe. Nevertheless, for the major part of this work, it does not matter whether the one or the other density function is used. Attention must be paid when calculating the MAP estimator, as the point where a density function reaches its maximum is not invariant against transformation.

3.6.3 Probability distribution adapted to the gravity field

For the basis functions in regional gravity field analysis, the optimal locations are often assumed intuitively at places where the gravity field shows prominent structures or distinctive points. For example, Balmino (1972) located the basis functions under the extreme values taken from a map of gravity anomalies. To make available knowledge of the gravity field usable in this work, an additional type of spherical distribution was designed, which is adapted to the gravity field.

As stated earlier, a probability density function has to fulfill the conditions (3.6) and (3.7), namely it has to be positive and to integrate to one. It should further be easy to evaluate the density and to generate realizations of the corresponding distribution. The new distribution is constructed on the basis of a gravity field model, which can be evaluated continuously all over the globe. A density value is calculated by first evaluating the desired functional of the gravity field model at a specific point. Next, the absolute value is taken, and the result is normalized. An approximate normalization constant can be calculated by numerical integration. If a geographical grid with spacing $\Delta\lambda$, $\Delta\vartheta$ is chosen for discretization, the normalization constant becomes

$$\iint_S |f(\lambda,\vartheta)|\sin\vartheta d\lambda d\vartheta \approx \sum_i \sum_j |f(\lambda_i,\vartheta_j)|\Delta\lambda(2\sin\vartheta_j \sin\frac{\Delta\vartheta}{2}), \tag{3.96}$$

where S is the support of the density function, and f is the desired functional (geoid height, gravity anomaly, deflection of the vertical or similar). In summary, the density function of the new distribution reads

$$p(\lambda,\vartheta) = \begin{cases} \frac{|f(\lambda,\vartheta)|\sin\vartheta}{\sum_i\sum_j f(\lambda_i,\vartheta_j)\Delta\lambda(2\sin\vartheta_j\sin\frac{\Delta\vartheta}{2})} & \text{for } (\lambda,\vartheta)\in S \\ 0 & \text{otherwise.} \end{cases} \tag{3.97}$$

Samples can be obtained by rejection sampling with the spherical uniform distribution being used as proposal. The factor, which is needed to scale the proposal function to the level of the target density, can be determined jointly with the normalization constant.

4. Optimization of point grids in regional gravity field analysis

4.1 Description of the problem

In Ch. 2, the approach for regional gravity field analysis was presented that is used in the Astronomical, Physical and Mathematical Geodesy Group of Bonn University. In the standard form, a dense and uniform point grid is used to define the nodal points of the basis functions with an extra margin around the region of the observations (cf. Sec. 2.5.1). The present work aims at optimizing the nodal point grid.

Determining the locations of the basis functions jointly with their scaling coefficients represents a nonlinear problem. But if the positions are specified in advance, the problem becomes linear, and it can be written in simple matrix-vector notation. This is the case that has been considered so far (cf. Ch. 2). For this type of problem, where only a subset of the sought-for parameters is nonlinear, I use the term *quasi-linear*. This term is taken from Gundlich and Kusche (2008), who used Monte Carlo methods for the problem while solving a part of it analytically—an idea that was picked up again in the present work. Using the method of least-squares for nonlinear problems yields, except for few special cases, a non-convex objective function, i.e. a function with more than just one extreme value (Boyd and Vandenberghe, 2004). Local optimization strategies linearize the functional relationship at a certain point and this way try to approximate the non-convex by a convex function. However, this does not necessarily lead to the global optimum. The crucial point is whether there are sufficiently good approximate values available. As mentioned earlier, the optimization of the point positions will also affect the required number of basis functions. But the correct number will not be known in advance. We thus find ourselves in the unusual situation that the number of unknowns is one of the unknowns (Hastie and Green, 2012). Technically, this leads to a design matrix with a variable number of columns, which can not be dealt with by means of ordinary statistical tools.

To tackle these challenges, I use global optimization in the context of Bayesian statistics. Here sampling algorithms are used when the problem cannot be solved analytically. The amount of randomness helps to avoid sticking to local minima, making us less dependent on the approximate values. Moreover, as we know from Sec. 3.4.5, there is a solution available that can deal with the problem of variable dimension.

In the remainder of the section, it is described how reversible jumps are used for the optimization of point grids. For the sake of clear arrangement, the unknowns are collected in the following vectors

$$\boldsymbol{\beta}_1 = \begin{pmatrix} \vdots \\ a_k \\ \vdots \end{pmatrix}, \boldsymbol{\beta}_2 = \begin{pmatrix} \vdots \\ \boldsymbol{x}_k \\ \vdots \end{pmatrix}, \beta_3 = K. \tag{4.1}$$

4.2 The problem at the example of 8 hidden basis functions

A simple closed-loop simulation scenario will be used at different places throughout the chapter to visualize selected parts of the algorithm. The fact that the true RBF model is known facilitates the evaluation of the algorithm when compared to the work with real data or simulated data from a spherical harmonic model. The model consists of 8 RBFs, which were distributed across an area

of $15° \times 15°$. When selecting the locations and scaling coefficients of the basis functions, it was taken care that they are neither too close nor too small to be found again also in the presence of noise. The individual basis functions were set up from harmonic degree 30 to 250, and the shape coefficients were specified using the degree variances of the gravity field approximated by Kaula's rule of thumb. The model was forward simulated, and observations of the type radial gravity gradients were calculated (Eq. (2.42)), being close to the gravity gradients in zz-direction, though without the need for being rotated. The field was evaluated along a sub-cycle of the real GOCE orbit (1.11.2009–11.12.2009) using a 5 sec sampling interval. White noise with a standard deviation of 40 mE was added, which is indeed a multiple of the actual noise of the real gradients. However, since correlations were not taken into account, a higher value was considered to be useful. For the generated signal and the derived observations, see Figs. 4.1(a) and 4.1(b).

As motivated earlier in Sec. 1.1, I expect that adapting the model resolution in the form of the optimization of the point grid will reduce overfitting and thereby naturally leads to a stabilization of the problem, so that the solution is less affected by simplified assumptions in the prior information. This way of thinking should be illustrated again by means of the 8-point example. A set of basis functions was defined, and the scaling coefficients were calculated from the simulated observations in the usual way described in Ch. 2. The same kernel function was used as it was already done for the simulation of the data. The positions of the basis functions were specified in the area of the observations by means of the points of a regular triangular grid (triangular vertex, level 79); an additional margin was not considered because of the small signal in the edge region. In the course of solving the normal equations, a regularization term was taken into account, and the regularization parameter was estimated with the help of variance component estimation to about 0.007. The rms of the differences to the true solution in the study area is 0.37 m in terms of geoid heights (Tab. 4.1, see also Fig. 4.1(c)). For comparison, when the true nodal point grid was employed, the regularization parameter was estimated to 0.062, and the rms amounts to 0.032 m (Tab. 4.1, see also Fig. 4.1(d)). The model based on the standard grid thus performed worse by an order of magnitude. It is not that the model is not able to adequately represent the observations. As can be seen from the approximation results of error-free data in Tab. 4.1, the pure model error is very small (see also Fig. 4.1(e)). The actual reason is the ill-posedness of the problem. The model based on the standard grid reacts very sensitively to errors in the data. This is also clear from the high rms value that results from solving the normal equations without regularization (see Tab. 4.1). So using regularization is necessary, and because of the high uncertainty involved in the inversion of the normal equations, the solution is rather receptive for it, i.e. the prior information in this scenario has a high impact. With perfect prior knowledge of the signal in the respective area, one could basically also come to a perfect solution. However, by using the signal degree variances to describe the accuracy of the prior knowledge, one only prescribes a certain signal content per wavelength without any concrete spatial reference. In this way, possible spatial differences in the smoothness of the field can not be taken into consideration. By contrast, no regularization is required for the true grid, and when using regularization, the solution is comparatively resistant to changes of the regularization parameter (cf. the solution with the fixed variance component in Tab. 4.1).

4.3 The joint posterior density function

The joint distribution of the sought-for parameters can be described by the mixed density

$$p(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \beta_3|\boldsymbol{y}) \propto p(\boldsymbol{y}|\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \beta_3)p(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \beta_3) \tag{4.2}$$

$$\propto p(\boldsymbol{y}|\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \beta_3)p(\boldsymbol{\beta}_1|\boldsymbol{\beta}_2, \beta_3)p(\boldsymbol{\beta}_2|\beta_3)p(\beta_3). \tag{4.3}$$

Only in few simple cases, the posterior density can be derived from the above expression in proper form; more often, it has to be simulated. Although this does in principle also hold for our problem,

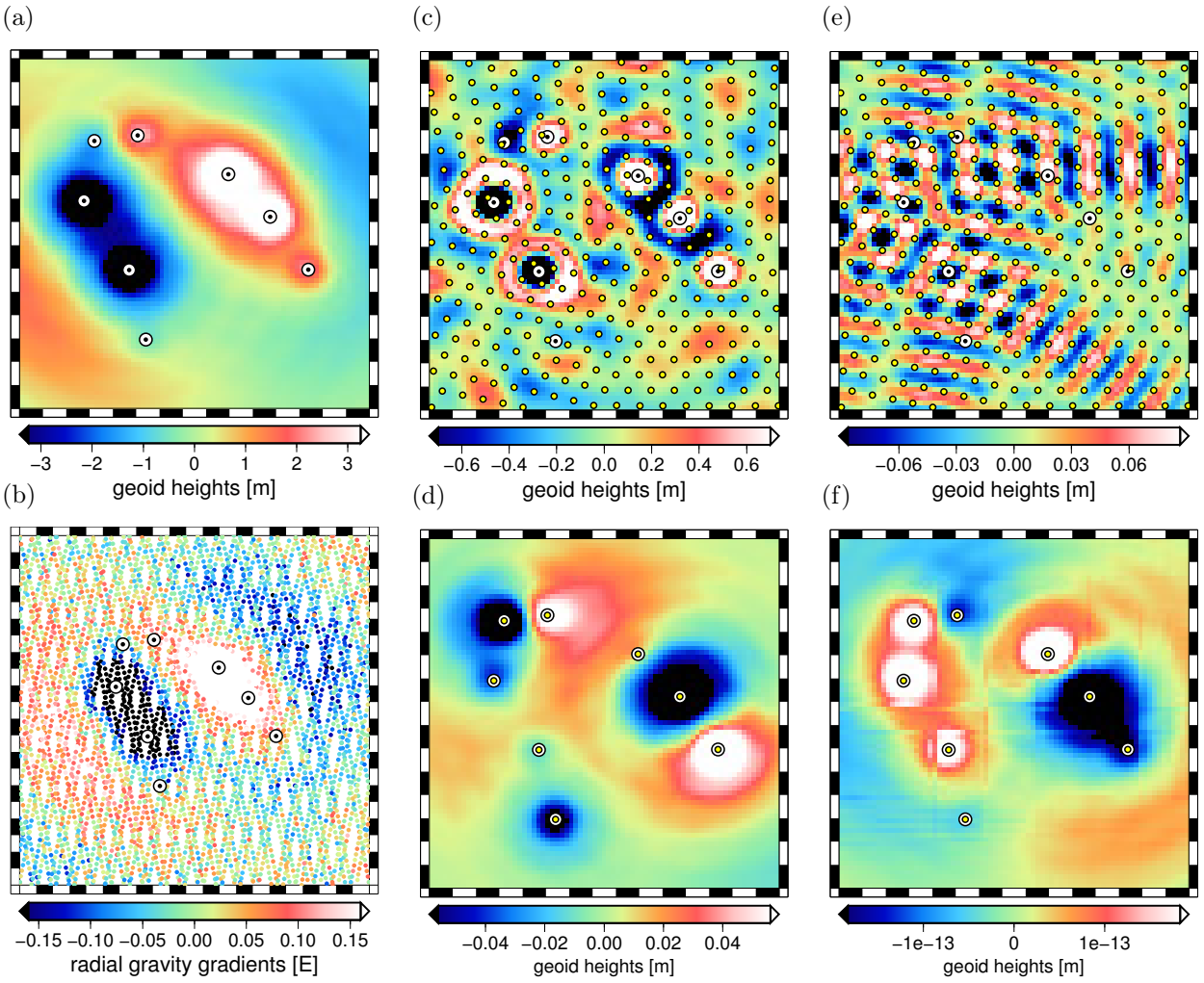


Figure 4.1: (a) Signal generated by the 8 hidden basis functions, (b) observations, (c) differences from the signal for the regular grid ($\sigma = 0.007$ with VCE), (d) differences for the true grid ($\sigma = 0.06$ with VCE), (e) differences for the regular grid (error free data), (f) differences for the true grid (error free data)

Table 4.1: Statistics for the 8-point example: rms of the differences in geoid heights between gravity field solutions based on different processing variants and the input field. Note that the modeling errors for the variant with the regular grid in the error free scenario could be further reduced down to $1e{-}5$ when choosing a larger, denser or simply other type of grid.

regular grid		**true grid**	
variant	**rms [m]**	**variant**	**rms [m]**
$\sigma = 0.007$ (with VCE)	0.37	$\sigma = 0.062$ (with VCE)	0.032
error free data	$1e{-}2$	error free data	$1e{-}13$
w/o regularization	7.75	w/o regularization	0.032
$\sigma = 0.062$ (fixed)	0.81	$\sigma = 0.007$ (fixed)	0.059

at least a part of it can be obtained analytically. Specifically, we can benefit from the quasi-linearity introduced in Sec. 4.1 because determining the scaling coefficients for a given point grid is a linear problem and, under certain conditions, the posterior density can be specified directly. So we do not have to sample from the scaling coefficients, but we can integrate them out and run the chain for the marginalized distribution. DiMatteo et al. (2001) proceeded in a similar way for the positioning of the basis functions in a spline approach. The marginal posterior is presented in the next section followed by a description of the individual quantities.

4.3.1 Integrating out the scaling coefficients

For the two first terms of Eq. (4.3), one can again apply Bayes' theorem:

$$p(\boldsymbol{y}|\boldsymbol{\beta}_1,\boldsymbol{\beta}_2,\beta_3)p(\boldsymbol{\beta}_1|\boldsymbol{\beta}_2,\beta_3) = p(\boldsymbol{\beta}_1|\boldsymbol{\beta}_2,\beta_3,\boldsymbol{y})p(\boldsymbol{y}|\boldsymbol{\beta}_2,\beta_3). \tag{4.4}$$

Introducing this in the original expression, it becomes clear that there is no term left that is conditional dependent on $\boldsymbol{\beta}_1$, so that it can be integrated out:

$$p(\boldsymbol{\beta}_2,\beta_3|\boldsymbol{y}) \propto p(\boldsymbol{y}|\boldsymbol{\beta}_2,\beta_3)p(\boldsymbol{\beta}_2|\beta_3)p(\beta_3). \tag{4.5}$$

The term on the left hand side is the marginal density for the point grid. When we want to work with this marginalized form, we have to know the marginalized likelihood. Under the usual assumptions of normality for the observations, $\boldsymbol{y}|\boldsymbol{\beta}_1,\boldsymbol{\beta}_2,\beta_3 \sim N(\boldsymbol{A}\boldsymbol{\beta}_1,\boldsymbol{Q}_y)$, and the prior knowledge, $\boldsymbol{\beta}_1|\boldsymbol{\beta}_2,\beta_3 \sim N(\boldsymbol{\mu}_{0\beta_1},\boldsymbol{Q}_{0\beta_1})$, we can, as the determination of the scaling coefficients for a given point grid is a linear problem, compute the corresponding density analytically:

$$p(\boldsymbol{y}|\boldsymbol{\beta}_1,\boldsymbol{\beta}_2,\beta_3) = \frac{1}{(2\pi)^{n/2}(\det \boldsymbol{Q}_y)^{1/2}} \exp\{-\frac{1}{2}[(\boldsymbol{y}-\boldsymbol{A}\boldsymbol{\beta}_1)^T\boldsymbol{Q}_y^{-1}(\boldsymbol{y}-\boldsymbol{A}\boldsymbol{\beta}_1)]\} \tag{4.6}$$

$$p(\boldsymbol{\beta}_1|\boldsymbol{\beta}_2,\beta_3) = \frac{1}{(2\pi)^{K/2}(\det \boldsymbol{Q}_{0\beta_1})^{1/2}} \exp\{-\frac{1}{2}[(\boldsymbol{\beta}_1-\boldsymbol{\mu}_{0\beta_1})^T\boldsymbol{Q}_{0\beta_1}^{-1}(\boldsymbol{\beta}_1-\boldsymbol{\mu}_{0\beta_1})]\}. \tag{4.7}$$

For reasons of clarity, we set $\boldsymbol{A} = \boldsymbol{A}_{\beta_2,\beta_3}$, $\boldsymbol{\mu}_{0\beta_1} = \boldsymbol{\mu}_{0\beta_1|\beta_2,\beta_3}$, and $\boldsymbol{Q}_{0\beta_1} = \boldsymbol{Q}_{0\beta_1|\beta_2,\beta_3}$. Combining Eq. (4.6) and Eq. (4.7) yields

$$\begin{aligned}
&(\boldsymbol{y}-\boldsymbol{A}\boldsymbol{\beta}_1)^T\boldsymbol{Q}_y^{-1}(\boldsymbol{y}-\boldsymbol{A}\boldsymbol{\beta}_1) + (\boldsymbol{\beta}_1-\boldsymbol{\mu}_{0\beta_1})^T\boldsymbol{Q}_{0\beta_1}^{-1}(\boldsymbol{\beta}_1-\boldsymbol{\mu}_{0\beta_1}) \\
&= \boldsymbol{y}^T\boldsymbol{Q}_y^{-1}\boldsymbol{y} + \boldsymbol{\mu}_{0\beta_1}^T\boldsymbol{Q}_{0\beta_1}^{-1}\boldsymbol{\mu}_{0\beta_1} + \boldsymbol{\beta}_1^T(\boldsymbol{A}^T\boldsymbol{Q}_y^{-1}\boldsymbol{A}+\boldsymbol{Q}_{0\beta_1}^{-1})\boldsymbol{\beta}_1 - 2\boldsymbol{\beta}_1^T\underbrace{(\boldsymbol{A}^T\boldsymbol{Q}_y^{-1}\boldsymbol{y}+\boldsymbol{Q}_{0\beta_1}^{-1}\boldsymbol{\mu}_{0\beta_1})}_{=(\boldsymbol{A}^T\boldsymbol{Q}_y^{-1}\boldsymbol{A}+\boldsymbol{Q}_{0\beta_1}^{-1})\hat{\boldsymbol{\beta}}_1 = \boldsymbol{Q}_{\beta_1}^{-1}\hat{\boldsymbol{\beta}}_1} \\
&= \boldsymbol{y}^T\boldsymbol{Q}_y^{-1}\boldsymbol{y} + \boldsymbol{\mu}_{0\beta_1}^T\boldsymbol{Q}_{0\beta_1}^{-1}\boldsymbol{\mu}_{0\beta_1} + (\boldsymbol{\beta}_1-\hat{\boldsymbol{\beta}}_1)^T\boldsymbol{Q}_{\beta_1}^{-1}(\boldsymbol{\beta}_1-\hat{\boldsymbol{\beta}}_1) - \hat{\boldsymbol{\beta}}_1^T\boldsymbol{Q}_{\beta_1}^{-1}\hat{\boldsymbol{\beta}}_1.
\end{aligned} \tag{4.8}$$

Again dependencies have not been mentioned explicitly, i.e. $\hat{\boldsymbol{\beta}}_1 = \hat{\boldsymbol{\beta}}_{1|\beta_2,\beta_3}$, and $\boldsymbol{Q}_{\beta_1} = \boldsymbol{Q}_{\beta_1|\beta_2,\beta_3}$. When we omit constant terms, we find

$$\boldsymbol{\beta}_1|\boldsymbol{\beta}_2,\beta_3,\boldsymbol{y} \sim N(\hat{\boldsymbol{\beta}}_1,\boldsymbol{Q}_{\beta_1}). \tag{4.9}$$

Any part of Eq. (4.8) that is not a part of the posterior for the scaling coefficients is now assigned to the marginal likelihood:

$$p(\boldsymbol{y}|\boldsymbol{\beta}_2,\beta_3) \propto \exp(-\frac{1}{2}E) \tag{4.10}$$

$$E = (\boldsymbol{y}-\mathbf{A}\hat{\boldsymbol{\beta}}_1)'\boldsymbol{Q}_y^{-1}(\boldsymbol{y}-\mathbf{A}\hat{\boldsymbol{\beta}}_1) + (\boldsymbol{\mu}_{0\beta_1}-\hat{\boldsymbol{\beta}}_1)'\boldsymbol{Q}_{0\beta_1}^{-1}(\boldsymbol{\mu}_{0\beta_1}-\hat{\boldsymbol{\beta}}_1). \tag{4.11}$$

For the prefactor of the marginal likelihood, the prefactor of the two original densities,

$$\frac{1}{(2\pi)^{n/2}(\det \boldsymbol{Q}_y)^{1/2}} \cdot \frac{1}{(2\pi)^{K/2}(\det \boldsymbol{Q}_{0\beta_1})^{1/2}}, \tag{4.12}$$

is divided by the factor of the posterior for the scaling coefficients, resulting in

$$\frac{(\det \boldsymbol{Q}_{\beta_1})^{1/2}}{(2\pi)^{n/2}(\det \boldsymbol{Q}_y)^{1/2}(\det \boldsymbol{Q}_{0\beta_1})^{1/2}}. \tag{4.13}$$

Instead of simulating the chain for the whole set of the parameters, it is sufficient to simulate over the marginal posterior for the point grid. In each step, only the point grid is simulated, and the associated scaling coefficients are calculated in a least squares adjustment as usual. By integrating out the parameters, one obviously reduces the sampling dimension. One saves sampling one third of the parameters on the price of having to set up and solve a normal equation system in every sampling step. It is difficult to assess which way is faster, as it additionally depends on the implementation and the platform, where the computations are performed. An obvious benefit is that one can integrate the usually available software package for regional gravity field analysis in a black-box manner; the Markov chain algorithm can be just built around the standard code. Finally, integrating the analytical solution in the proposal process helps the chain to move because it leads to good acceptance probabilities—details will follow in Sec. 4.4.1.

4.3.2 The prior on the number of basis functions

The prior on the number might be chosen to be uniform or to a priori put more weight onto fewer basis function, as realized by the Poisson or geometric distribution. When choosing the uniform distribution, the parameters a and b, which represent the minimum and maximum number of basis functions, have to be defined. It would be sensible to choose $a = 1$ and b according to the number of basis functions of the standard grid. Also for the other distributions, one has to define a feasible range. The resulting density functions do not have to be normalized, as a missing factor would cancel in the ratio. In addition, one has to define the free parameters, i.e. the slope of the curve in form of the parameter p for the geometric distribution, where the larger p the larger the descent, and the shape of the Poisson distribution, which can be adapted by the parameter λ. Both parameters are tuning parameters and can be set by experience.

4.3.3 The prior on the point grid

As little prior information as possible shall be put onto the point grid. It is assumed that an individual point is uniformly distributed over the sphere. In the context of regional applications dealing with regional data, only basis functions are important that are close to the observational data. It thus makes sense to limit the feasible area. This is implemented by using the limited uniform distribution introduced in Sec. 3.5.6. It is further assumed that the points are independently distributed. The joint distribution for all points to be used as prior for the point grid follows from multiplying the individual densities:

$$p(\boldsymbol{\beta}_2|\beta_3) = \prod_{k=1}^{K} p(\boldsymbol{x}_k) = \left(\frac{1}{A}\right)^K. \tag{4.14}$$

A is the normalization constant, which corresponds to the feasible area.

4.3.4 The prior on the scaling coefficients

The marginal likelihood has been received in the Sec. 4.3.1. Here, we want to talk about how the individual quantities should be set. $\boldsymbol{Q}_y$ is the covariance matrix of the observations. The stochastic

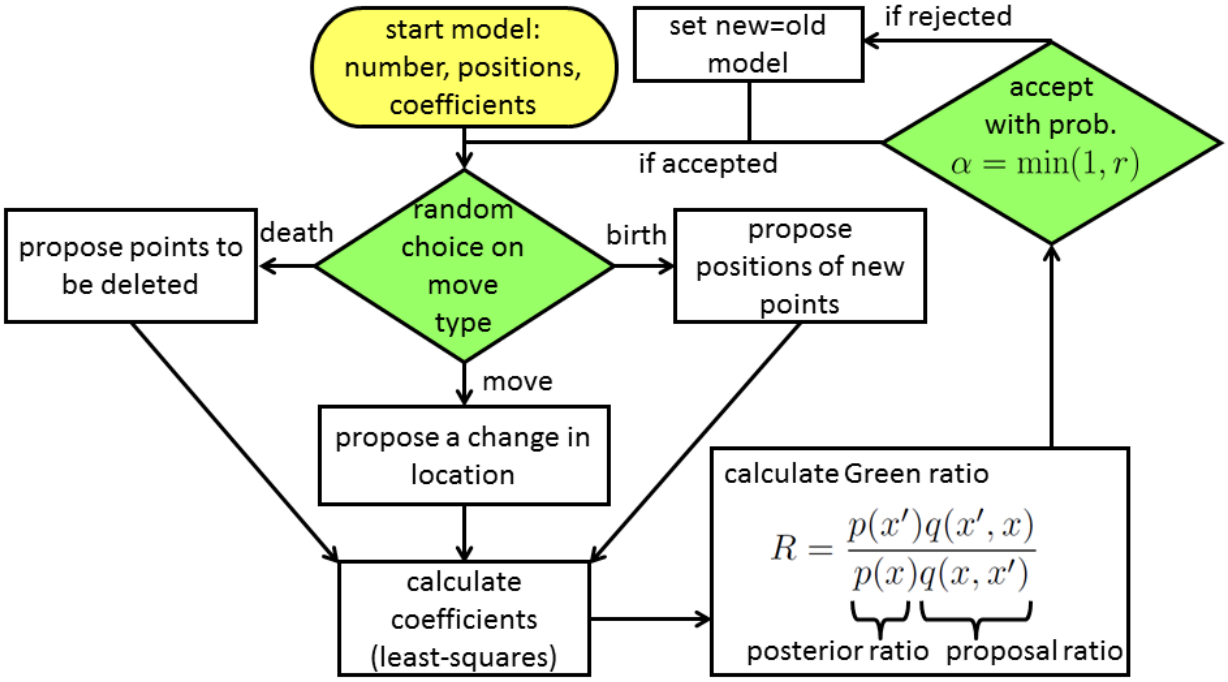


Figure 4.2: Algorithm overview

modeling for the GOCE gradients is done by estimating an empirical covariance function. It is set constant for the whole time of the chain. As was already discussed in Sec. 3.3, in the context of Bayesian statistics, the regularization matrix $\boldsymbol{R}$ can be associated with the inverse of the covariance matrix of the prior information, and $\boldsymbol{\mu}_{0\beta_1}$ is the zero vector. Applying an equivalent to Kaula regularization to the parameterization in RBFs leads in our approach to the regularization matrix being the unit matrix (cf. Sec. 2.5.3). Thus we get $\boldsymbol{Q}_{0\beta_1} = \sigma^2_{0\beta_1}\boldsymbol{I}$, where the variance component $\sigma_{0\beta_1}$ is usually determined by some kind of optimization procedure to get an optimal trade-off between a good data fit and a smooth solution. The optimization of the point grid applied in this thesis will lead to a more stable solution, which is additionally less sensitive to changes of the regularization parameter. This might lead to the idea that it is sufficient to set $\sigma_{0\beta_1}$ constant. How the Markov chain is affected by this assumption and if there are better alternatives will be subject of Sec. 4.9.

4.4 Implementation of reversible jumps for the optimization of point grids

4.4.1 Designing a well-mixing chain

The RJMCMC sampling algorithm applied in this thesis yields samples from the target distribution, i.e. the empirical density of the generated samples approximates the target distribution. Depending on the implementation, the number of samples required to visit the whole domain of the target distribution will differ. For our problem, we are limited in the number of samples, as in every step an adjustment problem has to be solved, which is time consuming. So we want our chain to visit all the interesting places in as few steps as possible; in other words, we want many accepted samples while proposing large steps.

To let the chain move, we need the acceptance probability to be high. This would be the case if the proposal had a similar high posterior support as the actual sample, i.e. if the sample had

a similar high posterior density value. For a move step, where a proposal is achieved by random variation of the actual parameter set, this can be easily achieved by proposing the new parameter values sufficiently close to the actual values. In contrast, when the proposal step is too large, we get potentially far, but the proposal will never be accepted, as it has a poor posterior support. On the other hand, one might not want to perform too small steps either, as although having a huge number of acceptances, the chain will move rather slowly, and the really interesting places might never be reached. In a nutshell, good proposals perform steps that are as large as possible while having a high acceptance rate. This would end up in what Andrieu et al. (2003) call a well-mixing chain. As a rule of thumb for a one dimensional example with samples being generated from a normal distribution, the spread of the proposal should be chosen in a way that 45% of the proposals are accepted. When the number of dimensions approaches infinity, it should be approximately 23% (Chib and Greenberg, 1995). Instead of randomly changing all parameters at the same time, one can also change a single parameter slightly stronger, which leads to the one-variable-at-a-time Metropolis-Hastings algorithm. In presence of correlations between the parameters, it would be sensible to aggregate the variables and change them together. As for our problem, at least one point on the unit sphere is changed, i.e. two coordinates together. Hastings (1970) adds that when only few parameters are changed, one should apply some kind of recurrence formula to take advantage of the calculations already performed in the step before. As for our problem, one could set up the normal equations in every step completely but only compute the entries that were changed by the change of the point positions. If the calculations are very time consuming, Hastings (1970) also advises to better vary all coordinates to a small amount in every step.

When doing a jump that changes dimension, the concept of close proposals can not be readily applied. The reason is that when adding a new parameter, there is nothing to which it should be close to. To nevertheless retain a high acceptance probability, we try to propose samples that have a high posterior density by themselves, not only compared to the previous sample. Therefore, I propose grid points at places which I think are probable, which is where the signal changes rapidly. This idea has been implemented in two different move types. Finally, using the analytically determined scaling coefficients facilitates considerably the jumps between dimensions, as they provide for the particular point the highest density value. In this way, a complicated setting of the coefficients and careful balancing with the other basis functions as is necessary in other approaches can be neglected (DiMatteo et al., 2001).

