Bayesian approach to ionospheric imaging with Gaussian Markov random field priors

Johannes Norberg

Doctoral dissertation, to be presented