4.4.2 Move types and probabilities

In Sec. 3.4.5, $p(K, K')$ was introduced, which is the probability to propose a move to the model $M_{K'}$ when currently being at the model M_K. The representation of the gravity potential by RBFs is a nested model, which means that any subset of model components represents itself a model, and in every model the components have the same meaning. To create a new model, one just has to add basis functions to the current model or to delete them, leaving the rest of the model unchanged. Thus randomly choosing the model boils down to choosing one of, in the simplest case, three different move types: birth of a basis function, death of a basis function and update of the current model, which consists of a change in the positions of the basis functions. The individual move types are attempted with the probabilities q_b, q_d and q_m, respectively. If one is still far from the optimum, and even a large change in the model would not change the likelihood function strongly, so that one nevertheless retains a high acceptance probability, it might be sensible to also allow for a birth or death of κ basis functions. This would be also useful when the target density for K is multimodal because changing several functions at the same time could facilitate the jumps between the models. The corresponding move probabilities are denoted as $q_{b\kappa}$ and $q_{d\kappa}$ with κ being the number of functions to be updated. The move probabilities must satisfy $\sum q = 1$, and further the conditions $q_{d\kappa}(\kappa) = 0$ and $q_{b\kappa}(K_{\max} - \kappa + 1) = 0$ have to be fulfilled, which is in accordance with

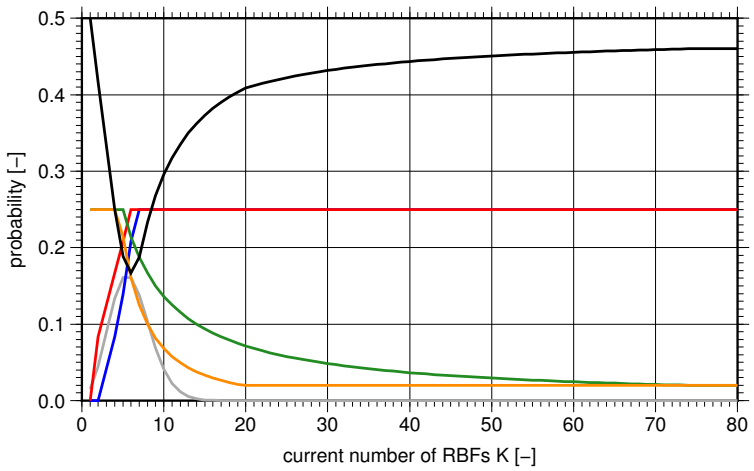


Figure 4.3: Probability for different move types derived from the Poisson distribution ($\lambda = 6$) presented as a function of K. The constants have been chosen as $c_1 = 0.25$, $c_2 = 0.02$ to realize at any state a high probability for a move and to guarantee a minimum probability for every move type, respectively. In particular, the minimum probability has been set to be by a factor of ten smaller than if the probability would have been distributed uniformly over the five move types. Blue: $q_{d2}(K)$, red: $q_{d1}(K)$, black: $q_m(K)$, green: $q_{b1}(K)$, orange: $q_{b2}(K)$, gray: pdf of the Poisson distribution for comparison.

the prior on K. The probabilities can either be prescribed, e.g. one can specify q_m and distribute the probability $1 - q_m$ on the other move types, which is e.g. followed by Lindstrom (2002), or we can use a method proposed by Green (1995), making use of the prior on K. In this approach, death steps are proposed particularly often when according to the prior the actual number of basis functions is too high, and birth steps are proposed often when the prior predicts more functions. This is implemented by choosing

$$q_{b\kappa}(K) = c_1 \min\{1, p(K+\kappa)/p(K)\}, \quad q_{d\kappa}(K) = c_1 \min\{1, p(K-\kappa)/p(K)\} \tag{4.15}$$

with c_1 being set so that the sum of the probabilities is less than one. Then the probability for a move step is set to $q_m = 1 - \sum_\kappa (q_{b\kappa} + q_{d\kappa})$. I have slightly changed this approach to assure that a minimal probability c_2 for every move type remains by adding a further condition:

$$q_{b\kappa} = \max\{c_2, q_{b\kappa}\}, \quad q_{d\kappa} = \max\{c_2, q_{d\kappa}\}. \tag{4.16}$$

An example for the resulting move probabilities in dependence of the actual model is presented in Fig. 4.3. When more than one move type is available for the transition to a particular model, the probability is distributed uniformly. When e.g. a local and a global birth step are realized within the same Markov chain, which might be sensible as it improves mixing, the individual steps would be attempted with $\frac{1}{2}q_b$. The choice between the available move types is made at random using inverse transform sampling. Assuming that the three basic move types are available, a uniform random number u is generated, and the death is adopted if $u < p_d$, else a move is adopted if $u < p_d + p_m$, and a birth is adopted otherwise.

4.4.3 Move

This move type is an ordinary move in the sense of the random walk Metropolis algorithm, i.e. the new sample is generated in dependence of the previous sample with the help of a symmetric proposal distribution. That way, the chain can move freely across the feasible area, which is important to actually find the optimum. For the proposal, the isotropic Fisher distribution is used being the

equivalent to the normal distribution on the sphere. As part of the move, one firstly proposes a point to be moved, sets up the Fisher distribution and generates a new point, which then replaces the original point. For this step, the proposal reads

$$q(\boldsymbol{\beta}_2, \boldsymbol{\beta}_2'^*) = q(\boldsymbol{\beta}_2, \boldsymbol{x}_k') = \frac{1}{K}\, p_F(\boldsymbol{x}_k, \boldsymbol{x}_k'). \tag{4.17}$$

As the nodal point of the Fisher distribution varies for every proposal, the notation was slightly changed by taking up the nodal point coordinates into the argument of the Fisher distribution. For the backward step, one finds

$$q(\boldsymbol{\beta}_2', \boldsymbol{\beta}_2^*) = q(\boldsymbol{\beta}_2', \boldsymbol{x}_k) = \frac{1}{K}\, p_F(\boldsymbol{x}_k', \boldsymbol{x}_k). \tag{4.18}$$

Because of the symmetry of the Fisher distribution, both terms are equal and cancel each other out in the ratio.

More generally, to move κ grid points dependent on the current grid, the proposal density is

$$\begin{aligned} q(\boldsymbol{\beta}_2, \boldsymbol{\beta}_2'^*) &= q(\boldsymbol{\beta}_2; \boldsymbol{x}_k', \boldsymbol{x}_l') \\ &= \frac{1}{K}\, p_F(\boldsymbol{x}_k, \boldsymbol{x}_k') \frac{1}{K-1}\, p_F(\boldsymbol{x}_l, \boldsymbol{x}_l') + \frac{1}{K}\, p_F(\boldsymbol{x}_l, \boldsymbol{x}_l') \frac{1}{K-1}\, p_F(\boldsymbol{x}_k, \boldsymbol{x}_k') \\ &= \frac{2}{K(K-1)}\, p_F(\boldsymbol{x}_k, \boldsymbol{x}_k') p_F(\boldsymbol{x}_l, \boldsymbol{x}_l') \end{aligned} \tag{4.19}$$

for the example of $\kappa = 2$. The factor can be identified with a binomial coefficient. Suppose that κ points are to be selected randomly from K available points. One does not wish to draw the same point for a second time, which is said to be 'without putting back' in the jargon of combinatorics. One further is not interested in which point is chosen at the beginning as long as the same configuration is received ('without order'). Then the number of possibilities to choose κ out of K is specified by the binomial coefficient. The reciprocal of this with $\kappa = 2$ is just the factor that we have noticed in Eq. (4.19). In the backwards step, the two new points are chosen, the Fisher distribution is set up, and the original points are generated. As was the case above, because of the symmetry of the Fisher distribution, this is equal to Eq. (4.19) with the tick marks being switched.

For the limiting case of moving one basis function, although being the same for any other choice, the probability of accepting this step boils down to the marginal likelihood ratio:

$$\begin{aligned} R &= \frac{p(\boldsymbol{y}|\boldsymbol{\beta}_2', \beta_3') p(\boldsymbol{\beta}_2'|\beta_3') p(\beta_3') q(\beta_3', \beta_3) q(\boldsymbol{\beta}_2', \boldsymbol{\beta}_2^*)}{p(\boldsymbol{y}|\boldsymbol{\beta}_2, \beta_3) p(\boldsymbol{\beta}_2|\beta_3) p(\beta_3) q(\beta_3, \beta_3') q(\boldsymbol{\beta}_2, \boldsymbol{\beta}_2'^*)} \\ &= \frac{\frac{(\det \boldsymbol{Q}'_{\beta_1})^{1/2}}{(2\pi)^{n/2} (\det \boldsymbol{Q}_y)^{1/2} (\sigma'_{0\beta_1})^K} \exp\left(-\tfrac{1}{2}E'\right) \left(\tfrac{1}{A}\right)^K p(K) q_m(K) \tfrac{1}{K}\, p_F(\boldsymbol{x}_k', \boldsymbol{x}_k)}{\frac{(\det \boldsymbol{Q}_{\beta_1})^{1/2}}{(2\pi)^{n/2} (\det \boldsymbol{Q}_y)^{1/2} (\sigma_{0\beta_1})^K} \exp\left(-\tfrac{1}{2}E\right) \left(\tfrac{1}{A}\right)^K p(K) q_m(K) \tfrac{1}{K}\, p_F(\boldsymbol{x}_k, \boldsymbol{x}_k')} \\ &= \frac{(\det \boldsymbol{Q}'_{\beta_1})^{1/2} (\sigma_{0\beta_1})^K \exp\left(-\tfrac{1}{2}E'\right)}{(\det \boldsymbol{Q}_{\beta_1})^{1/2} (\sigma'_{0\beta_1})^K \exp\left(-\tfrac{1}{2}E\right)}. \end{aligned} \tag{4.20}$$

Note that we put dashes on quantities that, although not being sampled itselves, depend on sampled quantities.

4.4.4 Global birth

In order to jump to a model with a higher number of parameters, a birth move is required. Here, we propose a global birth, meaning that new points are inserted independent of the current grid

over the whole feasible area. The number of basis functions to be inserted is a tuning parameter of the approach and should be adapted so that a reasonable number of proposals is accepted.

One intuitively assumes that the point grid for the arrangement of the basis functions should be oriented at the structures of the gravity field. This idea was already applied in earlier approaches. An early example is Balmino (1972), who set point masses to extremal points that were selected from a map of gravity anomalies. Another example is Antoni (2012), who chose places with large disturbing potential values as starting positions for his optimization algorithm. A number of other optimization or greedy approaches, which were summarized in the introductory part to this thesis (e.g. Barthelmes, 1986; Marchenko and Abrikosov, 1995; Wittwer, 2009), realize this principle indirectly by setting points at places where the largest residuals occur, so depending on the functional at places of large gravity anomalies or gravity disturbances. To exploit prior knowledge on the structure of the gravity field also in the present approach, I integrate the gravity field signal into the proposal process hoping that this will yield good proposals and a high acceptance rate. For this purpose, a special probability density function is derived from a gravity field model (see Sec. 3.6.3). Theoretically, any gravity field functional could be used to define a density function. For example, when choosing the geoid, this would mean that comparatively many points are proposed where the geoid height is large. Proposal distributions based on different gravity field functionals are tested later on in the results chapter, Sec. 5.3.3.

To perform the global birth step, the desired number of points is sampled from the density function here denoted by p_G. Working with an unsorted vector, new points are simply added at the end in the order of their occurrence. Individual samples are taken as independent, so that the full proposal density is achieved by multiplication:

$$q(\boldsymbol{\beta}_2'^{*}) = q(\boldsymbol{x}'_{K+1}, \ldots, \boldsymbol{x}'_{K+\kappa}) = \prod_{k=1}^{\kappa} p_G(\boldsymbol{x}'_{K+k}). \tag{4.21}$$

To define the acceptance probability, we need the corresponding death step, which still has to be defined. The death step is performed by deleting the required number of basis functions. In consideration of how the birth step was defined, namely by just putting the new points at the end of the vector, this step is purely deterministic. It requires neither to sample random numbers nor to evaluate a proposal density. This detail has often been a source of errors in previous studies (Jannink and Fernando, 2004). Roodaki et al. (2012) uncovered that Andrieu and Doucet (1999), using the same birth step as we do, introduced the term $1/K+1$ as proposal probability for the death of one model component. The additional term would affect the acceptance probability in the same way as if an additional prior on the number would have been introduced with a preference for smaller models. The correct expression for the acceptance probability is

$$\begin{aligned} R &= \frac{\frac{(\det \boldsymbol{Q}'_{\beta_1})^{1/2}}{(2\pi)^{n/2}(\det \boldsymbol{Q}_y)^{1/2}(\sigma'_{0\beta_1})^{K+\kappa}} \exp\left(-\tfrac{1}{2}E'\right)\left(\tfrac{1}{A}\right)^{K+\kappa} p(K+\kappa) q_{d\kappa}(K+\kappa)}{\frac{(\det \boldsymbol{Q}_{\beta_1})^{1/2}}{(2\pi)^{n/2}(\det \boldsymbol{Q}_y)^{1/2}(\sigma_{0\beta_1})^{K}} \exp\left(-\tfrac{1}{2}E\right)\left(\tfrac{1}{A}\right)^{K} p(K) q_{b\kappa}(K) \prod_{k=1}^{\kappa} p_G(\boldsymbol{x}'_{K+k})} \\ &= \frac{(\det \boldsymbol{Q}'_{\beta_1})^{1/2}(\sigma_{0\beta_1})^{K} \exp\left(-\tfrac{1}{2}E'\right)\left(\tfrac{1}{A}\right)^{\kappa} p(K+\kappa) q_{d\kappa}(K+\kappa)}{(\det \boldsymbol{Q}_{\beta_1})^{1/2}(\sigma'_{0\beta_1})^{K+\kappa} \exp\left(-\tfrac{1}{2}E\right) p(K) q_{b\kappa}(K) \prod_{k=1}^{\kappa} p_G(\boldsymbol{x}'_{K+k})}. \end{aligned} \tag{4.22}$$

As explained earlier, one has to realize equal dimensions in the forward and backward step to compare samples of different dimension by means of their density values. This is one of the key points of the Metropolis-Hastings-Green algorithm. In the above ratio, the proposed state consists of K plus κ points. The actual state has K points. κ additional points are generated using the proposal. The dimension matching criterion is therefore fulfilled, and the term $1/\kappa$, which is the prior ratio, is balanced by the proposal. This is still more evident when the proposal is chosen as uniform distribution, as in this case the terms would cancel each other out.

4.4.5 Local birth

As an alternative to the global birth, one could also use a local birth, which introduces many basis functions at places where already many basis functions lie. Assuming that many points have already been accepted in areas of rough gravity field signal, the local birth step proceeds according to the same principle as the global birth. A point is chosen uniformly from the available points; thus, one naturally picks often those in rough areas. The Fisher distribution is set up, and a new point is drawn in the neighborhood.

The proposal density to generate κ new points this way reads

$$q(\boldsymbol{\beta}_2, \boldsymbol{\beta}_2'^*) = q(\boldsymbol{\beta}_2; \boldsymbol{x}'_{K+1}, \ldots, \boldsymbol{x}'_{K+\kappa}) = \frac{1}{K^\kappa} \sum_{k=1}^{K} p_F(\boldsymbol{x}_k, \boldsymbol{x}'_{K+1}) \cdot \ldots \cdot \sum_{k=1}^{K} p_F(\boldsymbol{x}_k, \boldsymbol{x}'_{K+\kappa}). \tag{4.23}$$

Here it is also allowed to repeatedly propose the same point as origin. Note that every new point could have been generated from any available point, which is taken into account by summing up over the individual density values. K^κ is the number of possibilities to choose κ out of K if one is interested in the order of arrival and repetitions are allowed.

With this proposal, the odds ratio follows to

$$R = \frac{(\det \boldsymbol{Q}'_{\beta_1})^{1/2} (\sigma_{0\beta_1})^K \exp\left(-\frac{1}{2}E'\right)\left(\frac{1}{A}\right)^\kappa p(K+\kappa) q_{d\kappa}(K+\kappa)}{(\det \boldsymbol{Q}_{\beta_1})^{1/2} (\sigma'_{0\beta_1})^{K+\kappa} \exp\left(-\frac{1}{2}E\right) p(K) q_{b\kappa}(K) \cdot \frac{1}{K^\kappa} \sum_{k=1}^{K} p_F(\boldsymbol{x}_k, \boldsymbol{x}'_{K+1}) \cdot \ldots \cdot \sum_{k=1}^{K} p_F(\boldsymbol{x}_k, \boldsymbol{x}'_{K+\kappa})}. \tag{4.24}$$

When the local birth is not used as an alternative to the global birth but in addition to it, which might improve the mixing, the death move has no unique revers move anymore, which is necessary to formulate the odds ratio for each move type individually (Roodaki et al., 2012). We have to consider the move types together, as we could get the same constellation from both. Consequently, we have to sum up the expressions for both moves while dividing up the move probability according to

$$R = \frac{(\det \boldsymbol{Q}'_{\beta_1})^{1/2} (\sigma_{0\beta_1})^K \exp\left(-\frac{1}{2}E'\right)\left(\frac{1}{A}\right)^\kappa p(K+\kappa) q_{d\kappa}(K+\kappa)}{(\det \boldsymbol{Q}_{\beta_1})^{1/2} (\sigma'_{0\beta_1})^{K+\kappa} \exp\left(-\frac{1}{2}E\right) p(K) \left(\frac{1}{2} q_{b\kappa}(K) \prod_{k=1}^{\kappa} p_G(\boldsymbol{x}'_{K+k}) + \frac{1}{2} q_{b\kappa}(K) \frac{1}{K^\kappa} \sum_{k=1}^{K} p_F(\boldsymbol{x}_k, \boldsymbol{x}'_{K+1}) \cdot \ldots \cdot \sum_{k=1}^{K} p_F(\boldsymbol{x}_k, \boldsymbol{x}'_{K+\kappa})\right)}. \tag{4.25}$$

4.4.6 Death

Within the death step, the model dimension is reduced. It thus represents the backward step for the birth steps mentioned before. When performing a death step, 1 to κ basis functions are randomly deleted, where we use a small κ when changing the number of basis functions has a large effect on the value of the likelihood function and a large κ in the opposite case.

The probability of accepting a death step is just the reciprocal value of the acceptance probability for the corresponding birth step; one just has to slightly adapt the notation from $K+\kappa$ to K. This is clear from the definition of the acceptance probability. For the death step, with the backward step being defined uniquely to be the global birth, this would be

$$R = \frac{(\det \boldsymbol{Q}'_{\beta_1})^{1/2} (\sigma_{0\beta_1})^K \exp\left(-\frac{1}{2}E'\right) p(K-\kappa) q_{b\kappa}(K-\kappa) \prod_{k=1}^{\kappa} p_G(\boldsymbol{x}_{K-\kappa+k})}{(\det \boldsymbol{Q}_{\beta_1})^{1/2} (\sigma'_{0\beta_1})^{K-\kappa} \exp\left(-\frac{1}{2}E\right)\left(\frac{1}{A}\right)^\kappa p(K) q_{d\kappa}(K)} \tag{4.26}$$

and correspondingly for the local birth.

The birth and death as defined before always add or delete the last point, which is not what we really want. To overcome this shortcome, we can just introduce another update performing a uniform choice between the different permutations of the point vector (Geyer, 2005). This step is always accepted, as the different permutations represent the same model and thus have the same probability. We must not even really perform the permutation; the only effect this update has is that a random point rather than the last point can be deleted within the death step. Another still easier explanation would be that we now choose from among the K points randomly which one to delete. This however goes along with not only choosing a new point during the birth step but also a random position, which again adds a factor $1/K$. Both terms cancel, which leads to the odds ratio being exactly the one given above (Roodaki et al., 2012).

4.5 The 8-point example revisited: demonstration of the validity of the procedure

A nice way to check whether the different steps were properly designed and implemented is to set the likelihood equal to one and run a Markov chain (see e.g. Sambridge et al., 2006). Then the algorithm should produce samples of the prior distribution, which can easily be verified. In the course of this thesis, a uniform prior is used for the number of basis functions, here limited to the range 1–10. Conditional to the number, the positions of the grid points are regarded as independent uniformly distributed over the model domain. All possibilities for determining the move types and move probabilities were tested. The results are presented exemplarily for one of the constellations. Fig. 4.4 shows the marginal posterior for the number of basis functions. Fig. 4.5 shows the distribution of the point positions for the grids with five points displayed in a common histogram. Both distributions reflect the chosen prior distribution, thereby demonstrating the validity of the procedure.

I want to say a few words about the idea of calculating a closed-loop simulation with error-free data. This would involve setting up a model, simulating the data from this model and then retrieving the model parameters in an inverse calculation, which should work up to numerical errors. It is difficult to perform such a calculation in the present case. This is for two reasons. First, the residual sum of squares, which is part of the argument of the marginal likelihood, is currently derived from the normal equations according to $\boldsymbol{v}^T\boldsymbol{v} = \boldsymbol{y}^T\boldsymbol{y} - \boldsymbol{\beta}^T\boldsymbol{n}$. The residual sum of squares, $\boldsymbol{v}^T\boldsymbol{v}$, is naturally much smaller than the square sum of the observations, $\boldsymbol{y}^T\boldsymbol{y}$. So here a small value is determined as difference of two large values. This only works in the presence of errors when the remaining residuals are still sufficiently large. For error-free data, this difference has no meaning, as numerically it does not differ from zero. The best we can hope to achieve is to find a model that cannot be distinguished from the true model, and in our case this approximation would be rather crude. Second, I find it difficult to specify a meaningful likelihood function for a problem with error-free data. If we used the same likelihood function as for the data containing errors, we would not be able to retrieve the true model within a limited amount of time, for there would be many possible models in the range of the specified accuracy. Alternatively, one might consider choosing a standard deviation close to zero. However, the resulting likelihood function is very peaky, causing the algorithm to get stuck in a local maximum, since a step increasing the residual sum of squares will practically never be accepted. To conclude, a closed-loop scenario with error-free data does not make sense. Anyway, such a calculation would only show if the procedure was properly conceived and implemented, which has already been shown by the previous calculation with the likelihood function having been set to one, but it does not say anything about the performance in a realistic framework in the presence of measurement noise.

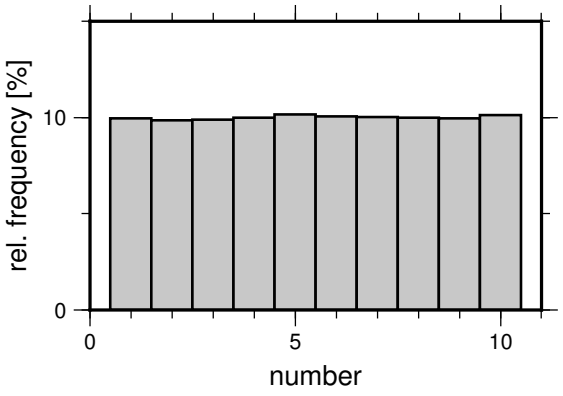


Figure 4.4: Marginal density of the number for the Markov chain with the likelihood being set to 1

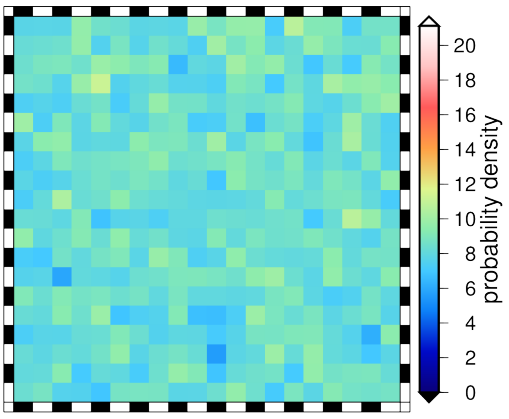


Figure 4.5: Distribution of the point positions for the grids with 5 points for the Markov chain with the likelihood being set to 1

4.6 The 8-point example revisited: convergence issues

After having been proven to work properly in the last section, the algorithm shall now be applied to simulate the probability distribution of the point grid for the example with the 8 hidden basis functions. To not express a preference for any configuration of basis functions, the prior for the number was chosen to be uniformly distributed in the range 1–650, and the prior for the point positions was defined by means of a spherical uniform distribution in the area of the observations. The variance factor that specifies the prior for the scaling coefficients was defined as 0.062—the variance factor of the true grid—and retained for the whole run of the random algorithm. The standard grid with 632 points was used as initial grid, and as before the same kernel function was used as for the simulation of the data. Coming from the start model, the algorithm chooses between a birth, death or move of one or more basis functions (see Fig. 4.2). At the beginning, the chain is still far away from the mode of the distribution in an area of flat density. If the proposal is not far enough away from the actual state of the chain, nearly every proposal would also be accepted. The chain would wander around, and even if in the end the number of accepted death steps is higher, it would take the chain a long time to reduce the number of basis functions in this way. For test purposes, I simulated such a Markov chain that allows for a birth or death of one basis function only. The acceptance rate lay at respectively 59% or 95%, and 4705 steps were required to reduce the number of basis functions to 52. For comparison, a simultaneous birth/death of 5 basis functions led to lower acceptance rates of respectively 15% or 82%, and the number of basis functions reduced to 52 already after 475 steps. Moreover, since death steps are obviously accepted particularly frequently in the initial stage of the chain, it makes sense to also propose them more

frequently. Therefore, not a uniform distribution was used as proposal for the number as was done for the two test runs, but the move probabilities were derived from a Poisson distribution with $\lambda = 8$ and $c_1 = 0.5555$, $c_2 = 0.1111$ in the way described in Sec. 4.4.2. In this way, the convergence of the chain was accelerated again, so that 52 basis functions were reached already after 295 steps. For the acceptance rates of different parts of the simulated chain, see Tab. 4.2. When the algorithm decided for a birth step, the new points were sampled globally with the help of a uniform distribution in the feasible area. In a death step, 5 points were chosen randomly from the available points and deleted. In a move step, 25 functions were moved, which in this stage of the chain is a sensible number in the sense of a fast convergence. After approximately 300 steps, the acceptance of the steps decreased (cf. Tab. 4.2). The number of basis functions was in this time reduced from 632 in the beginning to 20 basis functions. The chain finally arrived in a steeper part of the target density, where already small changes in the parameters change the target density value strongly. The proposal process was therefore adapted so that during a birth, death or move step only one single basis function was added, deleted or moved. For the move step, this resulted in an acceptance rate of 23%, which in the sense of mixing is considered to be ideal. Moreover, the individual move types were also selected with equal probabilities from there on.

In the following, the output of the chain shall be analyzed with regard to the convergence and mixing behavior of the chain. In practice, "it is not possible to say with certainty that a finite sample from an MCMC algorithm is representative of an underlying stationary distribution" (Cowles and Carlin, 1996). Yet, there are numerous methods that at least provide an indication. Graphical methods like time series plots or histograms have been used a lot; for an assessment of the dependencies within the chain, also autocorrelation plots are valuable. Frequently the Gelman-Rubin tool for convergence diagnostics is used (Gelman and Rubin, 1992). Several chains are simulated, and it is evaluated if the variance across the chains is equal to the variance within the chains, which would point to the convergence of the chain. An overview of tools to assess convergence and mixing is given by Mengersen et al. (1999) or Sinharay (2003). Both publications are restricted to tools for fixed dimensional MCMC. In RJMCMC convergence monitoring is still more difficult because the parameters might have a different meaning in different models or simply take on different values. Some people gave consideration to that and designed tools specifically for transdimensional problems, among them Brooks and Giudici (2000), who proposed a generalized version of the approach of Gelman and Rubin. They further proposed to monitor functions of the parameters that allow for a meaningful comparison across models. In a similar approach, Castelloe and Zimmerman (2002) made allowance for a multivariate convergence. In contrast, Brooks et al. (2003) applied test statistics mainly concentrating on marginal convergence of the model indicator. Finally, Sisson and Fan (2007) developed an approach that is specifically designed for models that are similar to the random point process, where the monitored function is e.g. the distance of a random point to the closest point.

In the present work, only the simple tools were applied. Timeseries plots enable to assess the convergence graphically. If the samples follow the target density, the moments of the distribution should not change over time. In this sense, a trend or jump in the timeseries plot or a long-term-change in the spread would indicate that the distribution of the samples does still converge in the direction of the target distribution. Fig. 4.6 shows timeseries plots for the number of the basis functions for different parts of the chain. In the initial phase of the chain, one can see the movement from 632 basis functions in the beginning to 8 basis functions after almost 1600 steps. Apart from the first several thousand samples, the chain looks stationary; there is no apparent trend. Looking at the zoom of the samples 300,000–400,000, one sees that the chain got stuck in certain areas over a long period of time, when the chain was e.g. in the area of the larger models, the number 8 was sometimes not visited for up to several thousand steps. The chain also visited the number 7 only in large time intervals to stay there for a couple of time afterwards. This leads to correlations among the samples. Hence, one can already learn a lot about the dependencies within the chain by just

looking at the wavelength of the dominant cycles in a time series plot. Accepting a birth/death step is in general more difficult and therefore less likely. For the 8-point example, the acceptance rate was about 6%; others reported similar low values, e.g. 7–21% were mentioned for a birth/death of one function by Lindstrom (2002) or 4–18% by Richardson and Green (1997). Against this background, I would say that the chain demonstrates a reasonably good mixing behavior also in comparison to examples from the literature (e.g. Lindstrom, 2002 or Gallagher et al., 2009).

The number of RBFs is only one of the parameters. As said above, it is not advisable to assess the convergence and mixing of the chain by standard tools, like time series plots, for the rest of the parameters because they are dependent on the model. As an alternative, I consider the rms of the differences to the true field. It is a function of all parameters and should therefore provide a comprehensive picture of the correlations; it is comparable across different models and invariant against the sorting of the parameters.

Fig. 4.7 shows the traceplots for the rms. In the initial part of the chain, one can see the rms converge from the relatively high initial value of 0.8 m in the direction of better solutions, which on average have a rms of 0.16 m. A burn-in period was chosen using the point in time where the chain reaches the mean rms. So one can be sure to have reached at least an area of medium high density. On this basis, I chose a burn-in period of 5,000 steps, and the first 5,000 samples were disregarded in all subsequent calculations. As for the number, no change in the stochastic behavior is visible in the timeseries plot for the full chain. Moreover, statistical values were calculated for any 100,000 samples, which did not give an indication of a still continuing convergence of the chain. With respect to the mixing of the chain, in comparison with the traceplots for the number, the cycles turn out broader, which points to larger correlations.

If the generated samples belong all to the same distribution, namely to the target distribution, the distribution would also look alike for any sufficiently long stretch of the samples. This can be tested easily by dividing the samples up and compare them on the basis of histogram plots. For the number, the two histograms calculated from half of the samples each look the same and have the same statistical values (Fig. 4.9). This also applies for a histogram based on the first 200,000 samples. The chain thus seems to be converged and produces samples from the target distribution. The corresponding histograms for the rms look similar but slightly differ from each other also in the statistical values (Fig. 4.10). This is not surprising because the rms reflects the mixing of all parameters.

The autocorrelation function specifies how strongly the samples depend on each other after a given number of steps. This is relevant because dependent samples do not contain as much information as they would if they were independent. So to achieve the same accuracy in parameter estimation, much more samples might possibly be necessary. Fig. 4.8 shows the autocorrelation function for the number and the rms both for the full chain and for the subset of samples that belongs to the range of typical models with 8–10 basis functions. For the number, correlations are weaker than for the rms, as could have been expected since the rms is affected by the mixing of all model parameters. With regard to the matter of an appropriate duration of the chain, Geyer (2011) states that “you need to run a large multiple of the time it takes the autocovariances to decay to nearly zero”, which in this scenario is fulfilled in any case.

4.7 Derivation of various estimates

4.7.1 Model comparison with Bayes factors and the question of parsimony

Bayes factors form the basis for model selection in Bayesian statistics. They can be derived directly from the output of the RJMCMC algorithm, making them particularly interesting in the course of this thesis.

Table 4.2: Number of proposed (*prop.*) and accepted (*acc.*) death, move and birth steps for different parts of the chain

	up to ∼300			**up to 500**			**from 500**		
	prop.	**acc.**	%	**prop.**	**acc.**	%	**prop.**	**acc.**	%
death	167	138	82.64	264	144	54.55	333697	20973	6.29
move	94	38	40.43	171	38	21.84	332118	81442	24.52
birth	34	22	64.71	61	23	37.71	333685	20955	6.28

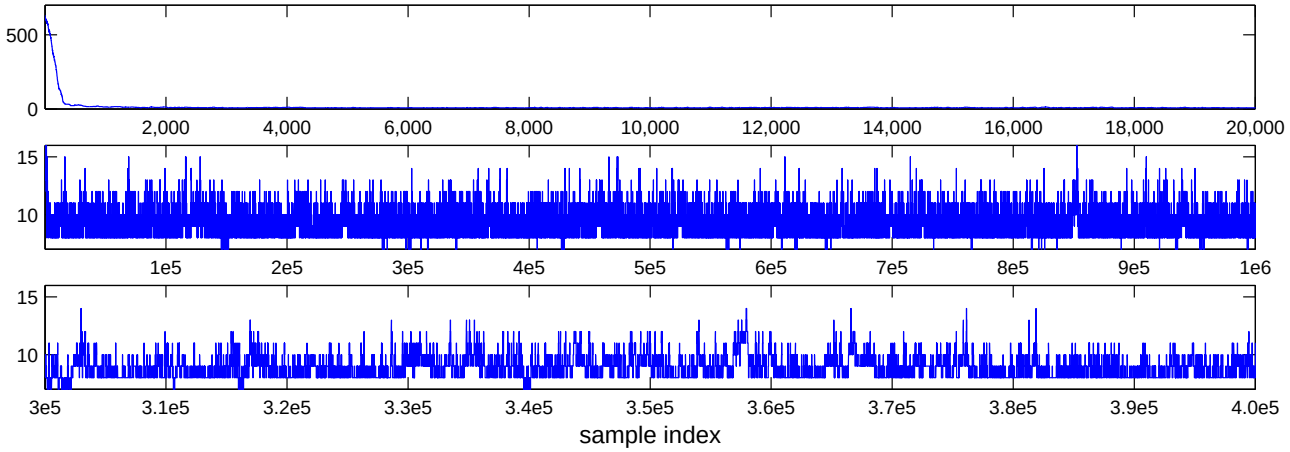


Figure 4.6: Time series plots for the number of RBFs for the beginning of the chain (*top*), the full chain (*center*), and the samples 300,000 to 400,000 (*bottom*)

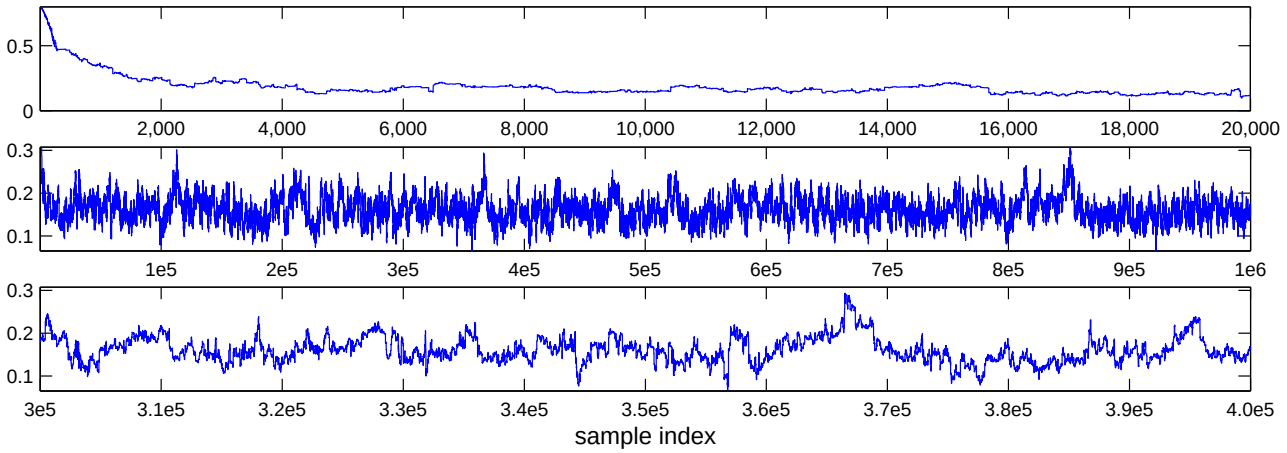


Figure 4.7: Time series plots for the rms between the individual solutions and the true field in terms of geoid heights in meter for the beginning of the chain (*top*), the full chain (*center*), and the samples 300,000 to 400,000 (*bottom*)

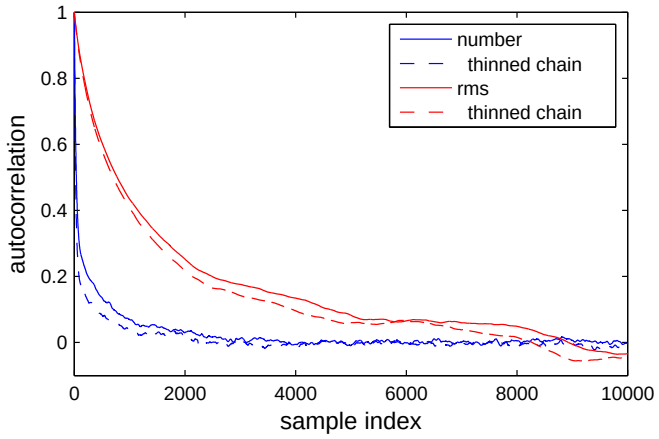


Figure 4.8: Autocorrelation function for the time series of the number of basis functions and the time series of rms values, both for the full chain and for a variant that is limited to the models with 8 to 10 basis functions

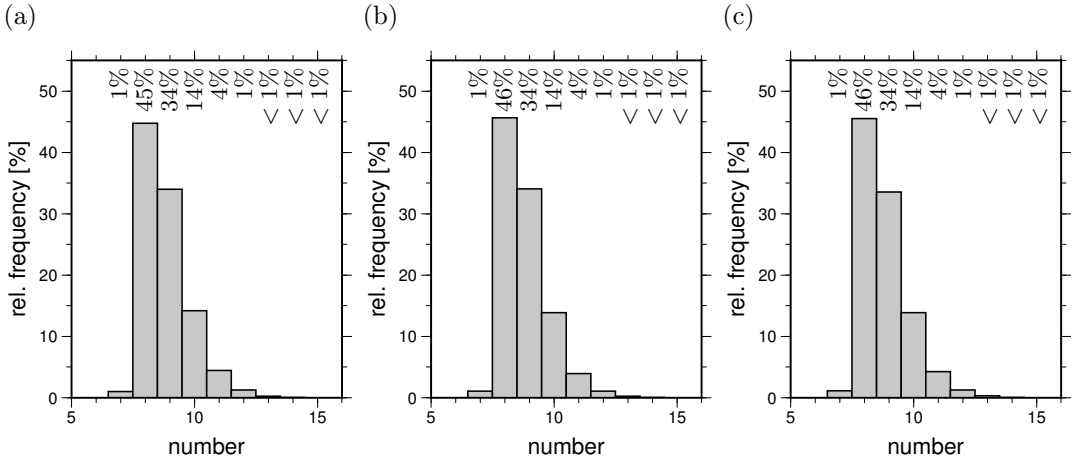


Figure 4.9: Histogram plots for the number of RBFs for (a) the first half of the chain, (b) the second half of the chain, and (c) the first 200,000 samples

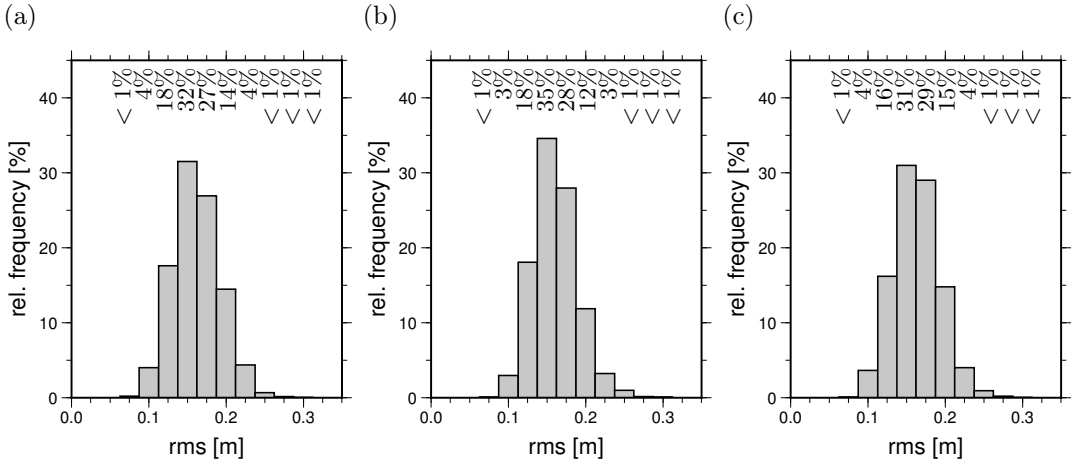


Figure 4.10: Histogram plots for the rms between the individual solutions and the true field in terms of geoid heights for (a) the first half of the chain, (b) the second half of the chain, and (c) the first 200,000 samples

Applying the Bayes theorem to the marginal density of the number, and building the ratio between the models $\beta_3 = K_1$ and $\beta_3 = K_2$, we get

$$\frac{p(K_1|\boldsymbol{y})}{p(K_2|\boldsymbol{y})} = \frac{p(\boldsymbol{y}|K_1)}{p(\boldsymbol{y}|K_2)} \frac{p(K_1)}{p(K_2)} \tag{4.27}$$

with the Bayes factor $B_{12} = \frac{p(\boldsymbol{y}|K_1)}{p(\boldsymbol{y}|K_2)}$. In words, the posterior odds is equal to the Bayes factor times the prior odds. The Bayes factor is the ratio of the evidences, where the evidence specifies how probable the data are for a given model. If $B_{12} > 1$, this means that the data are more likely to occur under model 1 than under model 2; one could also say the data support model 1. Some authors (among them Jeffreys, 1961; Robert et al., 2009; Kass and Raftery, 1995; Jarosz and Wiley, 2014) have categorized the range of values of the Bayes factor and stated how strong in their opinion the support for the particular model is. This is summarized in Tab. 4.3.

Table 4.3: Classification of the values of the Bayes factor and judgement of how large for a specific value the support for the particular model is according to Jeffreys (1961, Appx. B) and Kass and Raftery (1995).

Bayes factor	to Jeffreys means	to Kass & Raftery mean
1–3	worth a bare mention	worth a bare mention
3–10	substantial	positive
10–20	strong	positive
20–30	strong	strong
30–100	very strong	strong
100–150	decisive	strong
>150	decisive	very strong

In the prior, one can formalize his own preference for a model. For example, to avoid an accumulation of basis functions, we could choose a prior like the geometric or Poisson distribution, which give small models with few parameters a high probability. But even if we choose a uniform distribution as prior, simple models are favored over complex ones. As said above, the evidence is the probability of getting the observations under the assumption of K basis functions taking all possible combinations of parameters into account. Models with many parameters are very flexible and can represent any possible set of observations, so their probability mass will spread over a large area. In the small area, where also the simple model yields a good prediction, it is thus more probable after normalization; see Fig. 4.11. Herein the principle of parsimony in the sense of Occam's razor is realized in a natural way, and overparameterization is counteracted. Note that this is completely independent of a possible preference for simpler models being expressed by the prior.

For a better idea of which type of model is preferred in the Bayesian approach, we have to look at the formulas. I follow a similar line of reasoning as MacKay (2005). Let us assume for a moment that we are only interested in the number and scaling factors of the basis functions, while the point grid is fixed. The evidence for this problem is proportional to

$$p(\boldsymbol{y}|\beta_3) \propto \left(\frac{\sigma_{\boldsymbol{\beta}_1}}{\sigma_{0\boldsymbol{\beta}_1}}\right)^K \exp\{-\frac{1}{2}(\boldsymbol{v}^T\boldsymbol{v} + \frac{1}{\sigma^2_{0\boldsymbol{\beta}_1}}\boldsymbol{\beta}_1^T\boldsymbol{\beta}_1)\}. \tag{4.28}$$

This is a simplified version of Eq. (4.10), in which it was assumed that the prior and posterior are determined by their standard deviation $\sigma_{0\boldsymbol{\beta}_1}$ and $\sigma_{\boldsymbol{\beta}_1}$ only. Further, $\boldsymbol{v}^T\boldsymbol{v}$ is the sum of squares of the decorrelated residuals, and $\boldsymbol{\beta}_1^T\boldsymbol{\beta}_1$ is the regularization term. The evidence is the probability of

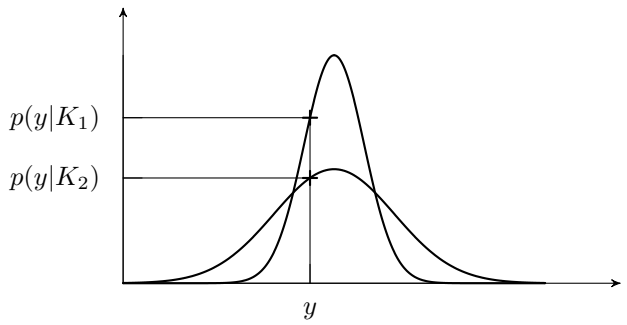


Figure 4.11: Illustration of how the principle of parsimony is incorporated in Bayesian model comparison (taken from MacKay, 2005, and modified). The figure shows the evidence of the observations for the model K_1 being simple in the sense that it has only few parameters and for the model K_2 being more complex. K_2 can be adapted to a wide range of different observations. As a consequence, it has less predictive power at the place where the observations actually lie.

the data given the model. For a high value of the evidence, the model has to describe the data well in the first place. What is meant is a good data fit and a smooth solution. If two models describe the data equally well, the Bayesian approach prefers simple models; very flexible models, whose parameters can vary a priori over a large range (large $\sigma_{0\boldsymbol{\beta}_1}$), get a small evidence. The same is true for models which yield a good approximation only when their parameters are precisely adjusted. For instance, in the regional analysis with the regular grid, the solution depends heavily on the choice of the regularization parameter. Such a solution would be penalized. Solutions for which the parameters need to be known only coarsely (large $\sigma_{\boldsymbol{\beta}_1}$) are supported. For large models (large K), this preference for stiff models and vague parameters is even stronger.

4.7.2 The label switching problem

The label switching problem arises when a model is invariant to permutations in the labeling of the model components. A prominent example is mixture modeling, where one tries to explain the data of a complicated distribution by a mixture of simpler distributions of the same kind. This is similar to our problem, as we also want to approximate data by a linear combination of equally-typed basis functions. Label switching may cause problems in the inference of parameter values. Consider the toy example of only two basis functions that shall be arranged on the real line. Switching the basis functions or switching the labels being assigned to the basis functions, which is totally the same, does not change the values of the likelihood function, as the model remains basically the same. The likelihood function will thus have two symmetric modes. If one assumes all values to be a priori equally probable, which is done here (cf. Sec. 4.3.3), and do so for each of the parameters, then also the posterior will be multimodal and symmetric. Having simulated the density by for example MCMC techniques, calculation of the Bayes estimate by averaging over the marginal distributions as usual will fail. The reason lies in the symmetry of the posterior, which has identical marginals and leads to identical estimates, which are useless for inference.

To solve for the labeling problem, one could define a special type of prior distribution, which enforces a specific order and thereby precludes label switching from the beginning (e.g. Richardson and Green, 1997). One could also sort the output of the sampling algorithm according to e.g. the numerical values of the parameters, which boils down to the same thing (Jasra et al., 2005). However, it is not always obvious how one should sort the parameters, for example when one works with points on a multidimensional surface instead just on the real line. Moreover, Stephens states that using identifibility constrains does not always completely solve the problem (Stephens, 2000). He proposed to use more sophisticated ways of relabeling.

An easy way out is to simply use the MAP instead of the Bayes estimator, that is to choose the parameter combination with the highest density. By summarizing the posterior density by the MAP estimator, some characteristics will get lost, such as a possible multimodality that cannot be explained by the permutation of labels. So using the MAP estimator cannot be considered as the general solution (Jasra et al., 2005).

Yet another way is to not work on parameter level but to use a function of the parameters that is invariant under the permutation of the basis functions; here, the labeling problem is not an issue.

4.7.3 The Bayes estimator

As explained in Sec. 4.7.2, the determination of the Bayes estimate requires the sorting of the random grids and the associated scaling coefficients. This is realized by sorting the points of the random grids in such a way that their absolute distance to the grid points of the MAP estimate becomes minimal; the scaling coefficients are then reordered accordingly. In the current implementation, after the obvious matches have been assigned, any permutation of the remaining points is tested in a brute force manner. However, this method turned out to be very time-consuming and also led to errors in case of multimodal densities.

Since it is not possible to average over vectors of different lengths, the Bayes estimator is determined conditional to a specific number K of basis functions. K can be specified by the mode of the marginal distribution of the number like for the MAP estimator, or one could identify adequate models by means of their Bayes factors. In contrast to the MAP, however, all samples of the selected number are used to derive the estimates.

As stated in the fundamentals chapter, Eqs. (3.17) and (3.22), the Bayes estimator corresponds to the expected value and can be approximated by the sample mean:

$$\hat{\boldsymbol{\beta}}_{2,B} = \int_{\mathcal{B}_2} \boldsymbol{\beta}_2 \, p(\boldsymbol{\beta}_2|\boldsymbol{y}) d\boldsymbol{\beta}_2 \tag{4.29}$$

$$\approx \frac{1}{N} \sum_{i=1}^{N} \boldsymbol{\beta}_2^{(i)}. \tag{4.30}$$

Here N is the number of samples for which $\beta_3 = K$. The calculation of the mean is realized by summing over the position vectors in Cartesian coordinates and back-projection onto the reference sphere.

To take advantage of the linearity of the scaling coefficients, let us write the expected value in the slightly different form

$$\hat{\boldsymbol{\beta}}_{1,B} = \iint_{\mathcal{B}_1\mathcal{B}_2} \boldsymbol{\beta}_1 \, p(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2|\boldsymbol{y}) d\boldsymbol{\beta}_1 d\boldsymbol{\beta}_2, \tag{4.31}$$

which can be understood as special case of Eq. (3.18). Applying the definition of conditional density, $p(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2|\boldsymbol{y}) = p(\boldsymbol{\beta}_1|\boldsymbol{\beta}_2, \boldsymbol{y}) p(\boldsymbol{\beta}_2|\boldsymbol{y})$, yields

$$\hat{\boldsymbol{\beta}}_{1,B} = \int_{\mathcal{B}_2} p(\boldsymbol{\beta}_2|\boldsymbol{y}) \int_{\mathcal{B}_1} \boldsymbol{\beta}_1 \, p(\boldsymbol{\beta}_1|\boldsymbol{\beta}_2, \boldsymbol{y}) d\boldsymbol{\beta}_1 d\boldsymbol{\beta}_2. \tag{4.32}$$

For a given point grid, i.e. conditional to $\boldsymbol{\beta}_2$, determining the scaling coefficients is a linear problem. Further, since it was decided on a conjugate prior, the problem becomes tractable. Specifically this

means that the expected value $\hat{\boldsymbol{\beta}}_{1|\boldsymbol{\beta}_2}$, which corresponds to the second integral equation, can be calculated analytically. This was derived in Sec. 3.3. Inserting this into the above equation, we get

$$\hat{\boldsymbol{\beta}}_{1,B} = \int_{\mathcal{B}_2} \hat{\boldsymbol{\beta}}_{1|\boldsymbol{\beta}_2}\, p(\boldsymbol{\beta}_2|\boldsymbol{y}) d\boldsymbol{\beta}_2 \tag{4.33}$$

$$\approx \frac{1}{N}\sum_{i=1}^{N} \hat{\boldsymbol{\beta}}_{1|\boldsymbol{\beta}_2{}^{(i)}}. \tag{4.34}$$

So for the Bayes estimator, we do not need random samples of the scaling coefficients, but we can take the mean of the analytical solutions of the random grids. Note that this is just another argument for the marginalization of the posterior density performed in Sec. 4.3.1, which allowed us to avoid sampling of the scaling coefficients.

4.7.4 The MAP estimator

The RJMCMC algorithm as formulated in this work yields samples of the marginal distribution of the point grid. What we are actually looking for is, however, the MAP estimate of the joint distribution, i.e. the sample that maximizes the joint density. We therefore recompose the density as follows:

$$p(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \beta_3|\boldsymbol{y}) = p(\boldsymbol{\beta}_1|\boldsymbol{\beta}_2, \beta_3, \boldsymbol{y}) p(\boldsymbol{\beta}_2, \beta_3|\boldsymbol{y}). \tag{4.35}$$

Remember from Sec. 4.3.1 that the posterior of the scaling coefficients given the point grid is a normal distribution, the expected value of which being $\hat{\boldsymbol{\beta}}_1$. As we lack the required samples, we just insert the $\hat{\boldsymbol{\beta}}_1$ into Eq. (4.35). This has the added benefit that for this specific point grid they give the highest density value under all possible values.

However, the computation of the MAP estimator for a problem of variable dimension is not straightforward. As explained earlier in Sec. 3.4.5, one must not compare densities of different dimension. Suppose we draw one or two random points from the uniform distribution on the unit sphere in a specific region. Then the density value for two points will be higher even if choosing one point from an infinite number of possible points is intuitively more probable. Depending on the size of the area, this ratio can be readily reversed. However, what has not been payed attention on is the fact that when comparing different dimensions the density value is totally meaningless. When the comparison is made on the basis of probability values, the outcome is in accordance with our intuition.

Instead of relying on equation (3.19), one could use histogram techniques to find the MAP estimator. For every parameter, the feasible region is divided into bins, the samples are assigned to the bins, and the bin with the highest number of samples is chosen as MAP estimator. In view of the size of the problem, the histogram would be huge and the calculation expensive. Moreover, in this specific context, the labeling problem would again lead to problems. A certain set of parameters has a certain probability, and according to this probability samples are allocated. As a result of the implementation of the sampling algorithm, the order of the parameters in the parameter vector will change over time. The total number of samples for a specific model will consequently be spread over $K!$ clusters, where K is the dimension of the model at which the algorithm is at the actual point in time. Using the histogram method would yield difficulties because for another model with larger K, the samples would be distributed over a larger number of clusters; the absolute number would thus decrease.

I simplify the problem by splitting up the joint density according to

$$p(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \beta_3|\boldsymbol{y}) = p(\beta_3|\boldsymbol{y}) p(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2|\beta_3, \boldsymbol{y}). \tag{4.36}$$

Based on this formula, model choice and parameter choice will be done separately. I look at the marginal density for the number being calculated from the samples, i.e. I ignore the other parameters and set up a histogram, which is very easy, as it is a discrete parameter. The MAP is then the model that occurs the most often:

$$\hat{\beta}_{3,M} \approx \operatorname{mode} \beta_3^{(i)}. \tag{4.37}$$

This is justified by the fact that the label switching problem does not affect the total number of samples for a model. Then the MAP of the parameter vector conditional on K is derived again using

$$\hat{\boldsymbol{\beta}}_{1/2,M} \approx \underset{\boldsymbol{\beta}_{1/2} \in \boldsymbol{\beta}_{1/2}^{(i)}}{\operatorname{argmax}} p(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2 | \beta_3, \boldsymbol{y}). \tag{4.38}$$

This is possible because within one and the same dimension, one can again compare the samples by means of their density values.

4.7.5 HPD regions for the point positions

When summarizing the a posteriori distribution of the parameters, one should also specifiy the uncertainty of the point estimate, e.g. in the form of an interval that contains the parameters with a certain probability. If the parameters represent position coordinates, one often displays the uncertainty graphically. This is what we also want to achieve here. For this purpose, I look at the marginal distribution per point and model only, i.e. $p(\boldsymbol{x}_k | K, \boldsymbol{y})$, which is a two-dimensional distribution and thus easy to display. Dependencies between the points, which are represented by the joint distribution of the grid points, will of course be disregarded this way. To get a complete picture, the marginal distributions of the individual points are superimposed; this is comparable to the usual approach of displaying geodetic networks, where the error ellipses for all points are shown in a single plot.

To represent the uncertainties, HPD regions are used (cf. Sec. 3.2). An HPD region is a region that contains the value of a quantity with a specific probability, while every point within the region has probability density higher or equally high than every point without the region. HPD regions are well suited for general (also multimodal) distributions. They always contain the point of the highest density and therefore seem a straightforward choice to specify the uncertainty of the MAP estimator. To calculate the HPD region for a distribution that has been simulated or, in other words, for which a sample of random values is available, one has to sort the samples corresponding to their probability density and choose the highest $1 - \alpha\%$ (see Hyndman, 1996; Chen and Shao, 1999). As the marginal density for the individual points is not known as closed form expression, we approximate the density function by the histogram of the samples. The HPD region is then calculated by collecting the highest bins while accumulating the corresponding probability until the accumulated value exceeds the chosen level of confidence. The x-values of the bins taken together then define the HPD region.

4.7.6 The Bayes estimator on solution level

The Bayes estimator for the gravity field functional f is defined as

$$\hat{f}_B(\boldsymbol{\beta}) = \iiint_{\mathcal{B}_1 \mathcal{B}_2 \mathcal{B}_3} f(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \beta_3) p(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \beta_3 | \boldsymbol{y}) d\boldsymbol{\beta}_1 d\boldsymbol{\beta}_2 d\beta_3 \tag{4.39}$$

(cf. Eq. (3.18)). Applying $p(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \beta_3|\boldsymbol{y}) = p(\boldsymbol{\beta}_1|\boldsymbol{\beta}_2, \beta_3, \boldsymbol{y})p(\boldsymbol{\beta}_2, \beta_3|\boldsymbol{y})$, we have

$$\hat{f}_B(\boldsymbol{\beta}) = \iint_{\mathcal{B}_2\mathcal{B}_3} p(\boldsymbol{\beta}_2, \beta_3|\boldsymbol{y}) h(\boldsymbol{\beta}_2, \beta_3, \boldsymbol{y}) d\boldsymbol{\beta}_2 d\beta_3 \tag{4.40}$$

with

$$h(\boldsymbol{\beta}_2, \beta_3, \boldsymbol{y}) = \int_{\mathcal{B}_1} f(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \beta_3) p(\boldsymbol{\beta}_1|\boldsymbol{\beta}_2, \beta_3, \boldsymbol{y}) d\boldsymbol{\beta}_1. \tag{4.41}$$

Because of the linearity of the scaling coefficients, $f(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \beta_3) = \boldsymbol{A}\boldsymbol{\beta}_1$. When we further consider that for a linear function g, $E(g(x)) = g(E(x))$, we can write

$$h(\boldsymbol{\beta}_2, \beta_3, \boldsymbol{y}) = \boldsymbol{A} \int_{\mathcal{B}_1} \boldsymbol{\beta}_1 \, p(\boldsymbol{\beta}_1|\boldsymbol{\beta}_2, \beta_3, \boldsymbol{y}) d\boldsymbol{\beta}_1 \tag{4.42}$$

$$= f(\hat{\boldsymbol{\beta}}_1, \boldsymbol{\beta}_2, \beta_3). \tag{4.43}$$

For the last step it was considered again that the expected value of the scaling coefficients given the point grid, $\hat{\boldsymbol{\beta}}_{1|\boldsymbol{\beta}_2,\beta_3}$, here just $\hat{\boldsymbol{\beta}}_1$, can be calculated analytically. Substituting this into Eq. (4.40),

$$\hat{f}_B(\boldsymbol{\beta}) = \iint_{\mathcal{B}_2\mathcal{B}_3} f(\hat{\boldsymbol{\beta}}_1, \boldsymbol{\beta}_2, \beta_3) p(\boldsymbol{\beta}_2, \beta_3|\boldsymbol{y}) d\boldsymbol{\beta}_2 d\beta_3 \tag{4.44}$$

$$\approx \frac{1}{N} \sum_{i=1}^{N} f(\hat{\boldsymbol{\beta}}_1^{(i)}, \boldsymbol{\beta}_2^{(i)}, \beta_3^{(i)}), \tag{4.45}$$

we see that the Bayes estimator for f boils down to just the mean of the gravity field solutions generated from the individual samples of the point grid, $\boldsymbol{\beta}_2^{(i)}, \beta_3^{(i)}$, and the corresponding analytically derived scaling coefficients, $\hat{\boldsymbol{\beta}}_{1|\boldsymbol{\beta}_2^{(i)},\beta_3^{(i)}}$, here just $\hat{\boldsymbol{\beta}}_1^{(i)}$. This time N equals the full number of samples. The Bayes estimator on solution level is different from the gravity field solution for the Bayes estimator on parameter level, i.e. $\hat{f}_B(\boldsymbol{\beta}) \neq f(\hat{\boldsymbol{\beta}}_B)$. The reason is that f is nonlinear in the point grid, so that the above mentioned linearity of the expected value does not hold. Moreover, in making use of the entire sample, the Bayes estimator on solution level also accounts for the uncertainty in the choice of the model.

4.8 The 8-point example revisited: point estimates, model averaging & credible regions

The estimators described in the previous sections were tested on the 8-point example; the results are presented in the following. Note that only 200,000 samples of the chain simulated in Sec. 4.6 were used for the present results, which was shown to be a representative sample of the target distribution. For the MAP as well as for the Bayes estimator, the estimate for the number of the basis functions has to be determined separately from the other parameters on the basis of the marginal distribution. The marginal distribution for the number of the basis functions indicates that 8 RBFs are most likely, although 9 and 10 RBFs are almost equally likely regarding their Bayes factors of respectively 1.22 and 2.67 in relation to the MAP model (Fig. 4.12(a)). Exemplarily for the models with 8 and 9 RBFs, Fig. 4.13 shows the probability distribution of the grid points by means of the marginal distributions of the individual points together with the 95%-HPD regions and the true and estimated point positions. As one can see in Fig. 4.13(a), the algorithm retrieves the true point positions within the range of accuracy; for the question of why I did not test the algorithm in an error-free scenario, see Sec. 4.5.

Table 4.4: Results for the 8-point example: differences of various estimates to the true field in terms of geoid heights [m]

	8-point example				**second data set**			
	K=8	**9**	**8–10**	**all**	**K=8**	**9**	**8–10**	**all**
MAP estimator	0.141	0.141			0.153	0.188		
Bayes estimator	0.126	0.149			0.155	0.193		
+ least squares	0.123	0.137			0.134	0.158		
on solution level	0.123	0.125	0.124	0.124	0.135	0.143	0.140	0.142

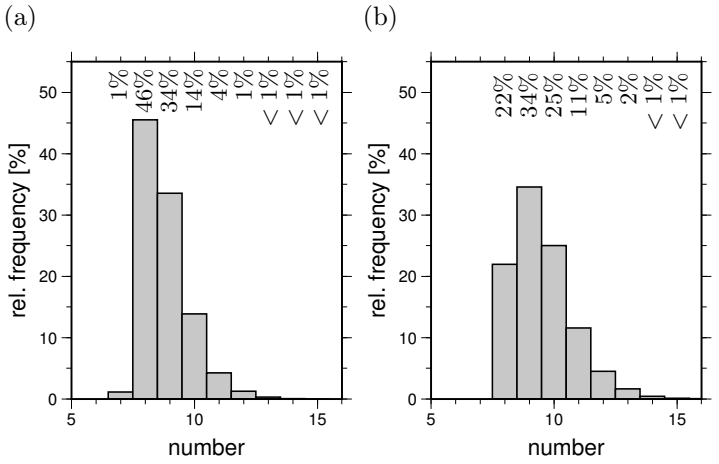
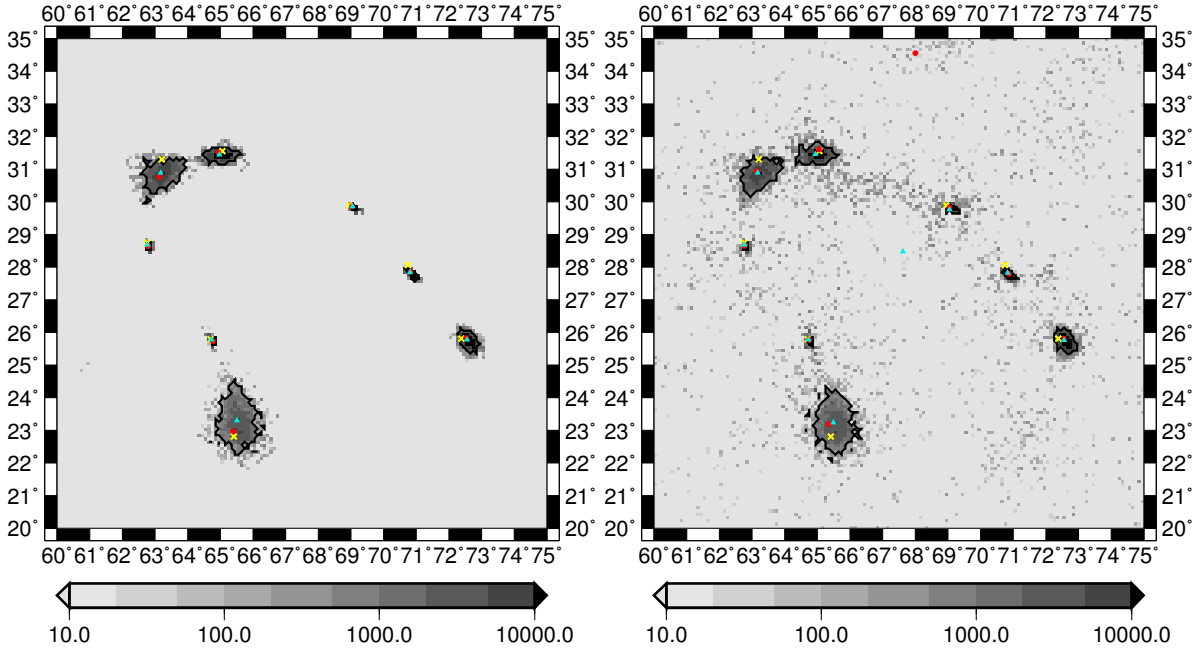


Figure 4.12: Marginal posterior of the number of RBFs for (a) the original data set and (b) the alternative data set

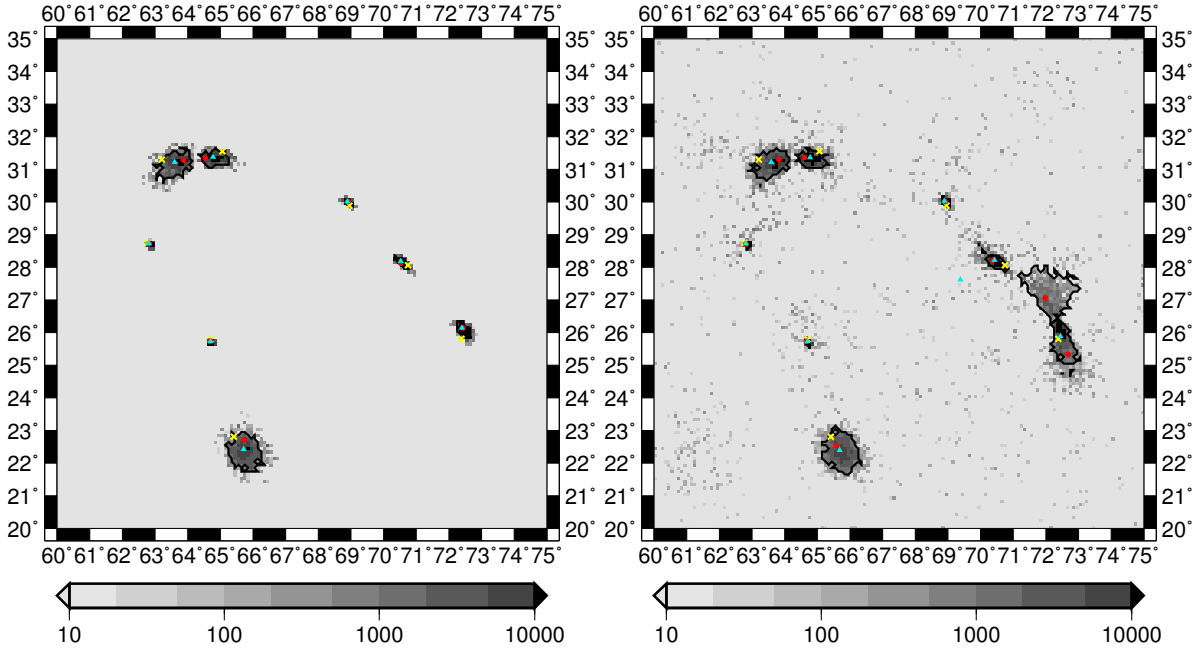
For comparison, a second data set was considered, which differs from the previous one in that another realization of noise was added. As a result, the marginal distribution for the number has its maximum not longer at the true number of 8 but at 9 RBFs (Fig. 4.12(b)). This might be surprising at first; however, the estimated number of basis functions, like any other ordinary model parameter, can of course deviate from the true value as a result of working with erroneous data. The 95%-HPD interval encloses the models with 8 to 10 RBFs, so the true value lies at least within the given limits of accuracy. Again, Fig. 4.14 shows the distribution of points for the models with 8 and 9 RBFs and the corresponding estimates. The results look very different from the previous example. While previously the 9th point occurred approximately uniformly distributed (Fig. 4.13(b)), it is now part of a constellation of 3 RBFs, which together represent the signal of actually 2 RBFs (Fig. 4.14(b)). Furthermore, sorting the random grids was not successful for this example. Because the true position of the point to the right lies halfway between the two points of the MAP, the random points were alternatively assigned to the one or to the other, whereas the respective other point was assigned the remaining uniformly distributed random point. Such kind of mixture distribution causes problems in calculating HPD regions or the Bayes estimator. Therefore, the MAP estimator conditional on 8 RBFs has been utilized for generating the results presented in Fig. 4.14(b). This way, sorting was indeed more successful, as by now the uniformly distributed random points were always assigned to the upper of the two MAP points. However, the multimodality of the posterior distribution is not fully remedied by the different sorting strategy but necessarily occurs when several different parameter constellations explain the observations well.

Apart from the MAP estimator and the Bayes estimator on parameter level, the Bayes estimator on solution level was computed for different (combinations of) models. Moreover, for the Bayes estimator on parameter level, an additional variant was considered, where the coefficients were



(a) $p(\boldsymbol{x}|K = 8)$ (b) $p(\boldsymbol{x}|K = 9)$

Figure 4.13: Marginal density for the position of a point in a common graph for all points. Supplementary, the black contour lines show the 95%-HPD regions, and the yellow cross, red dot and blue triangle mark the true point, MAP estimate and Bayes estimate, respectively.



(a) $p(\boldsymbol{x}|K = 8)$ (b) $p(\boldsymbol{x}|K = 9)$

Figure 4.14: As Fig. 4.13 but for the second data set

estimated from the observations in a least-squares sense. The estimates were compared to the true model in terms of differences in geoid heights; the rms values can be found in Tab. 4.4. In the documented results as well as in a number of further calculations, the MAP estimator did most of the time comparatively poorly. The Bayes estimator normally shows the smaller rms. The reason is that a Markov chain is not the appropriate procedure to find the highest point of a density function. It is, by contrast, a sampling approach, which generates samples according to the probability over the entire domain of the density function. Indeed, most samples fall into the area around the MAP compared to other areas of the same size. But in comparison to the entire domain, it is only a small portion of the samples, and one cannot expect to catch the point of the highest density precisely. The Bayes estimator corresponds to the expectation value of a probability distribution and can be approximated by the mean of the samples, which is what MCMC is typically used for. It is easy to see that in case of a normal distribution, where the expectation value matches the point of the highest density, the Bayes estimator is more precise because of the inherent averaging, while the MAP is only a snapshot, which comprises the whole uncertainty of an individual sample. As a test, I calculated the value of the density function for the Bayes estimator conditional to 8 RBFs and found that it would have been chosen instead of the MAP if it had been part of the sample. This is no surprise, as the density functions in the Figs. 4.13 and 4.14 resemble the normal distribution. To conclude, the fact that the Bayes estimator is superior to the MAP is indeed primarily a problem of the sampling being too sparse.

For 8 RBFs, where the sorting of the grid points succeeded without any problems, the Bayes estimator on parameter level is better than for 9 RBFs; here the rms values are higher (cf. Tab. 4.4). The reason lies probably in the process of sorting. The sorting was done according to the smallest absolute distance to the MAP. Where the assignment between the random grid points and the MAP grid points was unique or particularly clear, the points were assigned directly. All other points were considered in a brute force optimization, in which every permutation was tested, and the one with the smallest distance selected. Thereby it may happen that a point is assigned incorrectly and with it the corresponding scaling coefficient. Accordingly, the solution improved considerably, when for the Bayes estimator of the point grid, a new set of scaling coefficients was determined in the least- squares sense. So the degenerated accuracy might indeed be a problem of the sorting and should improve with a better sorting strategy. Moreover, for the second data set and 9 RBFs, the multimodality of the posterior because of the missing order of the basis functions could not be completely dissolved by sorting the way described above. Even when using the modified grid for sorting, as proposed earlier, we still get a mixture distribution, which is due to the nature of the problem where two completely different constellations of parameters give sensible results. The Bayes estimator on parameter level takes the mean of the probability distribution and thus ends up at a not very meaningful position (Fig. 4.14(b)). On the contrary, for the Bayes estimator on solution level, neither the order of parameters nor a possible mixture distribution matters. However different the parameters may be, they all provide a good (and thus also similar) solution, otherwise they would not have been chosen.

For the Bayes estimator on solution level, calculating the average over the 8-point model alone yields a slightly better rms value than the other variants for both data sets (Tab. 4.4). Admittingly, it is the true number of basis functions that has been used to generate the data of the simulation scenario, and the situation might change for real data applications, where a true model does not exist. For real data, the parameter space is much bigger, and much more samples would be required to get a representative picture of the target distribution. At the same time, autocorrelation is high, which decreases the effective sample size. As a consequence, it would be good to not rely on a single model but to average over e.g. the models which according to the theory of Bayes factors are almost equally likely.

To sum up, it is problematic to derive the MAP estimator from the output of a Markov chain. The problem will become even worse in a more realistic framework. Here even fewer samples would

fall into the area around the MAP if it is found at all. This phenomenon is referred to as the curse of dimensionality and can be visualized nicely by means of a normal distribution, where the majority of the probability mass concentrates in a certain radius at some distance from the origin as the dimension increases. A better way to find the MAP is to use a simulated annealing algorithm instead, which could easily be implemented in the frame of the present approach by just a small change in the target density. The Bayes estimator, on the contrary, has disadvantages in cases where the posterior distribution is multimodal. This problem might occur frequently for real data, as there will certainly be many different configurations of basis functions that explain the data similarly well. Moreover, sorting the random grids by testing any possible permutation in a brute force manner is very expensive. The problem to find the permutation that minimizes the distance to the MAP is very similar to the traveling sales man problem, for the solution of which efficient algorithms exist. In the form of the current implementation, however, it is not applicable in real data examples because of the enormous computational time effort. An attractive way out is to use the Bayes estimator on solution level, which unfortunately also has a drawback in that it does not yield a parametric solution. For computing another gravity field functional or to evaluate the field at another place, one has to make use of the entire probability distribution, i.e. the whole sample. Although being totally in line with the idea of the Bayesian approach, this is not very practical, e.g. for dissemination as a data product.

4.9 Adaption for real data application—the variance factor

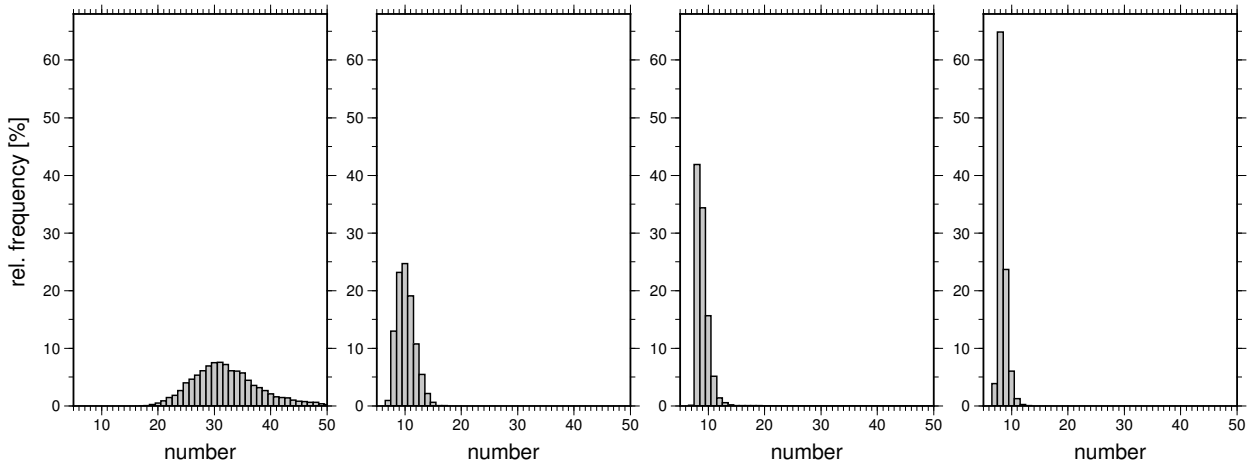
The determination of the variance factor in the prior of the scaling coefficients—or the reciprocal regularization factor in the least-squares method—has occupied me from the beginning. I have always been concerned that a good point grid will get a low marginal likelihood value if the variance factor is not adequately determined. On the other hand, we saw that the solution for the true grid was not sensitive to changes of the regularization parameter. This led me to believe that the value of the regularization parameter is not so important after all. Therefore, it was treated up to this point as a constant in all derivations, and in the previous calculations the variance component from the variance component estimation for the true grid was used. This approach shall now be examined. To this end, different values for the variance factor were specified, and a Markov chain of 50,000 steps was simulated for each. The result was that changing the variance factor changes the distribution of the number both the mode and variability (see Fig. 4.15). For a change in direction of the variance component for the standard grid, i.e. for a smaller σ, the algorithm chooses a larger model, and the uncertainty increases. For a larger σ, it is the other way around. One possible interpretation is that when the uncertainty of the prior information is higher, the algorithm deals with the complexity of the model in another way by reducing the number of unknowns. To sum up, even if the solution after optimization is largely unaffected by the choice of the regularization parameter, this does not apply for the optimization itself because the target density may change considerably. So if one wants to optimize the model, one has to determine σ in a clever way or to estimate it together with the model parameter. Otherwise, if one sets σ to an arbitrary value and lets the algorithm choose the number, there is not much of a difference compared to the case that one chooses the number and optimizes the regularization parameter as usual.

σ is estimated in the classical way using a hierarchical approach. This means σ is simply treated as further unknown parameter, called hyperparameter, and the prior of the scaling coefficients is extended by another prior, the hyperprior: $p(\boldsymbol{\beta}_1|\sigma)p(\sigma)$. The extended target density after marginalization reads $p(\boldsymbol{\beta}_2, \beta_3, \sigma|\boldsymbol{y}) = p(\boldsymbol{y}|\boldsymbol{\beta}_2, \beta_3, \sigma)p(\boldsymbol{\beta}_2|\beta_3)p(\beta_3)p(\sigma)$. Apart from the additional prior, it differs from the original expression, Eq. (4.5), in that the variance factor in the prefactor and in the argument of the marginal likelihood is now variable. The sampling is adapted as follows: in addition to the random grid, a new variance factor is proposed in each step; it is considered in the estimation of the scaling coefficients and in the evaluation of the sample by means of the odds

ratio. The form of the Bayes estimator does not change (the estimate however does). For the joint MAP estimator with σ, we have to look for the highest density value of the joint target density, $p(\boldsymbol{\beta}_1|\boldsymbol{\beta}_2, \beta_3, \sigma|\boldsymbol{y})p(\boldsymbol{\beta}_2, \beta_3, \sigma|\boldsymbol{y})$, using the two-step procedure introduced earlier.

Like Gelman (2006), I prefer to work with σ rather than with the precision parameter τ, which is often used in Bayesian statistics. Since I do not know much about the variance, I would like to use a non-informative prior. The inverse gamma prior, which is often used for this purpose, has problems with data sets in which small values of the variance factor may occur, as recognized by Gelman. He recommends using a uniform distribution for the standard deviation instead or, for few parameters and in regression, what he calls a weakly informative prior, such as the half-Cauchy distribution. This kind of prior is intended to suppress completely unreasonable solutions, but beyond that it is not very informative. Polson and Scott (2012) even advise to consider the half-Cauchy distribution as default prior for scale parameters in Bayesian hierarchical models. Following these recommendations, it is also applied here. For the scale parameter of the half-Cauchy distribution, Gelman recommends choosing a value that is slightly higher than what is expected for the standard deviation. The scale parameter corresponds to the width of the distribution at half of its height, so that larger values for σ are still possible.

σ is simulated in a random walk. To accelerate the sampling, only positive values are proposed using a normal proposal truncated at zero. This is easy to implement by sampling values from the non-truncated normal distribution until a value greater than zero comes out. The truncated density is equal to the non-truncated density apart from the different support and a normalization constant. Since the proposal is placed at different positions in the forward and backward step, the normalization constant does not cancel in the ratio and has to be considered to get the correct odds ratio. An alternative to this is to sample from a joint proposal for σ and the number. In this way, possible correlations would be taken into account. Since I do not want to change the proposal process for the number, I think of the proposal for σ as conditional distribution with mean and standard deviation depending on the given number. This would result in an independent step. In an independent step, the proposal does not cancel in the ratio. So here we have to account for both the normalization constant and the differing density values.



(a) $\sigma = 0.001$ (b) $\sigma = 0.003$ (c) $\sigma = 0.006$ (d) $\sigma = 0.01$

Figure 4.15: Distribution of the model parameter obtained with different values of the variance factor

5. Computations and results

In this chapter, gravity field models from GOCE data are presented, which were determined by means of the described three different techniques for gravity field recovery. The sections 5.1 and 5.2 show the results of using the global and regional parametrization as described in Ch. 2, and Sec. 5.3 shows the results of the reversible jump algorithm for the adaption of the resolution in regional modeling as described in Ch. 4.

5.1 The global gravity field model ITG-Goce02

This section is about the spherical harmonic analysis of GOCE data based on the short arc approach as described in Ch. 2. The resulting gravity field model was published under the name ITG-Goce02. The following sections are largely taken from the associated paper Schall et al. (2014). Only, some explanations were shortened to avoid repetition, and more recent comparison models were included.

5.1.1 Introduction

The GOCE (Gravity field and steady-state Ocean Circulation Explorer, Drinkwater et al. (2007)) satellite mission was launched in March 2009, with the goal to accurately measure the Earth's mean global gravity field with high spatial resolution. Its onboard satellite gravity gradiometer (SGG) observes the second derivatives of the gravitational potential, and GPS satellite-to-satellite tracking (SST) measurements are used for geo-locating the gradiometer observations as well as for the determination of the long-wavelength part of the gravity field. Since begin of the mission, several global GOCE gravity field models have been published by ESA's High-level Processing Facility (HPF, Rummel et al. (2004)) using three different processing strategies; an overview of the first results obtained by the three methods can be found in Pail et al. (2011). In principle, determining the coefficients of a spherical harmonic gravity field expansion from the linearly related gravity gradients is a straightforward procedure. However, GOCE gravity field analysis techniques have to deal with problems such as colored observation noise, instabilities due to signal attenuation and the polar gap, outliers, data gaps, and they have to be able to process a large amount of observations and solve for many parameters. Different analysis strategies cope with these challenges in different ways. We briefly review the second release of the official ESA models, as they cover almost the same time span of GOCE observations as the present study and will therefore be used for comparison in Section 5.1.3.

The direct approach (DIR, Bruinsma et al., 2010) is based on a least-squares solution with the SST part of the solution being calculated by the classical orbit perturbation approach (see, e.g. Reigber (1995)). SGG observations are bandpass filtered to include only frequencies within the gradiometer measurement bandwidth. To avoid instabilities in the gravity field solution caused by the polar gap problem, spherical cap regularization (Metzler and Pail, 2005) is applied using the GRACE gravity field model ITG-Grace2010s. As a result, prior gravity field information is introduced into the resulting GOCE model, such that it should not be regarded as a GOCE-only solution. The time-wise approach (TIM, Pail et al., 2010) considers the gradient and orbit observations as time-series measured along the satellite orbit and assembles and solves the full normal equation system; the energy integral approach (see O'Keefe (1957)) is applied for orbit analysis. For the gradiometer observations, full data decorrelation in the entire measurement spectrum is achieved by ARMA filtering. Instabilities are counteracted by applying Kaula regularization towards a zero model to the zonal and near-zonal coefficients and to the very high-degree coefficients. The resulting gravity

field solution thus represents a pure GOCE-only model, as no prior gravity field information is introduced. This is also the case for the space-wise approach (SPW, Migliaccio et al., 2011), which utilizes the spatial correlation of the gravity field. After applying Wiener filtering along the orbit to reduce the highly time-correlated noise, least-squares collocation is used to interpolate the gradient observations onto a spherical grid in mean satellite altitude. Regularization is implicitly included in this method by the choice of the covariance function applied in the collocation step. Afterwards the gravity field coefficients are derived by numerical integration. The SST part of the model is determined via the energy integral approach.

Here we have implemented a completely independent data analysis technique, the short arc approach. The concept of this approach was developed by Schneider (1968) for orbit determination and modified by Reigber (1969) to be applied in gravity field estimation. The approach was used by Mayer-Gürr et al. (2005) for the processing of CHAMP observations and then adapted by Mayer-Gürr (2006) for the computation of GRACE gravity field models. It was successfully applied in the determination of the ITG-Grace time series (Mayer-Gürr et al., 2010). Simulation studies using this approach for GOCE in the context of regional modeling were carried out by Eicker (2008), and Schall et al. (2011) reported the first application of the short arc approach to real GOCE gradient observations. Here, for the first time, we applied the approach to compute a full GOCE-only gravity field solution. The GOCE SST and SGG data are considered as time series along the orbit and are subsequently subdivided into short arcs. We then assemble a normal equation matrix individually for each arc, allowing for an effective decorrelation of the correlated observations, since it is possible to set up a full empirical covariance matrix for each individual arc. Moreover, the possibility to start a new arc after each data gap compensates for the problem of discontinuities in the observation series to some extent. No additional data is discarded after data gaps, as the determination of the empirical covariance matrix does not require any warm-up time, as is the case when using ARMA filtering. By arc-wise re-weighting of the observations, the influence of outliers can be reduced, as suggested by Kusche (2003) and applied by Mayer-Gürr (2006). Furthermore, the subdivision of the satellite observations into short arcs offers a straightforward possibility for parallelization and therefore has advantages for handling of the large amount of GOCE data. For the processing of the SST data, we applied the integral equation approach (Mayer-Gürr, 2006) to short arcs of the kinematic orbit data. Kaula regularization towards a zero model was introduced to account for the instabilities due to the polar gap problem. The combination of the different observation groups and the regularization was performed on the basis of the normal equations, while determining the relative weighting by variance component estimation (Koch and Kusche, 2002). A reasonable weighting requires a comprehensive description of the individual error behavior of both SST and SGG observations, which we believe to have achieved by the empirical covariance matrices.

In the following, we report the application of the short arc approach to GOCE gradiometer and orbit data, resulting in the global GOCE-only gravity field model ITG-Goce02 estimated from effectively 7.5 months of data. It is evaluated in space and frequency domain against the official ESA models covering the same time span (release 2), by comparison to global gravity field models (EGM2008 (Pavlis et al., 2012) and GOCE time-wise solution of release 4) and to independent data sets (GNSS/levelling data and altimetry observations).

5.1.2 Processing strategy

5.1.2.1 Data sets

The model presented here is based on GOCE observations from within the data period 2009/11/01 to 2010/06/30. Input data sets are specified as follows:

- EGG_NOM_1b: gravity gradients in the gradiometer frame (EGG_GGT), attitude information (EGG_IAQ), common mode accelerations (EGG_CCD),
- SST_PSO_2 (Bock et al., 2011): kinematic precise science orbits (SST_PKI_2) used as observations, epoch-wise covariance matrices for the positions (SST_PCV_2), reduced dynamic precise science orbits (SST_PRD_2) for geo-locating the gravity gradients.

The same models for tidal effects and non-tidal atmospheric and ocean variations were applied to reduce time variable gravity effects as suggested by the GOCE processing standards (European GOCE Gravity Consortium (EGG-C), 2010), but without additionally reducing time variations observed by GRACE. As we did not reduce the permanent part of the Earth tides, our model is computed in a zero tide system. Non-gravitational forces were derived from the common mode acceleration observations of the gradiometer.

5.1.2.2 Noise model

Since both the orbit position errors and the gradiometer observation errors are highly correlated in time, the treatment of the observation errors is crucial for GOCE data analysis. We set-up a full variance-covariance matrix for each arc, by estimating an empirical covariance function from the observation residuals referred to a reference model. The covariance function is computed as the inverse Fourier transform of the power spectrum of the residuals to guarantee positive definiteness. We use the Cholesky decomposition of the covariance matrix to decorrelate the observation equations. The reference model should represent a good approximation of the signal in the relevant frequency band, we therefore applied ITG-Grace2010s for the SST part and a preliminary in-house GOCE model for the SGG part. It has to be pointed out, that these reference models are only utilized for a realistic estimation of the error behavior but not for regularization, to keep the final solution as independent of non-GOCE information as possible. In order to take into account possible changes in the error behavior over time, a new covariance function is estimated for each of the four GOCE calibration periods (for SGG data) and for each month (for SST). The use of arc-wise covariance matrices implicitly assumes independence of neighboring arcs, which is not strictly valid. Therefore, further empirical parameters are introduced into the SGG noise model. These account for the long term error behavior and effectively decorrelate subsequent orbit arcs. In our approach an unknown constant per arc and gradient tensor element is estimated. For the SST part of the solution, one constant offset per arc is co-estimated for each of the elements of the common mode acceleration vector. Furthermore, arc-wise weight factors resulting in a down-weighting of arcs containing outliers or groups of outliers are determined using variance component estimation.

5.1.2.3 Analysis of orbit data

Kinematic orbit positions, representing a purely geometrical orbit solution without containing any gravitational force model, were used in a 10 sec sampling after low pass filtering (Mayer-Gürr, 2006) from the original 1 sec sampling rate in order to determine the SST part of the GOCE gravity field model up to spherical harmonic degree $n = 130$. For each of the short arcs of maximum 30 min length an observation equation was established, following the integral equation approach. Regarding the length of the arcs we rely on the experiences made in the CHAMP and GRACE processing, see Mayer-Gürr (2006). The epoch-wise covariance matrices of the individual satellite positions, which are provided together with the orbit product, are introduced into the estimation of the empirical covariance function explained in Section 5.1.2.2.

5.1.2.4 Analysis of gradiometer data

The three main diagonal components of the gravitational tensor (V_{xx} , V_{yy} and V_{zz}) are used as observations after being resampled to a 5 sec rate to reduce the amount of data. The observations are then grouped into arcs of 15 min length. The arc length represents a compromise between modeling the noise as good as possible and not losing too much of the signal through introducing too many empirical parameters. We prefer the reduced dynamic orbit product (interpolated to the same 5 sec sampling interval) for geo-locating the SGG measurements in the observation equations to avoid data gaps present in the kinematic orbit product. Attitude observations, together with an Earth rotation model according to the IERS2003 conventions (McCarthy and Petit, 2004), are used to rotate the observation equations from the Earth fixed frame to the gradiometer frame. The normal equation for the SGG part of the gravity field model is then set up for spherical harmonic coefficients up to degree $n = 240$.

5.1.2.5 Combination and regularization

To account for instabilities of the normal equation system caused by the downward continuation process and the polar gap problem, Kaula regularization towards a zero model is introduced for the spherical harmonic degrees $n = 5 \dots 240$. The Kaula regularization acts primarily on the higher spherical harmonic degrees and its contribution to the solution can be quantified as less than 2% up to degree 190. No prior information from an exterior gravity field solution is included, therefore our model can be regarded as a GOCE-only solution. The relative weighting of the two observation types (SST and SGG) and the regularization is optimally determined by variance component estimation as proposed by Koch and Kusche (2002).

5.1.3 Results

Using the data set and processing strategy described above, the global GOCE gravity field model ITG-Goce02 was determined up to degree $n = 240$. The solution is available via the International Centre for Global Earth Models (ICGEM). In the following, this model will be compared to the official ESA GOCE models of release 2 (GO_CONS_GCF_2_DIR_R2, GO_CONS_GCF_2_-TIM_R2, and GO_ CONS_GCF_2_SPW_R2), using almost the same data period, by evaluation against global reference models (EGM2008 and GOCE time-wise solution of release 4) and against independent data sets (GNSS/levelling data and altimetry observations).

5.1.3.1 Global comparison in frequency domain

Fig. 5.1 evaluates the different GOCE models in terms of difference degree amplitudes compared to the gravity field model EGM2008, which is based on a GRACE model combined with various terrestrial data sets and can therefore be assumed as having superior accuracy in the low and high degrees of the frequency spectrum. The lower degrees (below $n = 70$) appear to be dominated by errors in the SST data. Here ITG-Goce02 features smaller errors compared to TIM_R2 and SPW_R2, which we ascribe to the better performance of the integral equation analysis procedure compared to the energy integral approach, because the latter method reduces the amount of available orbit information, as also discussed in Ditmar and van Eck van der Sluijs (2004) and Pail et al. (2011). The low error in the low-degree range of DIR_R2 is a direct effect of the reference model, which results in the low degrees being more accurate compared to GOCE-only models. In the higher degrees (see zoom-in in Fig. 5.2), above approximately $n = 160$, ITG-Goce02 performs very similarly to the time-wise model, whereas the direct and especially the space-wise approach show

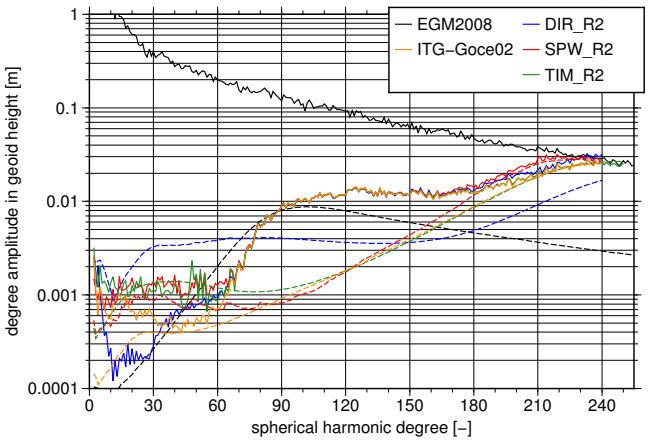


Figure 5.1: Comparison of different GOCE models to EGM2008 in terms of difference degree amplitudes (solid lines), omitting near zonal coefficients to exclude the effect of the polar gap (Sneeuw and van Gelderen, 1997). Corresponding formal errors are plotted as dashed lines. Figure taken from Schall et al. (2014, Fig. 1).

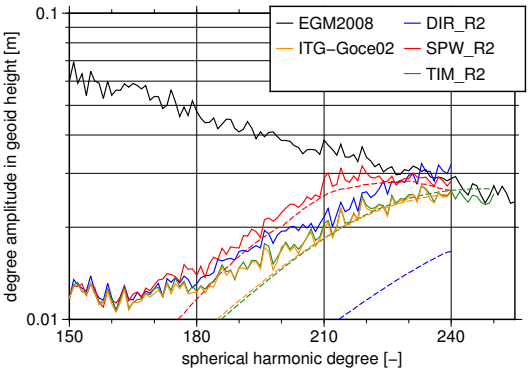


Figure 5.2: The same as Fig. 5.1, but zoomed-in into degrees above $n = 150$ (taken from Schall et al., 2014, Fig. 2).

slightly larger differences. The formal errors of ITG-Goce02 match the difference to EGM2008 very well in the frequency range $n = 30 \ldots 60$ and above $n = 200$, which is probably due to our use of full empirical covariance matrices for each short arc. In the frequency band $n = 60 \ldots 180$ the differences in the degree amplitudes are dominated by errors in EGM2008, caused by low-accuracy terrestrial data in some regions, which is also supported by the large formal errors of the reference model in this part of the spectrum (see also Hashemi Farahani et al. (2013)). The ARMA filtering process applied in the time-wise approach and a Monte Carlo approximation of the estimation errors carried out in SPW_R2 lead to a similarly realistic formal error spectrum. In the direct approach there is no agreement between formal errors and differences to the reference field, which could be related to an insufficient modeling of the observation noise.

5.1.3.2 Global comparison in space domain

For each of the GOCE models, differences were calculated compared to the GOCE-only model of release 4 following the time-wise approach (GO_CONS_GCF_2_TIM_R4). This model covers a significantly larger time span with an effective data volume of approximately 26.5 months and can therefore be regarded as reference solution of superior accuracy. It should be mentioned that for TIM_R4 also the integral equation approach has been applied to derive the SST part of the solution. The global differences excluding the polar gap (i.e. $-80° < \varphi < 80°$) were calculated up to degree $n = 200$ and are displayed in Fig. 5.3. A comparison of the different plots shows smaller differences for ITG-Goce02 and TIM_R2 than for SPW_R2 and DIR_R2. RMS values were derived for the global differences and for specific regions in the Himalayan ($70° < \lambda < 100°, 20° < \varphi < 40°$), in the Pacific ($-150° < \lambda < -90°, -60° < \varphi < 0°$), and in the Indian Ocean ($60° < \lambda < 105°, -45° < \varphi < -15°$). The results in Tab. 5.1 again show smaller differences for ITG-Goce02 and TIM_R2 with TIM_R2 being slightly better in the Himalayan and ITG-Goce02 featuring smaller discrepancies in

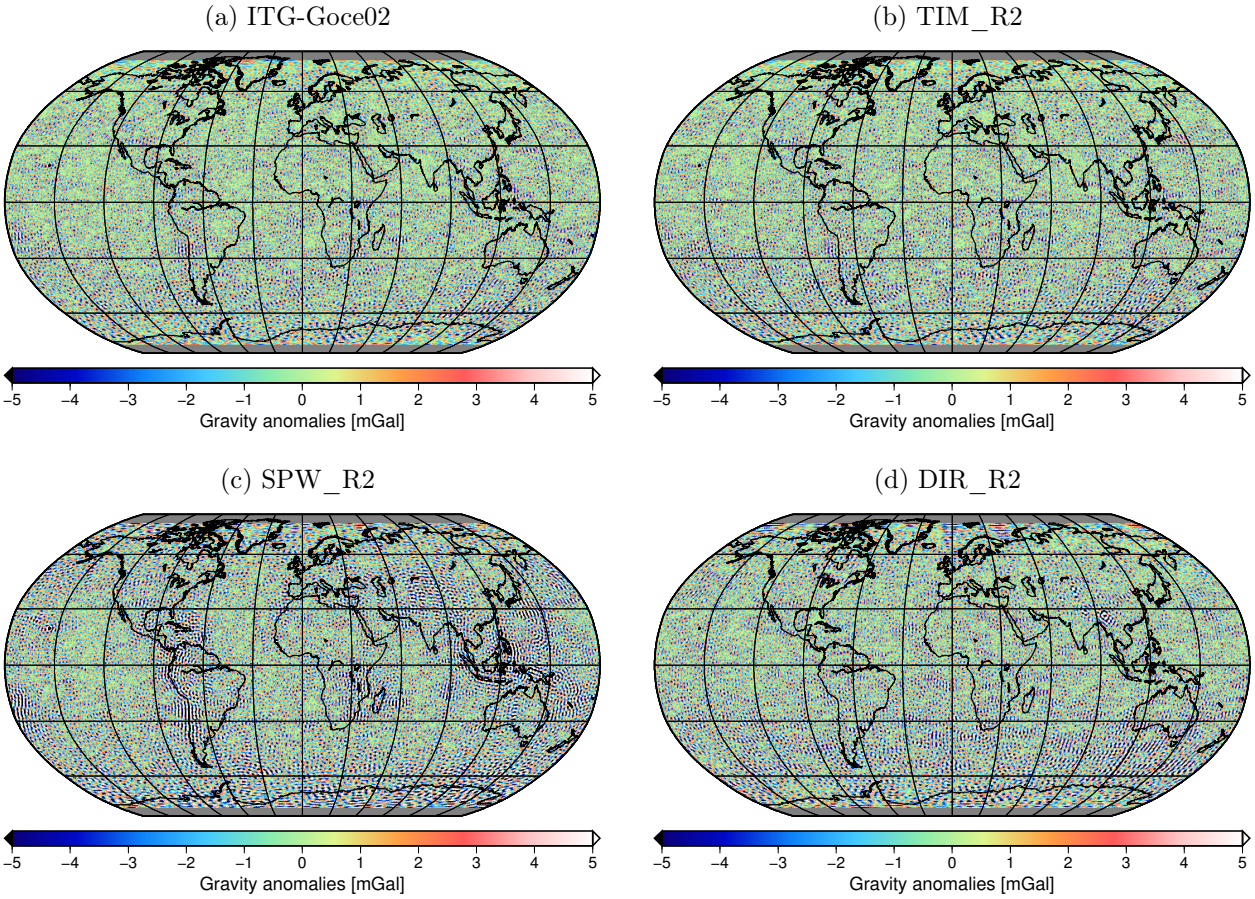


Figure 5.3: Differences in terms of gravity anomalies between the different GOCE models and the release 5 GOCE model GO_CONS_GCF_2_TIM_R5. Models are evaluated up to degree $n = 200$. Figure updated from Schall et al. (2014, Fig. 3).

the oceanic areas. This can be related to the slightly stronger regularization of ITG-Goce02 which allows a better fit to the smooth ocean signal and results in a marginal signal loss in the rough mountainous regions. Very similar results are obtained when the release 4 model computed by the direct approach (GO_CONS_GCF_2_ DIR_R4) is used as reference model (not shown here).

Area	ITG-Goce02	TIM_R2	SPW_R2	DIR_R2
global	1.93	1.99	2.90	2.35
Himalaya	1.77	1.74	3.72	3.91
Pacific Ocean	2.01	2.19	2.20	2.49
Indian Ocean	1.94	2.08	2.32	2.50

Table 5.1: RMS of differences between different GOCE solutions (release 2) and the release 5 GOCE model GO_CONS_GCF_2_TIM_R5. Models are evaluated in terms of gravity anomalies [mGal] up to degree $n = 200$. Table updated from Schall et al. (2014, Tab. 1).

5.1.3.3 Comparison to independent data sets

GNSS/levelling data

Fig. 5.4 shows the RMS differences between height anomalies, derived from the different GOCE gravity field models and 924 GNSS/levelling point measurements over Germany (Rülke et al., 2013)

after removal of a constant bias, displayed as a function of model resolution (expansion degree). GOCE models were truncated at different spherical harmonic degrees in 5-degree steps and filled up with coefficients of EGM2008 up to degree $n = 2190$. This investigation shows again a good agreement of ITG-Goce02 with the time-wise model, while DIR_R2 and SPW_R2 exhibit significantly larger differences above approximately degree $n = 140$. Tab. 5.2 shows the RMS values between the GOCE solutions and different international GNSS/levelling data sets provided by ICGEM (2013). Here the models were not filled up with EGM2008 coefficients; the values thus include the full omission error of the GOCE models. Results are broadly in agreement with the Germany data set and confirm the closeness of ITG-Goce02 and TIM_R2.

Area	ITG-Goce02	TIM_R2	SPW_R2	DIR_R2
USA (6169 pts.)	0.429	0.436	0.457	0.443
Canada (1930 pts.)	0.354	0.355	0.376	0.374
Europe (1235 pts.)	0.434	0.434	0.473	0.449
Australia (201 pts.)	0.371	0.375	0.376	0.391
Japan (816 pts.)	0.511	0.515	0.553	0.519

Table 5.2: RMS of differences between different GOCE solutions and GNSS/levelling points in terms of geoid heights [m]. Table taken from Schall et al. (2014, Tab. 2).

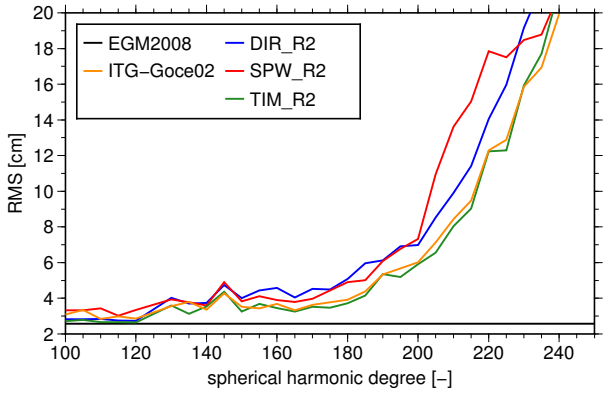


Figure 5.4: Comparison of different GOCE models to GPS/levelling data in Germany. Truncation degree of models is given on the abscissa, above this degree each model is filled up with coefficients of EGM2008. Figure taken from Schall et al. (2014, Fig. 4).

Altimetry observations in the North Sea

Geoid heights computed from the different GOCE models were compared to altimetrically observed sea surface heights at 112 cross-over points of ERS-2 and ENVISAT in the North Sea. To account for the time-varying dynamic topography and tides, water heights from the numerical ocean circulation and tide model BSHcmod (Dick et al., 2001) were removed from the altimetry sea surface heights. BSHcmod exhibits a higher spatial resolution compared to global ocean models and includes also non-linear tides, which play an important role in the North Sea. The RMS values of the differences between GOCE and altimetrically observed geoid heights, after reducing a constant offset, are again plotted as a function of model cut-off degree in Fig. 5.5 (with higher degrees having been filled up with EGM2008). The level of disagreement between satellite-derived geoid and in-situ observations is higher than in the GPS/levelling experiments, owing to the limitations of the technique and the possibly insufficient reduction of time-variable effects. But results point in the same direction:

ITG-Goce02 has similar differences as TIM_R2 in the higher degrees above approximately $n = 140$. While DIR_R2 shows larger discrepancies, SPW_R2 has significantly smaller differences in the very high degrees above $n = 220$, which may be related to an over-regularization as indicated in Migliaccio et al. (2011).

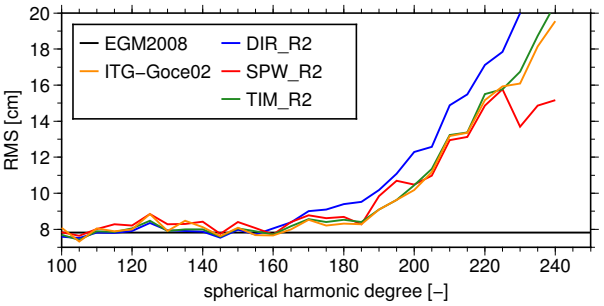


Figure 5.5: Comparison of different GOCE models to altimetry data in the North Sea. Truncation degree of models is given on the abscissa, above this degree each model is filled up with coefficients of EGM2008. Figure taken from Schall et al. (2014, Fig. 5).

5.2 Regional models with a uniform arrangement of basis functions

Having shown in the previous section that our global GOCE solution is competitive in the sense that it has comparable accuracy to the official ESA products, in this section the gain of the regional method following Eicker (2008) is demonstrated as compared with the global spherical harmonic solution. The results were already published in Eicker et al. (2014), where the regional method was applied for the first time and to the best of my knowledge as first regional approach in general to GOCE-level-1b data. The article included the results of three different test sites. Here I shall limit myself only to the one in the area of the South Sandwich deep sea trench ($-45/-20/-60/-35$; cf. Eicker et al. 2014, Sec. 3.3.2), because due to the inhomogeneous signal content, it is a challenge even for regional modeling and therefore suitable to demonstrate the value of using several regularization areas on the one side and of adapting the model resolution as in the following section on the other side. The validation is carried out by comparing with EGM08.

5.2.1 Processing strategy

The regional approach described in Ch. 2 was applied to GOCE gradiometer data. The same data and background models were used as for the calculation of the SGG-part of the global spherical harmonic solution ITG-Goce02 (see Sec. 5.1), and in principle also the same processing was applied except that the parametrization was changed from spherical harmonics to radial basis functions. ITG-Goce02 was reduced as reference field up to d/o 160. The truncation degree was chosen relatively high to avoid that oscillating signal enters the observations as a result of an early truncation within the degrees affected by the GOCE polar gap. The expansion of the RBF-kernel was limited to degree 240 corresponding to the maximum degree of the spherical harmonic solution. Up to degree 160, the shape coefficients were determined by the formal error degree variances of the reference field ITG-Goce02 and above by Kaula's rule of thumb. As proposed by Eicker (2008), a triangular vertex grid of level 81 was employed as point grid for the arrangement of the basis functions, which is approximately equivalent to the resolution of the global solution. I want to point out that in practice this choice is somewhat arbitrary. Often the best level of resolution is tested by comparing with a model of superior accuracy, so that the standard uniform point grid in some sense can be

regarded as being optimized by hand. Only GOCE data within the area of interest were used plus an additional margin of 3 degree to mitigate edge effects. For the same reason, the point grid was set up in an area that overlaps the area of the observations by another 3 degrees.

5.2.2 Results

While Kaula regularization in spherical harmonic analysis suggests a global mean signal content as prior information and thereby leads to errors in rather smooth or rough areas, in regional analysis the prior information can be adapted for the investigated area. The use of a regional mean signal content as prior information may improve the solution e.g. by about 50% for a homogeneous area in the open ocean (cf. Eicker et al., 2014, Sec. 3.3.1). In contrast, the chosen test site in the area around the South Sandwich Trench in the South Atlantic Ocean exhibits very inhomogeneous signal characteristics; the southern part of the region shows a strongly high frequency gravity field signal along the deep sea trench, and in the northern part the signal is rather smooth. In such an inhomogeneous region, in principle the same problems occur in regional analysis as when applying Kaula regularization on the spherical harmonic solution, and indeed the regional solution with only one regularization term did not improve with respect to the global solution ITG-Goce02. The signal of the global solution is presented in Fig. 5.6(a) and the differences to EGM08 in Fig. 5.6(c) with an rms of 6.30 mGal. To get more out of the given data, the region was subdivided into two regularization areas north and south of 51 degree, and a variance component was estimated for each of the two subregions (as was proposed by Eicker 2008). The resulting gravity field model is presented in Fig. 5.6(b) and the differences to EGM08 in Fig. 5.6(d). The rms of the differences amounts to 5.10 mGal, which is an improvement of 19% compared to the global solution. Looking at the two regularization areas individually, the rms improves from 8.17 to 7.59 mGal in the southern part (7%) and from 5.28 to 3.51 mGal in the northern part (34%). To sum up, the regional method following Eicker (2008) causes an improvement of the prior information—with the division into distinct regions, this also applies to inhomogeneous areas—and thereby leads to an improvement in the individual subregions.

Table 5.3: RMS in terms of gravity anomalies and geoid heights of the differences between GOCE solutions and EGM2008 for the South Sandwich Trench separately for the complete patch (N+S), the northern part (N) and the southern part (S). For each functional the relative improvement achieved by the regional refinement is given. Taken and adapted from Eicker et al. (2014).

	ITG-Goce02		Regional		Improve	
	Δg	N	Δg	N	Δg	N
Area	[mGal]	[cm]	[mGal]	[cm]	[%]	[%]
N+S	6.30	19.14	5.10	15.51	19	19
N	5.28	16.35	3.51	11.07	34	32
S	8.17	24.36	7.59	22.61	7	7

5.3 Regional models from the optimization of point grids

We saw in the foregoing section that in regional gravity field analysis a better adaption of the prior information can be achieved, thereby reducing errors that are related to an inappropriate regularization. With the division into several regularization areas as proposed by Eicker, this also works for inhomogeneous areas. However, the division has to be specified explicitly, which is not always easy like for the southern part of the patch—the new study area used in the following. I

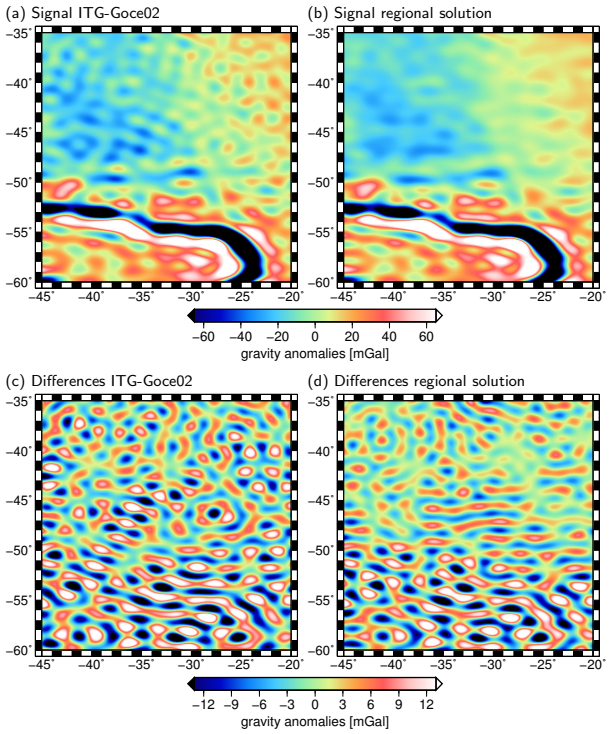


Figure 5.6: South Sandwich Trench: Signal (top) and differences compared with EGM2008 (bottom) of the global model ITG-Goce02 (left, RMS: 6.30 mGal) and the regional refinement using two different regularization parameters for the northern and southern part (right, RMS: 5.10 mGal), taken from Eicker et al. (2014).

chose a completely different approach. I do not try to improve the prior information by a further questionable division of the study area, but I adapt the model resolution and thereby hopefully stabilize the problem, which should make the solution less dependent on the prior information.

For the purpose of adapting the model resolution, in this chapter the number and positions of the basis functions are estimated in addition to the usual model parameters (i.e. scaling coefficients and variance factor). In contrast with the previous chapter, the point grid for the arrangement of the basis functions is thus not simply defined in advance but simulated in the course of a Markov chain. Furthermore, the variance factor is not determined by means of variance component estimation but simulated along with the point grid. However, estimating the scaling coefficients for a (fixed) random grid in a particular step of the chain works basically the same as before. Only, the area of the RBFs was chosen slightly smaller using an additional margin of 1.5° with respect to the area of the observations instead of 3° as before. The additional margin is meant to avoid edge effects, but these I think are comparatively small in satellite data analysis anyway. Moreover, the basis functions used in this scenario are relatively narrow, so that also a narrower margin is sufficient, as was confirmed by a test computation. Therefore, it makes sense to further restrict the feasible area of the RBFs (and to adapt the proposal accordingly), because meaningless proposals are rejected with high probability, slowing down the simulation of the chain.

Again the same data were used as in the calculation of the other models. However, as already mentioned, the new method was applied only to a small part of the previous study area, so the data were cut to a slightly different region. The reason is that an equation system has to be set up and solved in every step. This is time-consuming considering that a good Markov chain often requires several hundred thousand steps. Setting up the design matrix and, in this context, adding up the Legendre polynomials is the most time-intensive part. A larger study area would mean more observations and parameters and thus a larger design matrix, which would increase the computational time effort. Therefore, I have to content myself with an area of moderate size, although there is certainly still room for improvements in the implementation.

5.3.1 Processing of the 1st Markov chain and its convergence

In the following, the settings for the simulation of the first Markov chain are specified. Any configuration of basis functions was assumed a priori to be equally likely, provided that the nodal points lie within a reasonable distance from the observations and there are not too many of them. Accordingly, the prior of the locations of the basis functions was defined with the help of a spherical uniform distribution limited to the area of the observations plus 1.5°. For the prior on the number of basis functions, a uniform distribution in the range of 1–708 was chosen, whereby 708 is the size of the standard grid when using the wider margin of 3°. The hyperparameter σ in the prior of the scaling coefficients was simulated along with the other parameters, and I assigned a prior in the form of a half-Cauchy distribution. The included tuning parameter γ, which is recommended to be chosen slightly higher than what would be expected, was set to 0.02, which is slightly higher than the factor estimated in Ch. 5.2 for the southern part of the patch. This makes sense because the same kernel function was used as for the regional analysis in the last chapter and the prior information will not fundamentally change.

The first simulated chain was started from a triangular vertex grid of level 81—the same as also used for the calculation of the regional models in the previous chapter. Cutting the grid to the feasible area resulted in a uniform arrangement of 562 points. In the first step of the implemented approach, a random perturbation of the initial grid was proposed choosing from among three different move types with equal probability. The choice was made between a birth of 10 RBFs, the corresponding death of 10 RBFs and a move (i.e. a change in the position) of 30 RBFs. In a birth step, the positions of the new basis functions were simply proposed independently of the current configuration uniformly

distributed over the feasible area. In a move step, a Fisher distribution with a spread of 0.1° was applied. Additionally, in every step a new variance factor was proposed using a Gaussian proposal centered at the current value with a standard deviation of 0.001°. The aforementioned settings were determined in a series of trial runs so that an appropriate number of samples were accepted. In particular, a simultaneous birth/death of several basis functions proved to be necessary for a reasonable mixing. In a move step, the step width can generally be adjusted by either the number of functions moved or the amount of movement. Here it was decided to move many functions just a bit, as recommended for computing time intensive problems. After the random grid had been created, the normal equations were set up and solved in a regularized least squares adjustment using the proposed variance factor for regularization. Then it was decided whether to accept the proposed sample or to proceed with the current one on the basis of the odds ratio, and the whole procedure started again from the random perturbation of the point grid. This was repeated for overall 800,000 iterations.

Having been started from a regular point distribution, the chain may require some time to reach an area of at least moderate density. Nevertheless, there is no obvious running-in effect, like the exponential decrease in the simulations, in the time series plots for the number, variance factor and rms (Fig. 5.7). One can see, however, that the stochastic behavior changes over time. The sinusoidal oscillations at the beginning of the time series plots vanish, and after around 350,000 steps the chain looks more or less homogeneous. This is only interrupted by a systematic effect around the sample 600,000, which may be explained by the fact that at this point the lowest number of the entire chain—that is 182 RBFs—was adopted and the chain needed some time to find the way back from this improbable constellation to the right point positions. Thus, regardless of the mentioned distortion, it was decided to use a burn-in phase of 400,000 steps, in the hope of thereby removing the dependency on the start position.

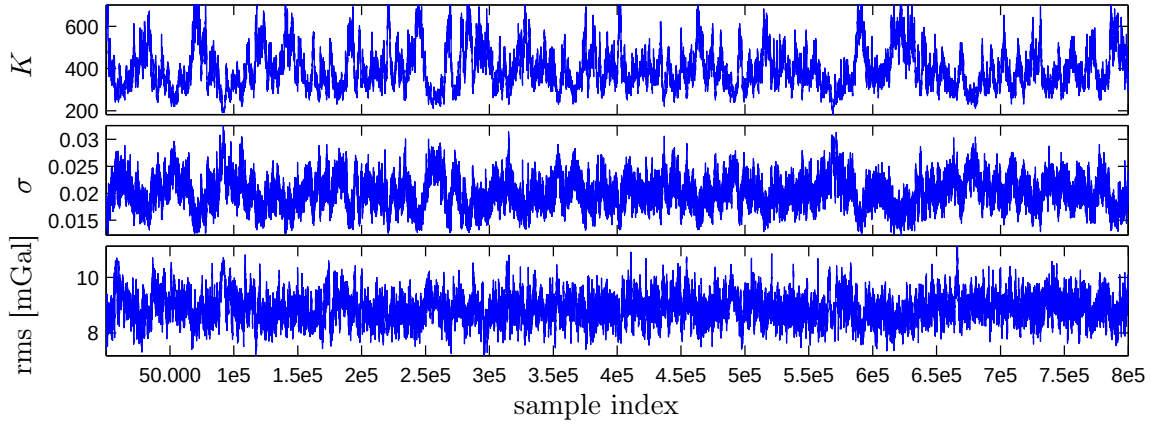


Figure 5.7: Time series plots for the number of basis functions (*top*), the variance factor (*center*) and the rms of the differences between the individual solutions and EGM08 in terms of gravity anomalies (*bottom*) for the 800,000 samples of the first simulated Markov chain

5.3.2 Does the point grid align with the gravity field structures?

In the introduction of this thesis, the question has been raised whether the point grid for the arrangement of the basis functions in regional gravity field analysis should not optimally be aligned with the structures of the gravity field. In the following, the question will be answered on the basis of the optimization results using the output of the first Markov chain. It should be emphasized again that the first chain was started from the uniform standard grid and a uniform distribution was used to propose the positions of new basis functions during a birth step. Hence, the chain was

not forced against the gravity field neither by the initial grid nor by the proposal process, although these settings only influence the convergence and mixing of the chain and not the actual target distribution anyway. Fig. 5.8 shows the distribution of the grid points for the example of 362 points as the result of sorting all grids with the corresponding number of points into a common histogram and applying the normalization of a probability density function. The reason for considering only grids with a specific number of points is that the distribution of the grid points may differ for different numbers. One can easily imagine that in the case of many basis functions, the actual position of a function is less important, and the probability distribution accordingly flatter, since errors can be compensated for by other functions. The resulting distribution of the grid points is obviously correlated with the gravity field signal. There is a close resemblance to Fig. 5.9, showing a probability density function derived from a geoid model in the way described in Sec. 3.6.3. Thus, the point grids that can predict the observations within their limits of accuracy follow the structures of the gravity field, i.e. the points concentrate at places where the signal is strong. This naturally also applies to the optimal grid, which lies somewhere in between. From this it can be concluded that a signal-adapted point grid is indeed the best choice for the arrangement of the basis functions in regional analysis, at least for our RBF model and the given data of nearly uniform coverage and accuracy.

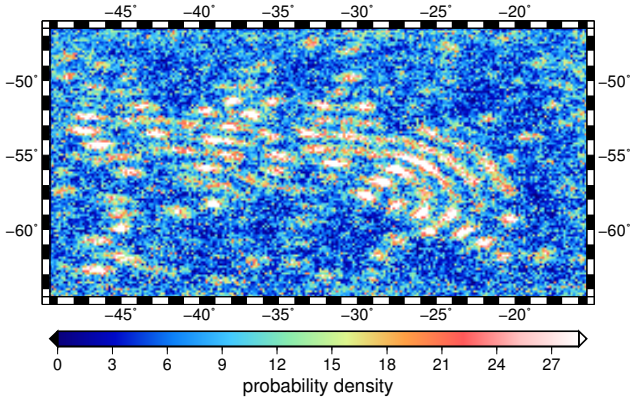


Figure 5.8: Empirical density for the distribution of the grid points with respect to the unit sphere (statistics: min=0, max=96.90, mean=9.50, rms=11.81)

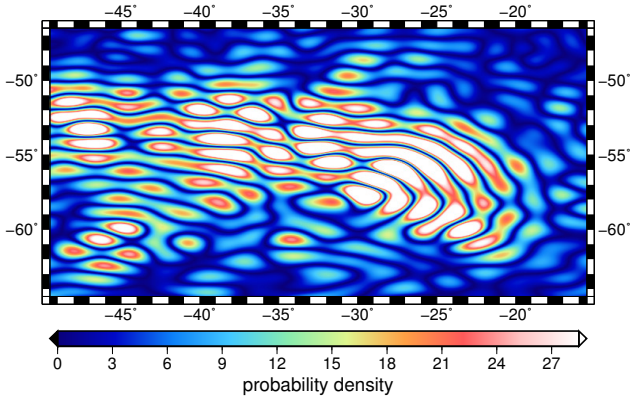


Figure 5.9: Model-based density on the unit sphere calculated using geoid heights from EGM08 in the range $n = 161..240$ evaluated on a sphere of radius $R = 6378\,\text{km}$ (statistics: min=1.76e-05, max=81.40, mean=9.51, rms=13.76)

5.3.3 Test of different proposal densities

As we saw in the previous section, the distribution of the grid points bears some similarity with the gravity field signal. It therefore seems to be obvious to exploit available information about the gravity field to improve the proposal process. It was with this in mind that a new proposal distribution for the positions of new functions in a birth step was developed, which involves information from a gravity field model (see Sec. 3.6.3 for the definition). In the following, it will be examined how well this new proposal distribution works in comparison with the uniform distribution used so far.

First, I tested the density that looked so similar to the distribution of the grid points (see Fig. 5.9). I started a short test run of 50,000 steps from the MAP estimate of the first chain. In this way, I started from a sample that is typical for the target distribution, thereby avoiding further running-in effects. As said earlier, a gravity-based distribution (model: EGM2008, range: $n = 161..240$ corresponding to the residual signal content, functional: geoid) was applied as proposal for the birth step in replacement of the uniform distribution; all other settings remained unchanged. The result was that the proposals got a much higher marginal likelihood value on average than when using the uniform distribution (Tab. 5.4). This shows that the new proposal process inspired by the gravity field creates point grids, with the help of which the observations can be approximated comparatively well. However, despite the higher likelihood values, which should actually encourage acceptance as can easily be seen from the odds ratio, the acceptance was much worse than with the uniform distribution. This may be explained by the fact that the proposal density values were even higher, indicating that the proposals, although in principle being reasonable, occurred too often with respect to the target distribution. For an independent chain in order to get a good mixing, one would ideally sample from the target distribution itself. Since in the current implementation only a few points are added in every step, it would be ideal to sample from the conditional density given the actual point distribution. This would yield a high acceptance of birth and death steps according to the ratio of the model probabilities. Although this is rather a theoretical reasoning, I expect a supposedly better proposal distribution to achieve a high acceptance—but at least a higher acceptance than the uniform distribution. The tested proposal distribution is therefore by no means optimal.

The problem is that after the convergence of the chain the point grid does not change fundamentally anymore but only as far as allowed by the accuracy of the observations. Although the tested proposal density contains information about where a point should generally be located, it does not consider whether a point already exists at this place. Therefore, in order to dampen the peaks of the proposal, I considered a linear combination of a gravity-derived and a uniform distribution as proposal. This is easy to implement by just setting up two alternative birth steps, leading to the formulas in Sec. 4.4.5. I tested two different linear combinations with a gravity-derived distribution using geoid heights. Moreover, vertical deflections were tested as another gravity field functional for comparison. The average likelihood values for the combined proposal distributions came out to be smaller than for the purely gravity-derived distribution but were still higher than for the uniform distribution (Tab. 5.4). One can see from the variants where the distribution derived from geoid heights entered with $1/2$ or $1/3$ that the likelihood values become the smaller the lower the contribution of the gravity-derived distribution. While the acceptance of the $1/2$-variant was not yet satisfactory, the acceptance of the $1/3$-variant was reliably higher than for the uniform distribution. Using the other gravity field functional did however not constitute an improvement.

Based on the experience gained from the test runs, another long Markov chain of 400,000 steps was simulated for the most promising proposal $1/3\,g$. Instead of simulating the variance factor in a random walk process as before, it was simulated independently making use of the knowledge already available about the distribution. Knowing that the variance factor and the number of RBFs are correlated, which will become obvious later on in this chapter in Fig. 5.15, a Gaussian proposal

Table 5.4: Statistics on the average marginal likelihood ratio value (*likelihood ratio*) and the acceptance rate (*acceptance*) for a birth coming from the indicated model (#) using different proposal densities. The median was used as robust measure of average, and the acceptance is given in percent. The following proposal densities were tested: a uniform proposal (u), a purely gravity-derived proposal based on geoid heights from the gravity field model EGM08 in the range $n = 161..240$ (g), the linear combination $1/2\,u + 1/2\,g$ (short: $1/2\,g$), the linear combination $2/3\,u + 1/3\,g$ (short: $1/3\,g$), and the linear combination $2/3\,u + 1/3\,d$ with d being an alternative gravity-derived distribution based on vertical deflections ($1/3\,d$). Moreover, an additional run was performed for the proposal $1/3\,g$, where the variance factor was simulated by means of an independent chain (*ind*). Note that only the first and last chain consist of 400,000 samples; with 50,000 samples all other chains are considerably shorter and the related numbers thus less reliable. Therefore, only those models were considered in the table that were visited sufficiently often by all chains.

	likelihood ratio						**acceptance**					
#	**u**	**g**	**1/2 g**	**1/3 g**	**1/3 d**	**ind**	**u**	**g**	**1/2 g**	**1/3 g**	**1/3 d**	**ind**
272	0.014	0.036	0.027	0.022	0.012	0.023	7.9	4.0	4.9	6.8	7.1	8.1
282	0.015	0.053	0.031	0.021	0.013	0.029	8.1	5.0	8.4	7.9	5.7	9.3
292	0.019	0.089	0.033	0.027	0.022	0.035	9.1	2.1	7.6	9.3	9.2	10.0
302	0.021	0.132	0.045	0.039	0.020	0.039	9.7	4.0	9.8	10.8	9.5	11.5
312	0.025	0.123	0.049	0.045	0.033	0.050	12.0	4.9	10.7	14.2	9.5	12.8
322	0.030	0.143	0.052	0.050	0.036	0.056	12.3	5.2	11.0	12.7	11.2	13.4
332	0.036	0.171	0.074	0.070	0.032	0.069	14.0	5.5	14.5	14.1	11.9	14.9
342	0.039	0.184	0.083	0.081	0.052	0.083	14.9	6.3	15.4	15.4	13.0	16.4
352	0.045	0.215	0.116	0.079	0.056	0.090	15.9	6.0	15.9	17.7	15.1	17.7
362	0.050	0.248	0.118	0.071	0.062	0.110	16.8	5.7	15.6	17.0	17.0	18.9
372	0.056	0.327	0.125	0.075	0.067	0.114	17.7	8.5	17.5	19.7	17.1	19.4
382	0.058	0.312	0.120	0.091	0.073	0.129	18.6	7.1	17.8	17.6	17.7	21.5
392	0.061	0.327	0.141	0.114	0.077	0.151	19.2	8.4	19.5	20.6	19.4	23.4
402	0.072	0.322	0.161	0.132	0.097	0.156	19.7	8.4	23.1	21.7	20.8	24.2
412	0.078	0.424	0.202	0.121	0.084	0.180	22.4	6.6	21.2	23.6	16.8	25.9
422	0.084	0.472	0.224	0.161	0.096	0.184	22.5	9.6	22.5	26.7	23.1	26.7
432	0.089	0.480	0.191	0.156	0.120	0.205	23.8	8.1	22.4	24.9	23.1	27.7
442	0.096	0.481	0.193	0.155	0.121	0.206	23.7	9.0	22.3	24.1	21.3	28.0
452	0.110	0.572	0.189	0.158	0.104	0.235	25.2	7.6	24.8	27.5	22.4	29.1
462	0.105	0.471	0.235	0.205	0.130	0.260	25.5	9.5	25.7	29.5	24.6	30.5
472	0.110	0.701	0.211	0.173	0.140	0.286	26.2	11.6	26.3	25.7	25.5	31.8
482	0.116	0.760	0.247	0.208	0.140	0.274	26.4	12.3	23.4	27.1	26.1	31.9

distribution was used, whose parameters were determined as a function of the number of RBFs with the help of the results of the first simulated Markov chain. By applying this alternative approach in the simulation of the variance factor, the acceptance could be improved once again (cf. *ind* in Tab. 5.4).

To sum up, the proposal density inspired by the gravity field represents an improvement over the use of a uniform proposal. In connection with the improved simulation of the variance factor, the likelihood values were up to two times higher, and the acceptance increased by 1-5%.

5.3.4 Modification of the kernel function

An alternative kernel function has been derived to see how a change in the shape of the basis function affects the algorithm. In the regional approach developed by Annette Eicker, the shape coefficients are defined as a function of the degree variances. However, degree variances represent the variance of the signal over the whole sphere; hence, they have global character. Even if the stochastic description of the signal can be adjusted for a desired region by estimating a variance factor from the data restricted to this region, which exactly is what constitutes the benefit of the regional method, the adjustment only refers to a missing scaling coefficient. Moreover, error degree variances of the global GOCE solution were applied to model the uncertainty of the reference field in the low to medium degrees. In doing so, correlations are lost, which for GOCE is particularly unpleasant because of the polar gap problem causing the bump in the representation of the degree variances.

I tried to tackle these issues by simple means: I multiplicated the individual sections of the degree variances curve by a factor that I derived from the comparison between the global variance and the regional variance in the study area. I also eliminated the bump in the range of the low to medium degrees by leaving out the (near) zonal coefficients in the calculation of the degree variances. Furthermore, I no longer used error degree variances but changed to difference degree variances with respect to the GRACE/GOCE combination model GOCO, since the error description of the global GOCE solution following the short arc approach is not reliable in the very low degrees.

An estimated variance component of approximately 1 is often regarded as an indicator of an appropriate stochastic modeling. Unfortunately, this can not be used as a criterion for the regional method applied here, because due to the derivation of the method, the size of the variance component can no longer be interpreted. In order to evaluate the changes made to the kernel function, I have to go one step further claiming that if the stochastic description of the signal was perfect, the variance component would be the same no matter how far the reference field was reduced. In this sense, a kernel function made up from difference degree variances between ITG-Goce02 and GOCO05 multiplicated by the factor 0.98 in the range $n = 2 - 160$, and EGM08 scaled by 1.62 above has proven to be the best of the tested combinations. Note that I did not intend to improve the kernel function in the first place or to provide guidance on how to do it, but I wanted to demonstrate the impact of the improvement on the algorithm also in view of a potential future optimization. In this connection, it is not important that GOCE/EGM information was already incorporated in the construction of the kernel function. Anyway, a correspondingly adjusted kernel function based on Kaula's rule and the formal errors showed similar good results.

With the alternative kernel function, another Markov chain of 400,000 steps was simulated and further 400,000 steps with σ being simulated independently. Since using another kernel function can be expected to be accompanied by another estimation for σ, the prior and proposal were adapted accordingly; the other settings remained unchanged.

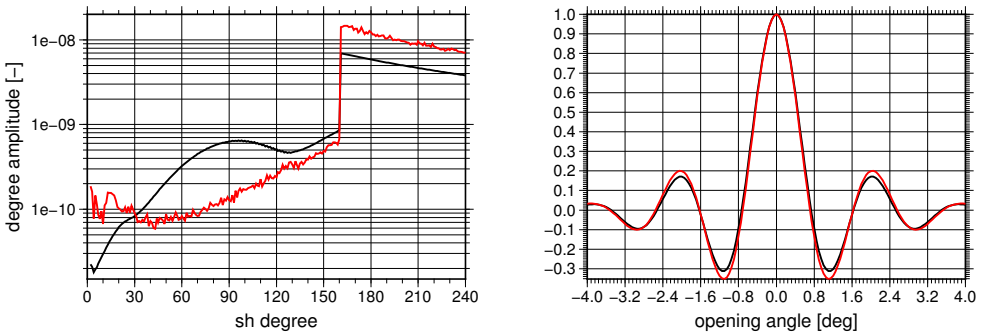


Figure 5.10: Usual (black) and modified (red) kernel function in frequency domain (left) and space domain (right)

5.3.5 Overview of the processing of the four chains

Tab. 5.5 summarizes the differences in the processing of the four simulated Markov chains. Common to all chains is the use of a uniform distribution as prior for the number and positions of the grid points. Furthermore, for all chains the same move type probabilities were chosen for a birth, death or move step. An inclusion of the a posteriori model probabilities would certainly further improve the mixing and could be a subject of future investigations. Also death and move steps were implemented the same for all chains according to the descriptions in Ch. 4. Note that the proposal densities listed in Tab. 5.5 for the birth step both realize a global birth, meaning that they propose points over the whole feasible area. The local birth with the help of the Fisher distribution was not used in the final calculations, as it produced many points outside the feasible area, slowing the chain down. The reason is that the Fisher distribution cannot simply be cut to a specific area, since the normalization constant, which would have to be applied, changes constantly with its position.

Table 5.5: Settings in the processing of four different Markov chains

#	**starting grid**	**birth step**	**simulation of σ**	**RBF-kernel**	**no. of steps**
1	triangular vertex of level 81 cut to the feasible area resulting in 562 points	uniform proposal	proposal: dependent, $q(\sigma, \sigma') = N(\sigma, 0.001)$, prior: half Cauchy with $\gamma = 0.02$	$n = 2..240$, formal errors of ITG-Goce02 up to 160, Kaula's rule above	400,000 + 400,000
2	MAP of the 1st chain with 362 points	gravity-based proposal ($^2/_3\,u + {}^1/_3\,g$)	proposal: indep., $q(\sigma') = p(\sigma\|K, y)$ from the 1st chain, prior: half Cauchy with $\gamma = 0.02$	"	400,000
3	last sample of the 2nd chain with 612 points	"	proposal: dependent, $q(\sigma, \sigma') = N(\sigma, 0.0005)$, prior: half Cauchy with $\gamma = 0.01$	diff. degree variances between ITG-Goce02 and GOCO05 w/o (near) zonal coeff. up to 160, EGM08 $\times 1.62$ above	400,000
4	MAP of the 3rd chain with 212 points	"	proposal: indep., $q(\sigma') = p(\sigma\|K, y)$ from the 3rd chain, prior: half Cauchy with $\gamma = 0.01$	"	400,000

5.3.6 Mixing behavior

In this section, we will look at the mixing of the simulated chains. When a chain is said to mix well, it is moving fast across the parameter space, leading to low correlations between the samples and thus to numerical efficiency. The lower the correlations, the better the samples can reflect the target distribution after a limited number of steps. To assess the mixing, we thus look at the time series plots, autocorrelation functions and histograms for each the number of basis functions, the variance factor and the rms of the differences to EGM08. To save computing time in the evaluation of the individual solutions, the rms was computed for every 10th sample only. This should not make a difference to the appearance of the time series plots and histograms. However, the values of the autocorrelation functions for the rms would still have to be scaled by a factor of 10 to make them comparable.

In general one can say that the correlations for the rms are much smaller than for the number and the variance factor. This can be seen from the shorter oscillation period in the time series plots for the rms. This can also be seen from the histogram plots, where a better picture of the distribution is obtained with the same amount of samples, or from the faster decrease of the autocorrelation functions. This is the opposite of what we saw in the simulations and all the more surprising because the rms is a function of all parameters and as such it is also affected by the mixing of all parameters. In fact, the correlations in the time series of the rms values are even smaller than in the simulations. This is particularly advantageous when calculating the Bayes estimator on solution level, which is based on taking the average of the solutions. In case of small correlations between the solutions, a good estimate might therefore be achieved already with a moderate number of samples.

Moreover, one can see that the histogram plots look worse for the alternative proposal process used for the chains 2 and 4 (as compared with the results for the chains 1 and 3, see Fig. 5.12). This points to higher correlations and a poorer mixing. This was not to be expected because the alternative proposal process was specifically designed to improve the mixing. As we saw in Sec. 5.3.3, a higher acceptance was achieved by the use of the gravity derived proposal for the point positions and the independent proposal for sigma. Although the acceptance was still moderate for the models listed in Tab. 5.4, we could already see that the acceptance increased with the number of basis functions used. So it is apparently less important where exactly new basis functions are introduced when there is a sufficient number of functions to capture the major part of the signal. In the range of the larger models, the acceptance was up to 35 and 45% for the chains 1 and 2, respectively. For the chains 5 and 6, which both make use of the EGM kernel function and otherwise differ only in the proposal distribution for sigma, the acceptance was up to 50 or even 60%. The higher acceptance is indeed an indication that the proposal for the point positions is better in the sense that it is closer to the target distribution. However, the model parameter is simulated in a random walk procedure, and here a too high acceptance rate is rather detrimental to the mixing. So the reason for the poorer mixing is the higher acceptance especially in the range of the larger models. When we leave out these models, limiting ourselves to the models in the HPD region, the histogram plots improve as shown by the example of chain 4 (Fig. 5.12). Looking at the values of the autocorrelation functions for the models in the HPD region, one can see that the correlations for the alternative proposal process even seem to be somewhat smaller, which would point to a slightly better mixing. However, to truly evaluate the gain of the alternative proposal process, one would have to choose a larger step width in the simulation of the model parameter in accordance with the increased acceptance rate.

While it is difficult to see from the present results whether the mixing has really improved by the use of the alternative proposal process, it is very obvious that limiting the range of models to the HPD region does reduce the correlations considerably. It would therefore probably make sense to limit the range of the feasible models more strictly in future calculations.

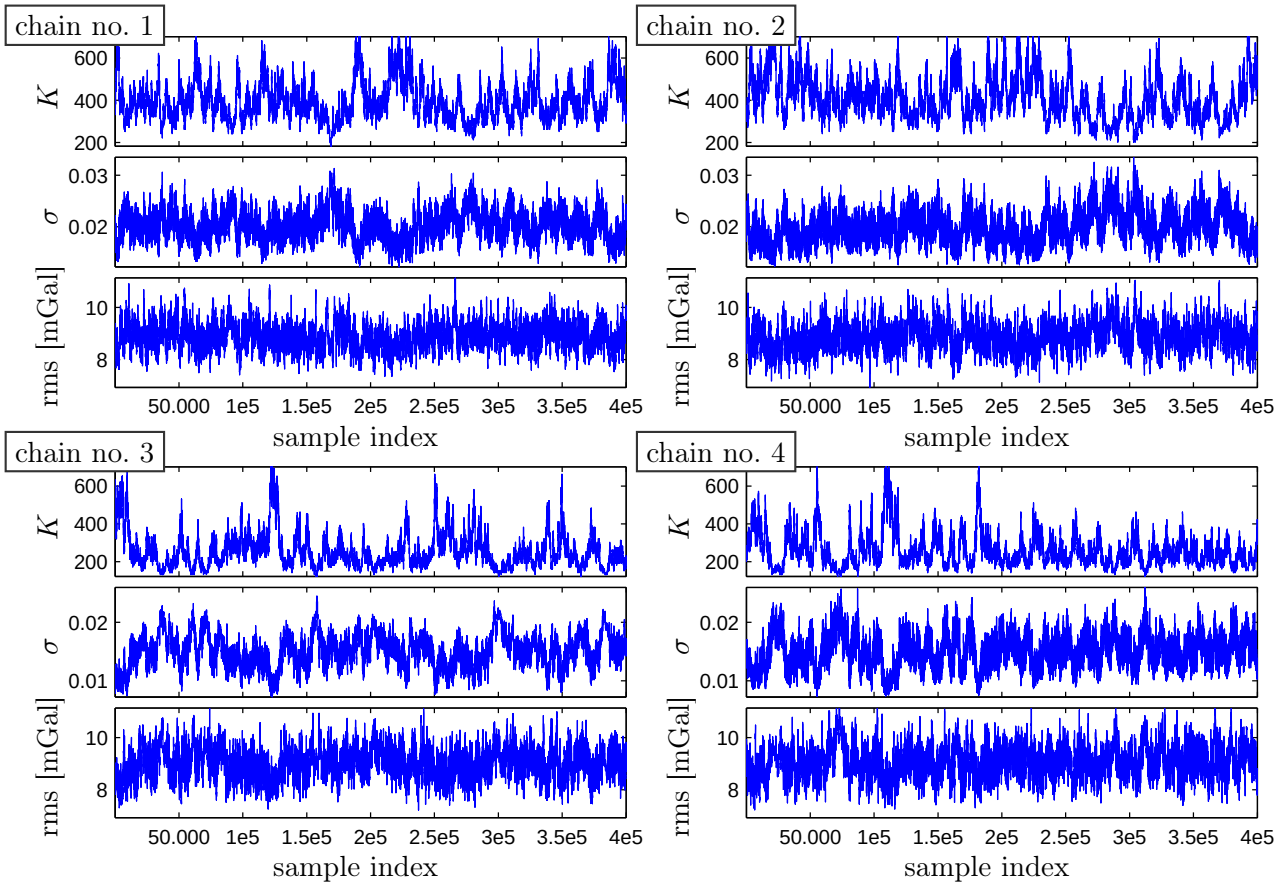


Figure 5.11: Time series plots for the number of the basis functions K, the variance factor σ and the rms of the differences of the individual solutions to EGM08 in terms of gravity anomalies for the four simulated chains.

Table 5.6: Number of steps until the correlation decreased to the indicated value (*autocorr.*) for the number of the basis functions K, the variance factor σ and the rms of the differences to EGM08 for the four simulated chains. The numbers for the rms were calculated from a time series made up of every 10th sample and therefore still need to be multiplied by 10.

	chain no. 1			**chain no. 2**			**chain no. 3**			**chain no. 4**		
autocorr.	**K**	**σ**	**rms**	**K**	**σ**	**rms**	**K**	**σ**	**rms**	**K**	**σ**	**rms**
0.3	3440	2888	59	2695	2624	72	4667	4017	83	3280	2576	95
0.2	4829	3734	92	6729	5902	109	5895	5373	163	3883	3255	182
0.1	8605	7650	151	7941	7613	340	8604	7500	380	5296	3983	325
0	17264	17394	764	31571	32062	732	15003	15539	1564	10977	5707	408

Table 5.7: The same as Tab. 5.6 but restricted to the models in the HPD region.

	chain no. 1			**chain no. 2**			**chain no. 3**			**chain no. 4**		
autocorr.	**K**	**σ**	**rms**	**K**	**σ**	**rms**	**K**	**σ**	**rms**	**K**	**σ**	**rms**
0.3	1431	1079	42	1349	1090	40	2204	2099	43	1419	1020	39
0.2	1905	1518	63	1773	1610	57	2940	2713	68	1886	1536	61
0.1	2788	2431	113	2484	2360	82	3933	3391	104	2494	2224	130
0	4076	4016	173	6051	5826	204	7183	6306	362	3057	2959	233

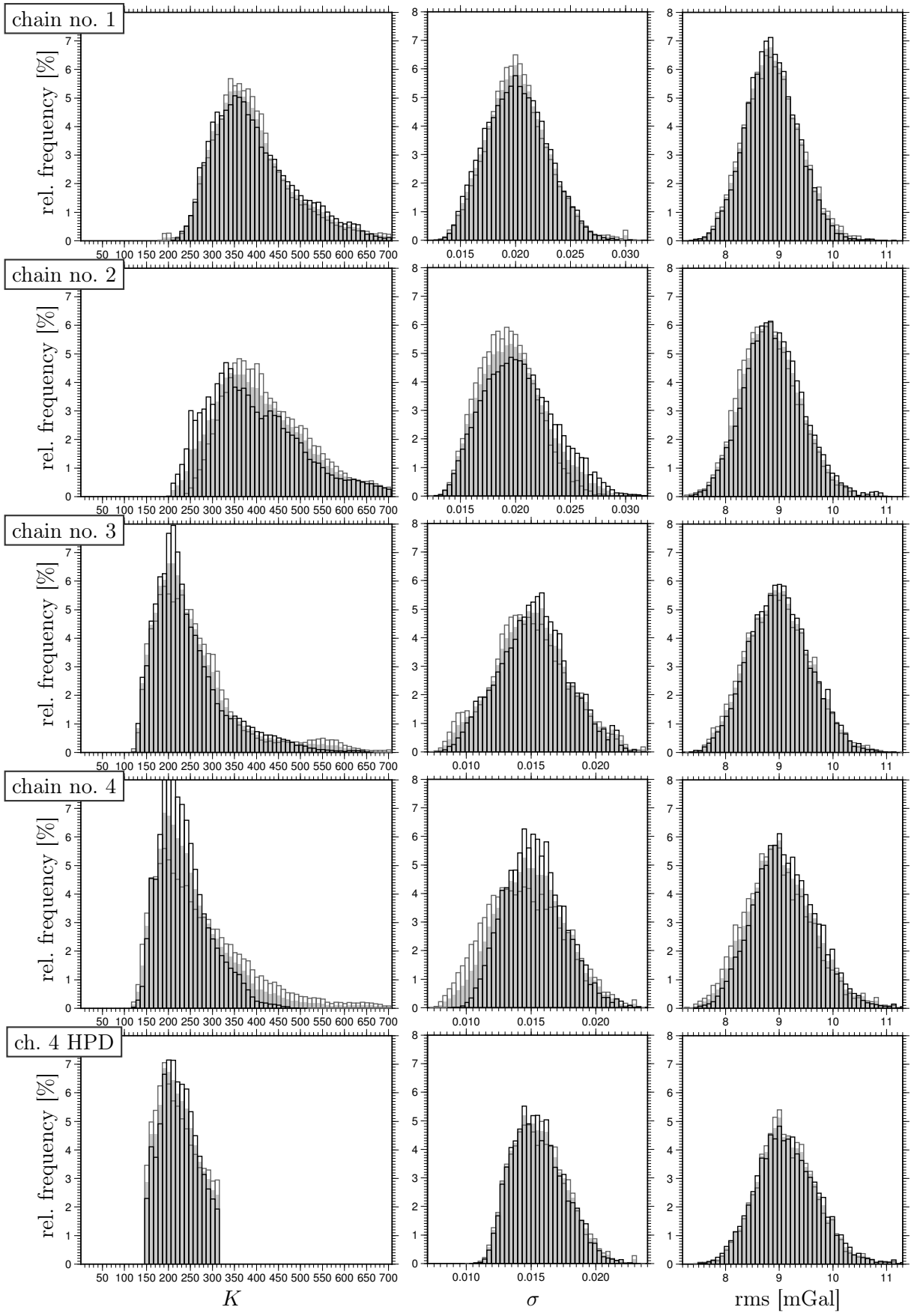


Figure 5.12: Histogram plots for the number of the basis functions K, the variance factor σ and the rms of the differences to EGM08 for the four simulated chains and for chain no. 4 restricted to the models in the HPD region. There is a histogram for the entire chain (filled) and one for both the first half (gray outline) and the second half (black outline).

5.3.7 Analysing the output of the chains

In this section, the results of the chains are presented. The focus is on the differences in the target distributions, as opposed to the previous section, which was only about the speed of convergence.

Fig. 5.13(a) shows the histogram of the model indicator (i.e. the number of basis functions) based on the output of the chains 1 and 2, which only differ in the proposal process applied but not in the target distribution itself. The point of the highest density, the MAP, is at 342 basis functions (cf. 708 for the standard regular grid and 562 for the initial grid with the smaller margin). The range of models with a Bayes factor $p(342)/p(K)$ of not greater than 3, which in the sense of the interpretation of the Bayes factors are just as likely as the MAP model, is $K = 272..512$. This corresponds to a 83% HPD interval. The HPD interval comprises 25 models or 250 basis functions; hence, there is a lot of uncertainty about the optimal number of basis functions. One reason for this is that the number of basis functions is strongly correlated with the variance factor, and as always with correlations, the uncertainty appears to be larger in the marginal distribution than it actually is. But also for a particular σ, the distribution is still wide. So the observations are obviously not very informative about the model. The addition of further basis functions seems to improve the sum of squared residuals over a wide range to such an extent that it outweights the degradation of the density due to the higher number.

Fig. 5.13(b) shows the histograms for the chains 3 and 4, both making use of the alternative kernel function from chapter 5.3.4. Here the MAP is at 202 basis functions; the models with a Bayes factor of up to 3 are 152..312, i.e. 17 models, which corresponds to a 79% HPD interval. Thus the type of kernel function used has a strong effect on the distribution of the number of the basis functions. In particular, a better kernel function in the sense of a better stochastic description of the signal, which I think I have achieved by the adaption to the local signal content, leads to a shift towards simpler models and to a narrower density function. It is intuitively clear that fewer basis functions are needed when using the adequate type of function. And since no artificial signal is introduced, further basis functions are very unlikely, leading to the steep descent of the density function.

In addition to the histogram plots, the Figs. 5.14(a) and 5.14(b) show the rms values of the samples depending on the number of the basis functions. All medium and large models can reach the same good data fit, as one can see from the good rms values reported for these models. From a certain point, however, the basis functions are no longer sufficient for a good approximation. The MAP lies at the point of intersection of the imaginary lines through the low rms values on the one hand and the increasing rms values on the other hand. It may therefore be interpreted as the lowest number of basis functions that still yields a good data fit. This is in accordance with the findings from chapter Sec. 4.7.1, in which the evidence was used to explain which models receive a high density value within the procedure.

Fig. 5.15 shows the relation between the number and the variance factor with different colors indicating different kernel functions. Obviously, the variance factor has a different size depending on the kernel function used. The variance factor has to be understood as a scaling factor to the kernel function. Thus the reason for the different size is simply that the two kernel functions differ on average, and another scaling factor is therefore necessary for the optimal adaption to the local signal content. Independent of the effect of the kernel function, one sees that the number and the variance factor are strongly correlated. Small models are associated with a large variance factor, that is with a low weight on the regularization term. This is clear because the prior information, which says that the coefficients are zero, is more accurate for many basis functions than for few. Here the coefficients will have to be larger to realize the same signal, resulting in a small value of the regularization parameter. Moreover, one can see that the smaller the number of the basis functions, the wider the distribution of the variance factor. This is probably due to the stabilization of the system caused by the reduction of the parameters (cf. Sec. 5.3.9). In a stable system, the

regularization has less influence on the solution, so that the regularization parameter can vary more widely within the limits of accuracy. The 3-dimensional representation of the distribution, Fig. 5.16, shows once again that if the variance factor was set to a specific value, the choice of the value would greatly influence the distribution of the number.

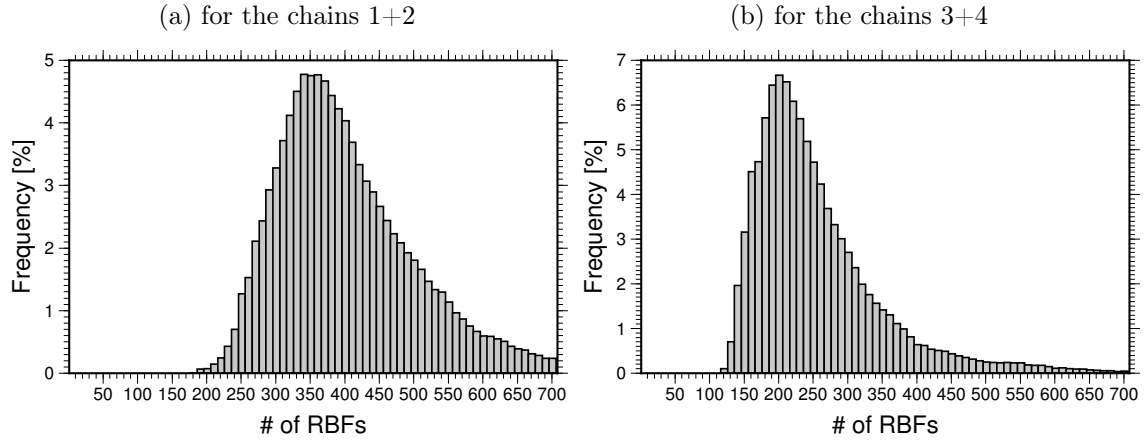


Figure 5.13: Posterior distribution for the number of basis functions

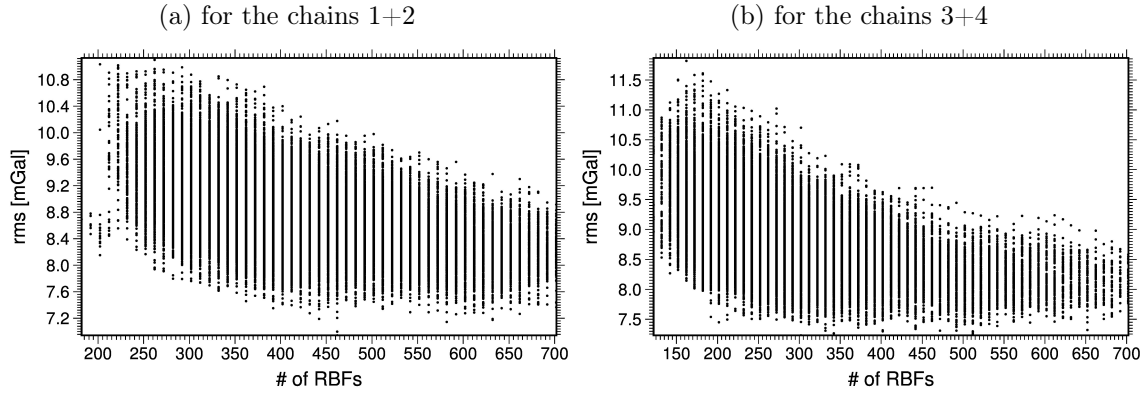


Figure 5.14: Scatter plot showing the rms of the samples as a function of the number of RBFs

5.3.8 Resulting gravity field models

5.3.8.1 The MAP estimator

The MAP estimator was determined for all models in the HPD region of the respective Markov chain. The resulting gravity field models were compared with EGM2008; the rms of the differences in terms of gravity anomalies fluctuates between 7.62 and 10.81 mGal, whereby the majority of the values lie in the range of 8 to 10 mGal. This is much larger than what had been achieved by the standard approach, which was 7.59 mGal. Earlier on, in the simulations, we had already seen that the MAP estimator yielded slightly worse results compared with the other estimates. This was explained by the fact that the sample with the highest density is not a very precise guess for the point of the highest density since the number of samples falling into the area of the MAP is small with respect to the total number. In a more realistic framework, the sampling is still far worse, which is because the size of the parameter space increases exponentially with the dimension. This phenomenon is known as the curse of dimensionality. In a nutshell, it is not useful to derive the MAP from the output of a Markov chain. It would be better to use a real approach for optimization

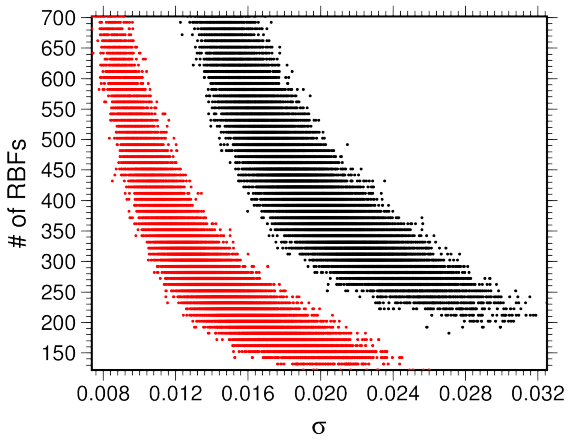


Figure 5.15: Scatter plot showing the relation between the variance factor and the number of RBFs for the chains 1+2 (black) and 3+4 (red)

instead like simulated annealing, which by the way could easily be implemented in the frame of the present approach by just a small modification in the expression of the target density.

5.3.8.2 The Bayes estimator on the level of the parameters

The calculation of the Bayes estimator on the level of the parameters involves the sorting of the random grids (cf. Sec. 4.7.3). This sorting is actually implemented by testing any possible permutation in a brute force manner, which is very time-consuming and could therefore not be applied to real data. The problem of finding the permutation that minimizes the distances to the MAP is very similar to the traveling salesman problem, for the solution of which efficient algorithms exist. This could be part of future work.

5.3.8.3 The Bayes estimator on solution level

The Bayes estimator on the solution level was calculated for the four simulated Markov chains; the results are shown in Tab. 5.8. Only every 10th sample was considered in the calculation of the mean in order to keep the computational effort in the evaluation of the individual solutions within acceptable limits. This does not affect the accuracy of the solution because the thinning of the chain does not only reduce the number of samples but also the correlations. Or in other words, in the presence of strong correlations, a single sample hardly contains any new information, so leaving it out should not make a difference to the solution. To demonstrate this, another solution based on every 20th sample was calculated exemplarily for the second chain, and the result was almost exactly equal (Tab. 5.8). In principle, the chains number 1 and 2 differ only in the proposal process applied to explore the target distribution, but the actual distribution is the same. This is also true for the chains 3 and 4, for the simulation of which a modified kernel function was employed. Markov chains for the same target distribution should obviously also come to the same conclusions. Deviations arise from errors in the numerical approximation of the moments of the distribution. For the chains number 1 and 2, the deviation in the rms values of the differences to EGM2008 in terms of gravity anomalies is 0.031 mGal. For the chains 3 and 4, the deviation is 0.076 mGal. This is a lot, indicating that the chains should probably have been simulated longer. It does not, however, affect the significance of the conclusions about the comparison with the standard approach later on, especially since the errors will still become less when taking the mean for the final results. The results of the chains 3 and 4 match worse. The reason could be that the correlations between the solutions are higher for the modified kernel function, as we saw in Sec. 5.3.6. Moreover, as was

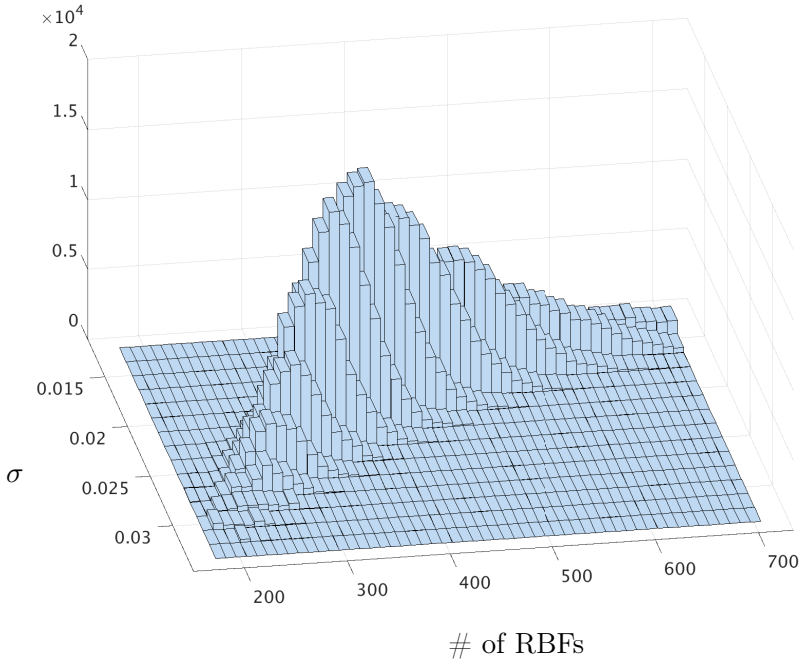


Figure 5.16: 3d histogram for the relation between the variance factor and the number of RBFs at the example of the chains 1+2

also done in the simulation scenario, the calculation of the mean was limited to the HPD region and the MAP model. Unfortunately, the effect of this manipulation of the distribution on the expectation value cannot be properly assessed because of the size of the numerical errors (see Tab. 5.8 exemplarily for the second chain).

In the following, the final results, which were generated by taking the mean of the individual results of the chains 1 and 2, and 3 and 4, are compared with the global spherical harmonic solution ITG-Goce02 and the regional solution based on the standard approach. It should be emphasized that the regional solution described in Ch. 5.2 having been calculated for the complete patch consisting of northern and southern part is utilized for the comparison, cut to the particular comparison area needed. The reason is that for the current smaller study area, which corresponds to the former southern part, the regional analysis did not yield satisfactory results because of difficulties in the variance component estimation. Additionally, a further regional solution based on the standard regular grid but using the modified kernel function is taken into account.

As already discussed at the beginning of this chapter, even in the study area considered here, the signal is not truly homogeneous. A uniform model resolution and regularization according to the mean signal content will therefore probably lead to noise remaining in smooth areas and a loss of signal in rough areas. To test for a possible improvement by the adaption of the model resolution, I divided the area into two comparison areas of rather smooth or rough signal. The division was done visually on the basis of a map of vertical deflections reflecting the slope of the geoid (cf. Fig. 5.17). As the smooth portion of the study area is small, I evaluate the models additionally in the area of the observations. The choice of another smaller comparison area should serve to reduce edge effects, which I think are small in satellite data analysis anyway. Tab. 5.9 shows the statistics for the signal of the models in the different areas, Tab. 5.10 for the differences to EGM2008, both in terms of gravity anomalies. Numbers in terms of geoid heights are not provided because of their similar meaning. In the smooth areas, the regional solution resulting from the optimization of the point grid is smoother than ITG-Goce02 and the regional solution from the standard approach. For

example, the rms of the signal in the area of the observations decreases from 18.35 mGal for ITG-Goce02 to 18.23 mGal for the regional solution with the standard grid and further to 17.85 mGal with the optimization of the grid. This is an improvement as one can see from the comparison with EGM2008, in which the rms of the differences decreases from 6.84 to 6.45 and further to 5.97 mGal. Note that the numbers for the regional method following the standard approach are too optimistic since in the considered region the solution benefits from the stronger regularization in the northern regularization area. Actually, a computation limited to the study area yielded worse results than the global approach. In the rough area, in contrast, the energy increases. For instance, the mean of the signal in the study area increases from 6.27 mGal for ITG-Goce02 or the regional solution with the standard grid to 6.41 mGal for the mean of the chains 3 and 4. At the same time the mean of the differences to EGM2008 is reduced by a factor of 3. Moreover, the min/max values raise from $-204.99/158.44$ for ITG-Goce02 to $-216.54/173.29$ for the mean of the chains 3 and 4, which is up to 15 mGal more signal at certain points. Fig. 5.18 shows the differences to EGM2008 in the space domain. For the regional solutions only the variant with the modified kernel function is included. Comparing the regional solution from the standard approach, Fig. 5.18(b), to the global spherical harmonic solution, Fig. 5.18(a), one sees that in the rough area the systematic differences along the trench decrease, whereas in the smooth south the differences grow. In the smooth north the solution probably still benefits from the stronger regularization in the northern regularization area, as pointed out before. In the regional solution resulting from the optimization of the point grid, Fig. 5.18(c), there is a clear decrease of the systematic signal-correlated structures but also a visible reduction of noise in the south or next to the trench in the East.

Tab. 5.11 shows the final results for the entire study area. The rms of the regional solutions improve from 7.59 mGal with the standard grid to 7.37 mGal with the optimization of the grid, both for the usual kernel function. As one can see, the modified kernel function yields better results, that is 7.42 mGal with the standard grid and 7.14 mGal with the optimization of the grid. In total, the optimization approach thus leads to an improvement of 13% over the global solution, which is almost twice the improvement of 7% reported for the regional method following the standard approach. In terms of geoid heights, the improvement is slightly less. This is not surprising given the fact that the benefit is mainly in the higher frequencies, and these are given greater weight in the evaluation of the anomalies.

Table 5.8: Rms of the differences between the Bayes estimator on solution level in different variants and EGM08 for the four simulated chains. The numbers in parentheses indicate differences between the chains 1 and 2 and between the chains 3 and 4.

	Δg [mGal]	**N [m]**
chain no. 1	7.353	0.2213
chain no. 2	7.384 (0.031)	0.2223 (0.0010)
every 20th	7.384 (0.000)	0.2223 (0.0000)
HPD	7.390 (0.006)	0.2225 (0.0002)
MAP	7.424 (0.040)	0.2236 (0.0013)
chain no. 3	7.103	0.2144
chain no. 4	7.179 (0.076)	0.2163 (0.0019)

5.3.9 Stability issues

Gravity field determination is an ill-posed problem. As explained in Sec. 2.5.3, the reasons are the downward continuation process and, specifically for GOCE, the polar gap. But the stability is also affected by the representation, which for RBFs depends on the choice of the maximum resolution,

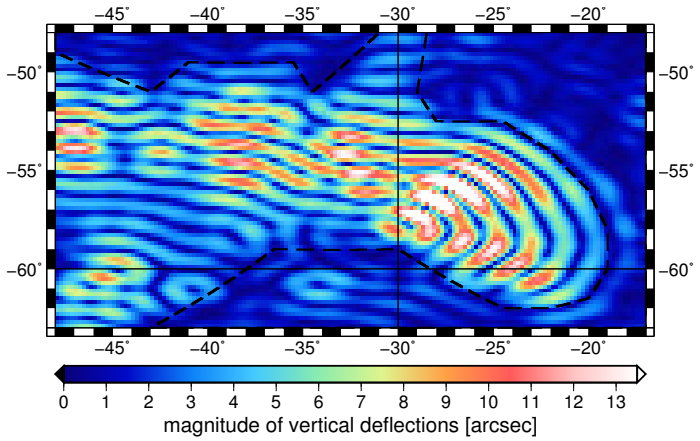


Figure 5.17: Division of the test site into smooth and rough areas with the help of the map of vertical deflections shown in the background

Table 5.9: Statistics (min/max/mean/rms) for the gravity field signal of the calculated GOCE models evaluated in different areas in terms of gravity anomalies [mGal]

	smooth part		**rough part**	
	study area (small)	**data area (med)**	**study area (small)**	**data area (med)**
ITG-Goce02	-16.76/56.83/14.17/20.38	-38.66/56.83/7.57/18.35	-204.99/158.44/6.27/52.10	-204.99/158.44/6.74/47.81
regional	-14.18/56.51/13.96/19.81	-33.47/56.51/7.61/18.23	-205.48/159.96/6.27/52.34	-205.48/159.96/6.71/47.79
+mod. kernel	-14.55/58.70/13.90/19.95	-34.84/58.70/7.57/18.32	-208.22/162.14/6.29/52.56	-208.22/162.14/6.76/48.04
chain no. 1+2	-16.85/59.48/13.70/19.69	-34.47/59.48/7.49/17.85	-213.52/171.46/6.40/52.72	-213.52/171.46/6.80/48.16
chain no. 3+4	-16.99/59.74/13.58/19.61	-34.66/59.74/7.45/17.93	-216.54/173.29/6.41/53.03	-216.54/173.29/6.86/48.49

Table 5.10: Statistics (mean/rms) for the differences of the computed gravity field models to EGM2008 in terms of gravity anomalies [mGal]

	total		**smooth part**		**rough part**	
	small	**med**	**small**	**med**	**small**	**med**
ITG-Goce02	0.07/8.17	0.00/7.52	1.89/6.99	0.23/6.84	-0.19 /8.32	-0.15/7.91
regional	0.04/7.59	0.00/7.27	1.69/6.66	0.28/6.45	-0.19 /7.71	-0.17/7.73
+mod. kernel	0.05/7.42	0.01/7.25	1.63/6.44	0.24/6.51	-0.17 /7.55	-0.12/7.66
chain no. 1+2	0.13/7.37	0.01/7.04	1.42/5.61	0.16/5.97	-0.06 /7.58	-0.08/7.61
chain no. 3+4	0.11/7.14	0.02/6.95	1.31/5.56	0.11/5.95	-0.05 /7.34	-0.03/7.49

Table 5.11: Rms (and improvement with respect to ITG-Goce02) of the differences between the calculated GOCE models and EGM2008

	Δg [mGal]	**N [m]**
ITG-Goce02	8.17	0.244
regional	7.59 (7%)	0.226 (7%)
+mod. kernel	7.42 (9%)	0.221 (9%)
chain no. 1+2	7.37 (10%)	0.222 (9%)
chain no. 3+4	7.14 (13%)	0.215 (12%)

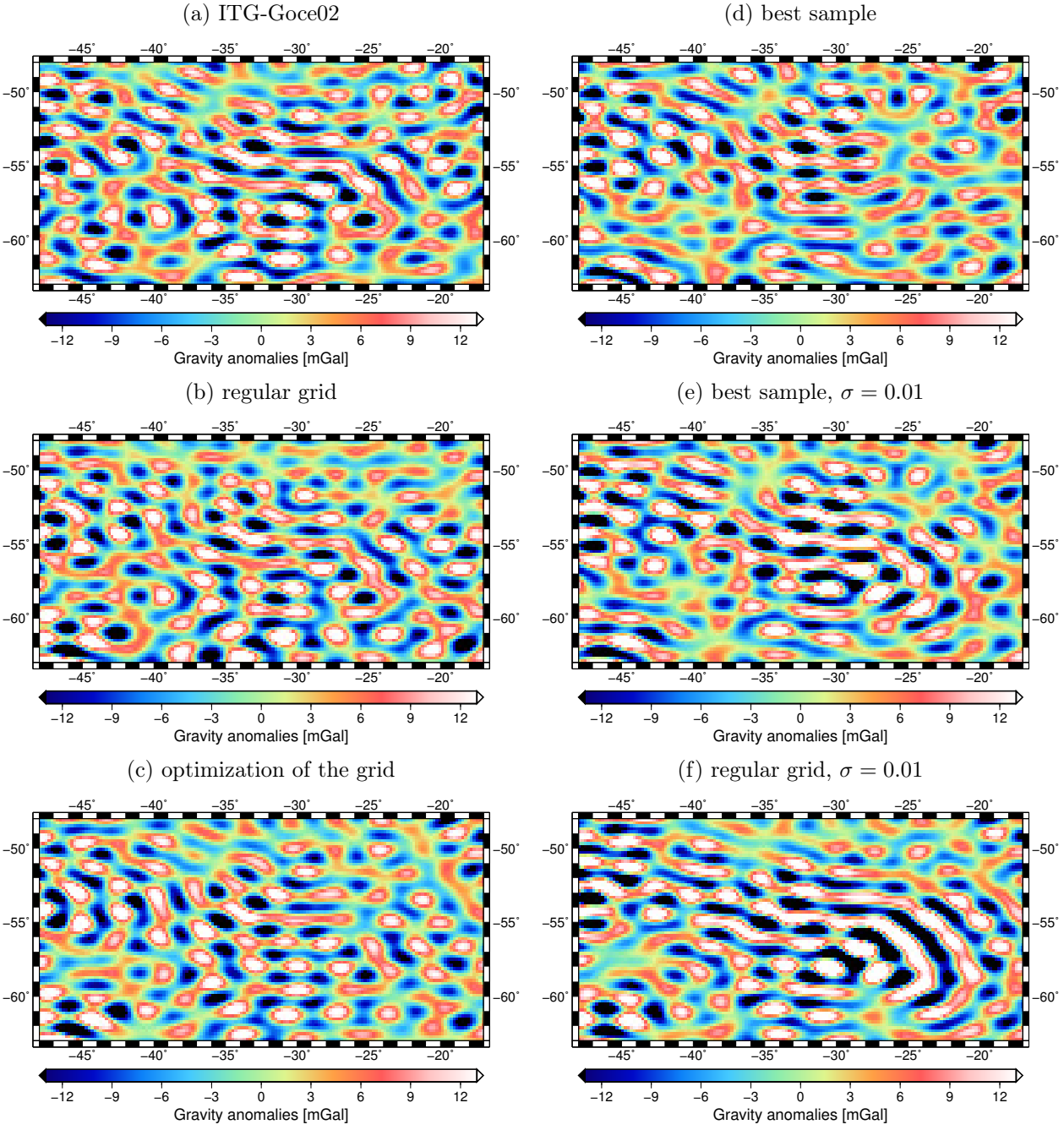


Figure 5.18: Differences between the calculated gravity field models and EGM08. The following models were considered: (a) the global spherical harmonic model ITG-Goce02, (b) the regional solution using the standard regular grid, (c) the regional solution resulting from the optimization of the grid, (d) the sample with the best agreement with EGM08, (e) the same as (d) but using a smaller variance factor, and (f) the same as (b) but using a smaller variance factor.

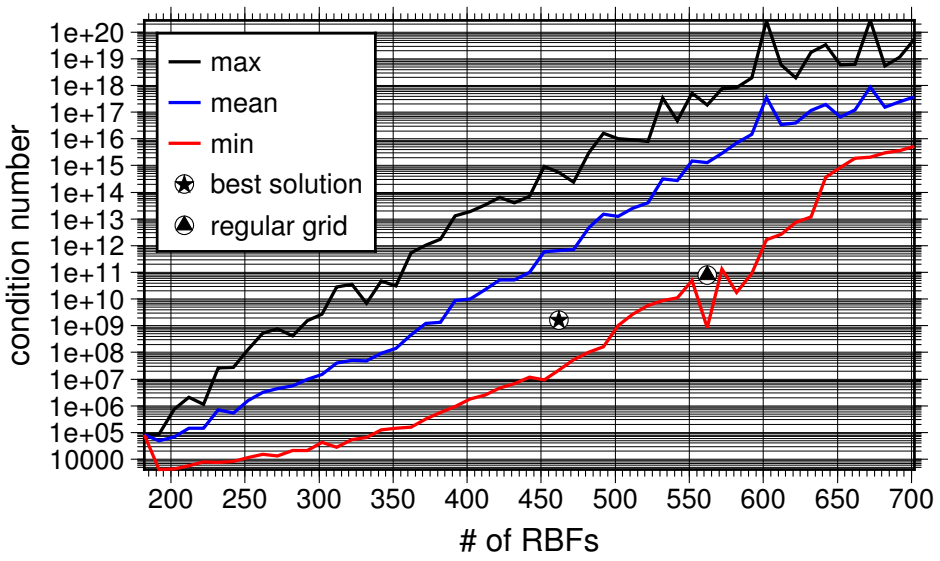


Figure 5.19: Stability over the run of the 1st Markov chain

the shape of the basis functions and their nodal point distribution. Kusche examined these factors under ideal conditions using the condition number as a measure of ill-posedness (cf. Kusche, 2002). He found that the geometry of the point grid is important for the condition of the normal equation matrix. According to him, the ideal point grid in the sense of the best stability is the uniform grid derived from the Platonic solids. However, it was also noted that reducing the number of basis functions has a strongly regularizing effect. So one might expect that the use of fewer well-distributed basis functions also leads to a stabilization of the problem. This shall be verified in the following.

To this end, condition numbers for the non-regularized normal equation matrices were calculated for the samples of the chains 1 (from the start model) and 2. In accordance with what was stated above, the condition depends on the number of RBFs and decreases when the number is reduced (Fig. 5.19). For the same number, the condition may vary depending on the actual arrangement of the basis functions by several orders of magnitude. To check for the impact of regularization, I decided to use the best sample in the sense of having the best agreement with EGM2008. I think this is in line with what is usually done in the standard approach, where the resolution of the point grid is chosen so that the solution fits best to a reference model. The chosen sample with 462 RBFs has a condition number that is smaller by a factor of about 50 than the condition of the initial solution based on the standard regular grid with 562 RBFs, though the initial solution is almost the best among the samples with the corresponding number.

The best solution is thus more stable than the standard solution. To show how this changes the degree the solutions being affected by regularization, I modified the regularization parameter by the same amount for both solutions and compared the results. A variance factor that is larger by a factor of about 2 yields rms values with respect to the data area of 8.51 and 9.99 mGal for the best solution and the standard solution, thus a smaller change for the best solution (Tab. 5.12). Accordingly, a variance factor that is smaller by a factor of 2, which corresponds to a stronger regularization, yields rms values of 7.86 and 9.01 mGal. As one can see in Figs. 5.18(e) and 5.18(f), the strong systematic, signal-correlated errors caused by giving too much weight to the erroneous prior information are smaller for the best solution with the optimized grid.

Table 5.12: Rms for different values of the variance factor

	$\sigma \approx 0.02$		$\sigma = 0.05$		$\sigma = 0.01$	
	small	**med**	**small**	**med**	**small**	**med**
best solution	7.01	6.94	8.27	8.51	8.78	7.86
standard solution	7.58	7.58	9.00	9.99	10.48	9.01

5.4 Discussion

The applied approach was developed for the processing of satellite observations divided into short arcs. It is therefore equally well suited for global and regional gravity field analysis. It should be emphasized that the same data and data preprocessing, and the same standards and background models were used for all generated gravity field models. All calculations were made with the same software package, and where possible also the same program settings were used. I therefore believe that the comparison is fair and that different results point to differences in the underlying processing strategies.

In Sec. 5.1 we saw that the global GOCE solution is of comparable quality as the official ESA models of the second generation. In Sec. 5.2 we saw that the regional method with uniform point distribution applied with exactly the same data can still improve this result by for example 7% in the southern part of the South Sandwich deep sea trench test area. At the beginning of this work, there was the hope that the adaption of the model resolution helps to prevent overparameterization and thereby leads to a stabilization of the problem, so that the solution is less affected by the imperfections of the prior information. Indeed, the number of the basis functions decreased drastically from 708 for the regular standard grid (or 562 for the grid with the smaller margin) to 202 for the variant with the alternative kernel function. In Sec. 5.3.9 I showed that at the same time the normal equation system becomes more stable and the prior information has less influence. Moreover, we saw that the smaller the number, the larger the variance factor, that is the lower the weight on the regularization term, which might even further reduce the influence (see Fig. 5.15). With regard to the resulting gravity field solutions, one can say that the adaption of the point grid leads to more energy in rough areas and a smoother solution in smooth areas and altogether to smaller errors than ITG-Goce02 or the regional solution with the standard grid (Sec. 5.3.8). Thus deficiencies of the solution, which I think are the result of an inappropriate prior information, become less. I want to add that, no matter which settings I tried in the regional analysis with standard grid, I could not achieve an equally good solution in both areas. All in all one can say that the procedure for the optimization of the point grid works and yields up to 6% better results than the standard approach; the improvement over the global solution is thus twice as large.

6. Conclusions and outlook

The aim of this thesis was to estimate an optimal point grid for the arrangement of the basis functions in a radial basis function approach to regional gravity field analysis in addition to the usual model parameters, which are the scaling coefficients and a variance factor. To achieve this, the RJMCMC algorithm of Green (1995) was implemented, which allows sampling from a posterior distribution that contains the number of parameters as one of its parameters. In this way, the number of the basis functions can be estimated from the data like an ordinary model parameter. This is a great advantage over existing approaches since I do not have to specify any kind of stop criterion to prevent the algorithm from introducing more and more basis functions in order to further reduce the residual sum of squares. In fact, I do not even have to specify a preference for specific models in the prior, but the tendency for simple, uncomplicated models is inherent in the Bayesian approach to model comparison.

The key points in the implementation of the approach are as follows: (1) I make use of the least squares estimate for the linear problem of determining the scaling coefficients for a given point grid. Thereby the sampling dimension is reduced, and I am able to integrate available software for the regional analysis into the procedure. Furthermore, it improves the acceptance of steps that change dimension and is therefore important for a fast mixing. (2) To further improve the mixing, I invented a proposal density for the birth step that is derived from a gravity field model. Using it, I could slightly increase the acceptance. However, the mixing of the chain is still a problematic issue. (3) Since I am mainly interested in the resulting gravity field model, I do not calculate the estimates on parameter level, but I formulate the Bayes estimator on the basis of the gravity field solutions constructed from the sampled parameters. In this way, the labeling problem is solved without the need for sorting the parameters.

Applying the method in regional gravity field analysis from GOCE data resulted in a significant reduction of the required number of basis functions compared to the use of a standard network, and the spatial distribution of the basis functions resembled the structures of the gravity field signal. The solutions showed less noise and more signal respectively in smooth and rough areas and improvements of up to 13% in comparison to competitive global and regional models based on the same processing strategy. As hypothesized in the introduction, I attribute this to the stabilization of the normal equation system resulting from the reduction of the number, making the solution more resistant to simplified assumptions in the prior information, which was confirmed by the results. Generally, one can conclude that it makes sense to concentrate on improving the model in regional analysis, and that information about the optimal model can be revealed from the data.

These findings (see also the results chapter) suggest the following extensions/improvements of the methods, and alternative ideas and applications for future research:

- With a few simple adaptions of the degree variances curve determining the shape of the basis functions, the acceptance was higher, the number of basis functions smaller, the uncertainty about the number lower, and the solution was better. Accordingly, I expect that really adapting the shape to the data is likely to bring further improvements. For this, one would have to introduce an additional parameter per basis function, controlling the shape of the function, and design a new move type to simulate it. For example, this could be the depth of the function under the reference sphere. Alternatively, one could also choose from among a discrete number of different types of basis functions.
- The adjustment of the variance factor in regional analysis allows to adapt the prior information to some extent to the signal in the local study area. In the present approach, the prior information about the smoothness of the field is already incorporated in the construction of

the basis function by choosing the shape coefficients as the degree variances according to Kaula's rule. Estimating the shape of the basis functions from the data might therefore be interpreted as a further adaption of the prior information to the local conditions. As an alternative to that, there are other ways to improve the prior information. A. Eicker already proposed earlier to split up the study area and estimate several regularization parameters for the individual sub-regions (Eicker, 2008). This could be easily combined with the RJMCMC approach developed in this thesis. Future work could also aim at a completely different construction of the regularization matrix using local constraints on the scaling coefficients of neighboring functions like Rowlands et al. (2010) or Watkins et al. (2015).

- In this work, the approach for optimizing the point grid was applied to data of the satellite mission GOCE. Although in principle the approach can be applied to any data, the benefit might be greatest for heterogeneous data. In this context, it would be interesting to use it for the analysis of GRACE(-like) observations. Because of the mission design, the highly accurate inter-satellite range measurements are only available along the polar satellite orbit; across this direction, the sensitivity is lower. The uniform model resolution connected with the use of spherical harmonics leads to correlations between sectorial coefficients, which manifest themselves in the well-known striping pattern of GRACE solutions. The optimization of the point grid in the framework of regional gravity field analysis would allow adapting the model resolution to the data, thereby counteracting these effects already in the data processing step. Compared to the use for GOCE data, using the approach for the calculation of monthly GRACE solutions would have the advantage that because of the lower number of observations and the lower resolution achievable, the individual steps would be faster, and we probably require less. It is conceivable to estimate the optimal grid on the basis of one month of data and then to re-use it for the processing of the other data, assuming that the spatial structures remain the same over time.

- However, to re-use the point grid for further tasks, a parametric solution would have to be available, which currently is not. The RJMCMC algorithm implemented here arranges the basis function-specific parameters in an unsorted vector. Determining the Bayes estimator as the mean of the parameter vectors therefore requires the sorting of the parameters. In my view, it would be sensible to sort the points of a random grid in such a way that the distance to a specific point grid (e.g. the MAP grid) becomes minimal. When the number of basis functions is high, this is not an easy task. But the problem is similar to the traveling salesman problem for which solution algorithms exist.

- The applicability of the approach is limited by the rather large numerical errors in the solutions. These errors are caused by the fact that the number of generated samples is not sufficiently large, as in every iteration of the algorithm, an equation system has to be set up and solved, which is time-consuming. The current version of the algorithm already considers that always only a few points change within a step by updating only the part of the normal equations that has changed compared to the previous step. However, in a birth or move step always the whole observation equations are set up again. Since the accumulation of the basis functions takes a significant amount of time, also the observation equations should be stored in future calculations, and/or function tables should be used to interpolate the values of the basis functions depending on the height and the opening angle.

- Moreover, correlations between the samples reduce the effective sample size. We saw that in the range of large/small models the acceptance probability was rather high/low, leading to a bad mixing and high correlations. It is thus advisable to further restrict the prior on the number in future calculations, for example to the models in the HPD region. This would only slightly change the target density and thus the derived estimates, but it would considerably reduce the correlations, as was very clear from the test computations.

Bibliography

Andrieu, C., de Freitas, N., Doucet, A., and Jordan, M. I. (2003). An Introduction to MCMC for Machine Learning. *Machine Learing*, 50:5–43.

Andrieu, C. and Doucet, A. (1999). Joint Bayesian Model Selection and Estimation of Noisy Sinusoids via Reversible Jump MCMC. *IEEE Transactions on Signal Processing*, 47(10):2667–2676.

Antoni, M. (2012). *Nichtlineare Optimierung regionaler Gravitationsfeldmodelle aus SST-Daten*. PhD thesis, Universität Stuttgart.

Antunes, C., Pail, R., and Catalão, J. (2003). Point mass method applied to the regional gravimetric determination of the geoid. *Studia Geophysica et Geodaetica*, 47(3):495–509.

Balmino, G. (1972). Representation of the Earth Potential by Buried Masses. *The Use of Artificial Satellites for Geodesy*, 15:121–124.

Barthelmes, F. (1986). *Untersuchungen zur Approximation des äußeren Schwerefeldes der Erde durch Punktmassen mit optimierten Positionen*. PhD thesis, Zentralinstitut für Physik der Erde.

Barthelmes, F. (1989). Local gravity field approximation by point masses with optimized positions. In: *Gravity field variations*, no. 102, pp. 157–167. Veröffentlichungen des Zentralinstituts für Physik der Erde.

Barthelmes, F. (2013). Definition of Functionals of the Geopotential and Their Calculation from Spherical Harmonic Models. Technical Report STR09/02, Revised Edition, GFZ Potsdam.

Barthelmes, F., Dietrich, R., and Lehmann, R. (1991). Representation of the Global Gravity Field by Point Masses on Optimized Positions Based on Recent Spherical Harmonics Expansions. Poster Presented at the XX. General Assembly of the International Union of Geodesy and Geophysics, Vienna.

Bentel, K. (2013). *Regional Gravity Modeling in Spherical Radial Basis Functions - On the Role of the Basis Function and the Combination of Different Observation Types*. PhD thesis, Norwegian University of Life Sciences.

Bingham, R. J., Tscherning, C., and Knudsen, P. (2011). An initial investigation of the GOCE error variance-covariance matrices in the context of the GOCE user toolbox project. In: *Proceedings of 4th International GOCE User Workshop*. European Space Agency.

Bock, H., Jäggi, A., Meyer, U., Visser, P., van den IJssel, J., van Helleputte, T., Heinze, M., and Hugentobler, U. (2011). GPS-derived orbits for the GOCE satellite. *Journal of Geodesy*, 85(11):807–818.

Box, G. E. P. and Muller, M. E. (1958). A note on the generation of random normal deviates. *The Annals of Mathematical Statistics*, 29(2):610–611.

Boyd, S. and Vandenberghe, L. (2004). *Convex Optimization*. Cambridge University Press.

Brockmann, J. M., Zehentner, N., Höck, E., Pail, R., Loth, I., Mayer-Gürr, T., and Schuh, W.-D. (2014). EGM_TIM_RL05: An independent geoid with centimeter accuracy purely based on the GOCE mission. *Geophysical Research Letters*, 41(22):8089–8099.

Brooks, S. P. and Giudici, P. (2000). Markov chain Monte Carlo convergence assessment via two-way analysis of variance. *Journal of Computational and Graphical Statistics*, 9(2):266–285.

Brooks, S. P., Giudici, P., and Philippe, A. (2003). Nonparametric convergence assessment for MCMC model selection. *Journal of Computational and Graphical Statistics*, 12(1):1–22.

Bruinsma, S. L., Marty, J. C., Balmino, G., Biancale, R., Förste, C., Abrikosov, O., and Neumayer, H. (2010). GOCE gravity field recovery by means of the direct numerical method. Presented at the ESA Living Planet Symp., Bergen, Norway.

Cai, J. and Grafarend, E. W. (2007). The Statistical Property of the GNSS Carrier Phase Observations and its Effects on the Hypothesis Testing of the Related Estimators . In: *Proceedings of ION GNSS, Forth Worth, TX*.

Carlson, R. E. and Foley, T. A. (1991). The parameter R^2 in multiquadric interpolation. *Computers & Mathematics with Applications*, 21(9):29–42.

Castelloe, J. M. and Zimmerman, D. L. (2002). Convergence assessment for reversible jump MCMC samplers. Technical Report 313, Department of Statistics and Actuarial Science, University of Iowa.

Chambodut, A., Panet, I., Mandea, M., Diament, M., Holschneider, M., and Jamet, O. (2005). Wavelet frames: an alternative to spherical harmonic representation of potential fields. *Geophysical Journal International*, 163(3):875—899.

Chen, M.-H. and Shao, Q.-M. (1999). Monte Carlo Estimation of Bayesian Credible and HPD Intervals. *Journal of Computational and Graphical Statistics*, 8(1):69–92.

Chib, S. and Greenberg, E. (1995). Understanding the Metropolis-Hastings Algorithm. *The American Statistician*, 49(4):327–335.

Claessens, S. J., Featherstone, W. E., and Barthelmes, F. (2001). Experiences with point-mass gravity field modelling in the Perth region, Western Australia. *Geomatics Research Australasia*, 75:53–86.

Cordell, L. (1992). A scattered equivalent-source method for interpolation and gridding of potential-field data in three dimensions. *Geophysics*, 57(4):629–636.

Cowles, M. K. and Carlin, B. P. (1996). Markov chain Monte Carlo convergence diagnostics: a comparative review. *Journal of the American Statistical Association*, 91(434):883–904.

Dampney, C. N. G. (1969). The equivalent source technique. *Geophysics*, 34(1):39–53.

Devroye, L. (1986). *Non-Uniform Random Variate Generation*. Springer.

Dick, S., Kleine, E., Müller-Navarra, S., Klein, H., and Komo, H. (2001). The operational circulation model of BSH (BSHcmod). *Berichte des Bundesamtes für Seeschifffahrt und Hydrographie*, (Nr.29/2001).

DiMatteo, I., Genovese, C. R., and Kass, R. E. (2001). Bayesian Curve-Fitting with Free-Knot Splines. *Biometrika*, 88(4):1055–1071.

Ditmar, P. and van Eck van der Sluijs, A. A. (2004). A technique for modeling the earth's gravity field on the basis of satellite accelerations. *Journal of Geodesy*, 78(1-2):12–33.

Dortet-Bernadet, J.-L. and Wicker, N. (2008). Model-based clustering on the unit sphere with an illustration using gene expression profiles. *Biostatistics*, 9(1):66–80.

Drinkwater, M. R., Haagmans, R., Muzi, D., Popescu, A., Floberghagen, R., Kern, M., and Fehringer, M. (2007). The GOCE gravity mission: ESA's first core explorer. In: *Proceedings of the 3rd International GOCE User Workshop, Frascati, Italy, 2006*. ESA SP-627.

Eicker, A. (2008). *Gravity Field Refinement by Radial Basis Functions from In-situ Satellite Data*. PhD thesis, Universität Bonn.

Eicker, A., Schall, J., and Kusche, J. (2014). Regional gravity modelling from spaceborne data: case studies with GOCE. *Geophysical Journal International*, 196(3):1431—-1440.

Elsaka, B. (2010). *Simulated Satellite Formation Flights for Detecting the Temporal Variations of the Earth's Gravity Field*. PhD thesis, Universität Bonn.

European GOCE Gravity Consortium (EGG-C) (2010). GOCE Standards. GO-TN-HPF-GS-0111.

Fischer, D. (2011). *Sparse Regularization of a Joint Inversion of Gravitational Data and Normal Mode Anomalies*. PhD thesis, Universität Siegen.

Fischer, D. and Michel, V. (2012). Sparse regularization of inverse gravimetry—case study: spatial and temporal mass variations in South America. *Inverse Problems*, 28:065012 (34pp).

Fraiture, L. (2012). Uniformly Distributed Random Directions in Bounded Spherical Areas. Part I: conventional Approaches for Attitude Purposes. `http://luc-fraiture.com/wp/wp-content/uploads/luc-fraiture-note8.pdf` (Retrieved: May 31, 2019).

Freeden, W., Gervens, T., and Schreiner, M. (1998). *Constructive Approximation on the Sphere*. Oxford University Press.

Gallagher, K., Charvin, K., Nielsen, S., Sambridge, M., and Stephenson, J. (2009). Markov chain Monte Carlo (MCMC) sampling methods to determine optimal models, model resolution and model choice for Earth Science problems. *Marine and Petroleum Geology*, 26:525–535.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian Analysis*, 1(3):515–534.

Gelman, A. and Rubin, D. B. (1992). lnference from Iterative Simulation Using Multiple Sequences. *Statistical Science*, 7(4):457–472.

Geyer, C. J. (2005). Markov Chain Monte Carlo Lecture Notes. `http://www.stat.umn.edu/geyer/f05/8931/n1998.pdf` (Retrieved: May 31, 2019).

Geyer, C. J. (2011). Introduction to Markov Chain Monte Carlo. In: Brooks, S., Gelman, A., Jones, G. L., and Meng, X.-L. (Eds.), *Handbook of Markov Chain Monte Carlo*. Chapman & Hall/CRC.

Grafarend, E. and Awange, J. (2012). *Applications of Linear and Nonlinear Models, Fixed Effects, Random Effects, and Total Least Squares*. Springer.

Green, P. J. (1995). Reversible Jump Markov Chain Monte Carlo Computation and Bayesian Model Determination. *Biometrika*, 82(4):711–732.

Green, P. J. and Hastie, D. I. (2009). Reversible jump MCMC. *Genetics*, 155(3):1391–1403.

Gundlich, B. and Kusche, J. (2008). Monte Carlo Integration for Quasi-linear Models. In: Xu, P., Liu, J., and Dermanis, A. (Eds.), *VI Hotine-Marussi Symposium on Theoretical and Computational Geodesy, Wuhan, China, 2006*, vol. 132 of *International Association of Geodesy Symposia*, pp. 337–344. Springer, Berlin, Heidelberg.

Hashemi Farahani, H., Ditmar, P., Klees, R., Teixeira da Encarnação, J., Liu, X., Zhao, Q., and Guo, J. (2013). Validation of static gravity field models using GRACE K-band ranging and GOCE gradiometry data. *Geophysical Journal International*, 194(2):751–771.

Hastie, D. I. and Green, P. J. (2012). Model Choice using Reversible Jump Markov Chain Monte Carlo. *Statistica Neerlandica*, 66(3):309–338.

Hastings, W. K. (1970). Monte Carlo Sampling Methods Using Markov Chains and Their Applications. *Biometrika*, 57(1):97–109.

Hayn, M., Holschneider, M., and Panet, I. (2013). Adaptative gravity modelling from GOCE gradient data over the Himalaya. AGU Fall Meeting Abstracts.

Hofmann-Wellenhof, B. and Moritz, H. (2006). *Physical Geodesy*. Springer.

Holschneider, M., Chambodut, A., and Mandea, M. (2003). From global to regional analysis of the magnetic field on the sphere using wavelet frames. *Physics of the Earth and Planetary Interiors*, 135(2-3):107–124.

Hyndman, R. J. (1996). Computing and Graphing Highest Density Regions. *The American Statistician*, 50(2):120–126.

ICGEM (2013). International Centre for Global Earth Models. `http://icgem.gfz-potsdam.de/ICGEM/`.

Ilk, K.-H. (1983). *Ein Beitrag zur Dynamik ausgedehnter Körper – Gravitationswechselwirkung*. Postdoctoral thesis, TU München.

Ilk, K.-H., Löcher, A., and Mayer-Gürr, T. (2008). Do We Need New Gravity Field Recovery Techniques for the New Gravity Field Satellites? In: Xu, P., Liu, J., and Dermanis, A. (Eds.), *VI Hotine-Marussi Symposium on Theoretical and Computational Geodesy, Wuhan, China, 2006*, vol. 132 of *International Association of Geodesy Symposia*, pp. 3–9. Springer, Berlin, Heidelberg.

Ivins, E. R., Watkins, M. M., Yuan, D., Dietrich, R., Casassa, G., and Rülke, A. (2011). On-land ice loss and glacial isostatic adjustment at the Drake Passage: 2003–2009. *Journal of Geophysical Research*, 116:B02403.

Jannink, J.-L. and Fernando, R. L. (2004). Note On the Metropolis-Hastings Acceptance Probability to Add or Drop a Quantitative Trait Locus in Markov Chain Monte Carlo-Based Bayesian Analyses. *Genetics Society of America*, 166:641–643.

Jarosz, A. F. and Wiley, J. (2014). What Are the Odds? A Practical Guide to Computing and Reporting Bayes Factors. *Journal of Problem Solving*, Special Issue, 7:2–9.

Jasra, A., Holmes, C. C., and Stephens, D. A. (2005). Markov Chain Monte Carlo Methods and the Label Switching Problem in Bayesian Mixture Modeling. *Statistical Science*, 20(1):50–67.

Jeffreys, H. (1961). *The Theory of Probability*. Oxford University Press, 3rd edition.

Jekeli, C. (2005). Spline Representations of Functions on a Sphere for Geopotential Modeling. Technical Report No. 475, The Ohio State University.

Kass, R. E. and Raftery, A. E. (1995). Bayes Factors. *Journal of the American Statistical Association*, 90(430):773–795.

Kent, J. T. (1982). The Fisher-Bingham Distribution on the Sphere. *Journal of the Royal Statistical Society. Series B (Methodological)*, 44(1):71–80.

Klees, R., Tenzer, R., Prutkin, I., and Wittwer, T. (2008). A data-driven approach to local gravity field modelling using spherical radial basis functions. *Journal of Geodesy*, 82(8):457—-471.

Klees, R. and Wittwer, T. (2007). A data-adaptive design of a spherical basis function network for gravity field modelling. In: Tregoning, P. and Rizos, C. (Eds.), *Dynamic Planet*, vol. 130 of *International Association of Geodesy Symposia*, pp. 322–328. Springer, Berlin, Heidelberg.

Koch, K. and Kusche, J. (2002). Regularization of geopotential determination from satellite data by variance components. *Journal of Geodesy*, 76(5):641–652.

Koch, K.-R. (1999). *Parameter Estimation and Hypothesis Testing in Linear Models*. Springer, 2nd edition.

Koch, K.-R. (2007). *Introduction to Bayesian Statistics*. Springer, 2nd edition.

Koch, K. R., Kuhlmann, H., and Schuh, W.-D. (2010). Approximating covariance matrices estimated in multivariate models by estimated auto- and cross-covariances. *Journal of Geodesy*, 84(6):383–397.

Kusche, J. (2002). *Inverse Probleme bei der Gravitationsfeldbestimmung mittels SST- und SGG-Satellitenmissionen*. No. 548 in Reihe C. Deutsche Geodätische Kommission, München.

Kusche, J. (2003). A Monte-Carlo technique for weight estimation in satellite geodesy. *Journal of Geodesy*, 76(11-12):641–652.

Kusche, J., Rudolph, S., Feuchtinger, M., and Ilk, K. H. (2001). Gradiometric data analysis using icosahedral grids. In: *International GOCE User Workshop. ESA/ESTEC, Noordwijk*.

Lehmann, R. (1993). The method of free-positioned point masses – geoid studies on the Gulf of Bothnia. *Bulletin Géodésique*, 67:31–40.

Liebsch, G., Schirmer, U., Ihde, J., Denker, H., and Müller, J. (2006). Quasigeoidbestimmung für Deutschland. *DVW Schriftenreihe*, 49:127–146.

Lin, M. (2016). *Regional gravity field recovery using the point mass method*. PhD thesis, Universität Hannover.

Lin, M., Denker, H., and Müller, J. (2014). Regional gravity field modelling using free-positioned point masses. *Studia Geophysica et Geodaetica*, 58(2):207–226.

Lindstrom, M. J. (2002). Bayesian estimation of free-knot splines using reversible jumps. *Computational Statistics & Data Analysis*, 41(2):255–269.

Löcher, A. (2010). *Möglichkeiten der Nutzung kinematischer Satellitenbahnen zur Bestimmung des Gravitationsfeldes der Erde*. PhD thesis, Universität Bonn.

Luthcke, S. B., Sabaka, T. J., Loomis, B. D., Arendt, A. A., McCarthy, J. J., and Camp, J. (2013). Antarctica, Greenland and Gulf of Alaska land-ice evolution from an iterated GRACE global mascon solution. *Journal of Glaciology*, 59:613–631.

Luthcke, S. B., Zwally, H. J., Abdalati, W., Rowlands, D. D., Ray, R. D., Nerem, R. S., Lemoine, F. G., McCarthy, J. J., and Chinn, D. S. (2006). Recent Greenland Ice Mass Loss by Drainage System from Satellite Gravity Observations. *Science*, 314(5803):1286–1289.

MacKay, D. J. C. (2005). *Information Theory, Inference, and Learning Algorithms*, Model Comparison and Occam's Razor, pp. 343–355. Cambridge University Press.

Marchenko, A. and Abrikosov, O. (1995). Geoid in the West Ukraine Area Derived by Means of Non-Central Multipole Analysis Technique. In: Sünkel, H. and Marson, I. (Eds.), *Gravity and Geoid*, vol. 113 of *International Association of Geodesy Symposia*, pp. 624–629. Springer, Berlin, Heidelberg.

Marchenko, A. N., Barthelmes, F., Meyer, U., and Schwintzer, P. (2001). Regional Geoid Determination: An Application to Airborne Gravity Data in the Skagerrak. Technical Report STR01/07, GFZ Potsdam.

Mardia, K. V. and Jupp, P. E. (1999). *Directional Statistics*. John Wiley & Sons.

Matsumoto, M. and Nishimura, T. (1998). Mersenne Twister: A 623-dimensionally equidistributed uniform pseudo-random number generator. *ACM Transactions on Modeling and Computer Simulation: Special Issue on Uniform Random Number Generation*, 8(1):3–30.

Mayer-Gürr, T. (2006). *Gravitationsfeldbestimmung aus der Analyse kurzer Bahnbögen am Beispiel der Satellitenmissionen CHAMP und GRACE*. PhD thesis, Universität Bonn.

Mayer-Gürr, T., Ilk, K. H., Eicker, A., and Feuchtinger, M. (2005). ITG-CHAMP01: a CHAMP gravity field model from short kinematic arcs over a one-year observation period. *Journal of Geodesy*, 78(7–8):462–480.

Mayer-Gürr, T., Kurtenbach, E., and Eicker, A. (2010). ITG-Grace2010 gravity field model.

McCarthy, D. and Petit, G. (Eds.) (2004). *IERS Conventions 2003*. No. 32 in IERS Technical Notes. Verlag des Bundesamts fuer Kartographie und Geodäsie, Frankfurt am Main.

Mengersen, K. L., Robert, C. P., and Guihenneuc-Jouyaux, C. (1999). MCMC convergence diagnostics: a reviewww. *Bayesian Statistics*, 6:415–440.

Meschkowski, H. (1962). *Hilbertsche Räume mit Kernfunktion*, vol. 113. Springer.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equation of State Calculations by Fast Computing Machines. *Journal of Chemical Physics*, 21:1087–1092.

Metzler, B. and Pail, R. (2005). GOCE data processing: the spherical cap regularization approach. *Studia Geophysica et Geodaetica*, 49(4):441–462.

Michel, V. and Telschow, R. (2016). The Regularized Orthogonal Functional Matching Pursuit for Ill-Posed Inverse Problems. *SIAM Journal on Numerical Analysis*, 54(1):262–287.

Migliaccio, F., Reguzzoni, M., Gatti, A., Sansó, F., and Herceg, M. (2011). A GOCE-only global gravity field model by the space-wise approach. In: *Proceedings of 4th International GOCE User Workshop, Munich, Germany*. ESA SP-696.

Moritz, H. (1980). *Advanced Physical Geodesy*. Herbert Wichmann Verlag, Germany and Abacus Press, Great Britain.

Naeimi, M. (2013). *Inversion of satellite gravity data using spherical radial base functions*. PhD thesis, Universität Hannover.

O'Keefe, J. A. (1957). An application of Jacobi's integral to the motion of an earth satellite. *The Astronomical Journal*, 62:265.

Pail, R., Bruinsma, S., Migliaccio, F., Förste, C., Goiginger, H., Schuh, W.-D., Höck, E., Reguzzoni, M., Brockmann, J. M., Abrikosov, O., Veicherts, M., Fecher, T., Mayrhofer, R., Krasbutter, I., Sansò, F., and Tscherning, C. C. (2011). First GOCE gravity field models derived by three different approaches. *Journal of Geodesy*, 85:819–843.

Pail, R., Goiginger, H., Mayrhofer, R., Schuh, W.-D., Brockmann, J. M., Krasbutter, I., Höck, E., and Fecher, T. (2010). GOCE gravity field model derived from orbit and gradiometry data applying the time-wise method. In: *Proceedings of the ESA Living Planet Symposium, Bergen, Norway*. ESA SP-686.

Panet, I., Kuroishi, Y., and Holschneider, M. (2011). Wavelet modelling of the gravity field by domain decomposition methods: an example over Japan. *Geophysical Journal International*, 184(1):203—-219.

Pavlis, N. K., Holmes, S. A., Kenyon, S. C., and Factor, J. K. (2012). The development and evaluation of the Earth Gravitational Model 2008 (EGM2008). *Journal of Geophysical Research*, 117(B4).

Polson, N. G. and Scott, J. G. (2012). On the half-Cauchy prior for a global scale parameter. *Bayesian Analysis*, 7(4):887–902.

Popko, E. S. (2012). *Divided Spheres: Geodesics and the Orderly Subdivision of the Sphere*. CRC Press.

Press, W. H., Teukolsky, S. A., Vetterling, W. T., and Flannery, B. P. (2007). *Numerical Recipes, The Art of Scientific Computing*. Cambridge University Press, 3rd edition.

Reigber, C. (1969). *Zur Bestimmung des Gravitationsfeldes der Erde aus Satellitenbeobachtungen*. Postdoctoral thesis, TU München.

Reigber, C. (1995). Gravity field recovery from satellite tracking data. In: F. Sansó, R. R. (Ed.), *Theory of Satellite Geodesy and Gravity Field Determination. Lecture Notes in Earth Sciences 25.* Springer.

Richardson, S. and Green, P. J. (1997). On Bayesian Analysis of Mixtures with an Unknown Number of Components. *Journal of the Royal Statistical Society. Series B (Methodological)*, 59(4):731–792.

Robert, C. P. (2007). *The Bayesian Choice*. Springer, 2nd edition.

Robert, C. P., Chopin, N., and Rousseau, J. (2009). Harold Jeffreys's Theory of Probability Revisited. *Statistical Science*, 24(2):141–172.

Roodaki, A., Bect, J., and Fleury, G. (2012). Note on the computation of the Metropolis-Hastings ratio for Birth-or-Death moves in trans-dimensional MCMC algorithms for signal decomposition problems. Technical report, Supelec Systems Sciences, Gif-sur-Yvette, France.

Rowlands, D. D., Luthcke, S. B., McCarthy, J. J., Klosko, S. M., Chinn, D. S., Lemoine, F. G., Boy, J.-P., and Sabaka, T. J. (2010). Global mass flux solutions from GRACE: A comparison of parameter estimation strategies—Mass concentrations versus Stokes coefficients. *Journal of Geophysical Research*, 115:B01403.

Rülke, A., Liebsch, G., Sacher, M., Schäfer, U., Schirmer, U., and Ihde, J. (2013). Unification of European height system realizations. *Journal of Geodetic Science*, 2(4):343–354.

Rummel, R., Gruber, T., and Koop, R. (2004). High level processing facility for GOCE: products and processing strategy. In: *Proceedings of the 2nd International GOCE User Workshop, ESA-SP569*.

Rummel, R., Yi, W., and Stummer, C. (2011). GOCE gravitational gradiometry. *Journal of Geodesy*, 85:777–790.

Sambridge, M., Gallagher, K., Jackson, A., and Rickwood, P. (2006). Trans-dimensional inverse problems, model comparison and the evidence. *Geophysical Journal International*, 167(2):528–542.

Schall, J., Eicker, A., and Kusche, J. (2014). The ITG-Goce02 gravity field model from GOCE orbit and gradiometer data based on the short arc approach. *Journal of Geodesy*, 88(4):403—-409.

Schall, J., Mayer-Gürr, T., Eicker, A., and Kusche, J. (2011). A global gravitational field model from GOCE gradiometer observations. In: *Proceedings of the 4th International GOCE User Workshop, Munich, Germany*. ESA SP-696.

Schmidt, M., Fengler, M., Mayer-Gürr, T., Eicker, A., Kusche, H., Sánchez, L., and Han, S.-C. (2007). Regional gravity modeling in terms of spherical base functions. *Journal of Geodesy*, 81(1):17–38.

Schmidt, M., Seitz, F., and Shum, C. K. (2008). Regional four-dimensional hydrological mass variations from GRACE, atmospheric flux convergence, and river gauge data. *Journal of Geophysical Research*, 113:B10402.

Schneider, M. (1968). A general method of orbit determination. *Royal Aircraft Translation*, (1279).

Sinharay, S. (2003). Assessing Convergence of the Markov Chain Monte Carlo Algorithms: A Review. Technical Report RR-03-07, Educational Testing Service, Princeton.

Sisson, S. A. and Fan, Y. (2007). A distance-based diagnostic for trans-dimensional Markov chains. *Statistics and Computing*, 17(4):357–367.

Sneeuw, N. and van Gelderen, M. (1997). The Polar Gap. In: Sansó, F. and Rummel, R. (Eds.), *Geodetic Boundary Value Problems in View of the One Centimeter Geoid*, pp. 559–568. Springer, Berlin, Heidelberg.

Stephens, M. (2000). Dealing with Label Switching in Mixture Models. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 62(4):795–809.

Stewart, J. (1976). Positive definite functions and generalizations, an historical survey. *Rocky Mountain, Journal of Mathematics*, 6(3):409–434.

Tscherning, C. (1977). A note on the choice of norm when using collocation for the computation of approximations to the anomalous potential. *Bulletin Géodésique*, 51(2):137–147.

Ulrich, G. (1984). Computer Generation of Distributions on the m-Sphere. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 33(2):158–163.

van Gelderen, M. and Kopp, R. (1997). The use of degree variances in satellite gradiometry. *Journal of Geodesy*, 71(6):337–343.

Watkins, M. M., Wiese, D. N., Yuan, D.-N., Boening, C., and Landerer, F. W. (2015). Improved methods for observing Earth's time variable mass distribution with GRACE using spherical cap mascons. *Journal of Geophysical Research (Solid Earth)*, 120:2648—-2671.

Wenzel, J. (2012). Numerically stable sampling of the von Mises Fisher distribution on S^2 (and other tricks). `https://www.mitsuba-renderer.org/~wenzel/files/vmf.pdf` (Retrieved: May 31, 2019).

Wieczorek, M. A. and Simons, F. J. (2005). Localized spectral analysis on the sphere. *Geophysical Journal International*, 162(3):655–675.

Wittwer, T. (2009). *Regional gravity field modelling with radial basis functions*. PhD thesis, TU Delft.

Wood, A. T. A. (1987). The simulation of spherical distributions in the Fisher-Bingham family. *Communications in Statistics - Simulation and Computation*, 16(3):885–898.

Wood, A. T. A. (1994). Simulation of the von Mises Fisher distribution. *Communications in Statistics - Simulation and Computation*, 23(1):157–164.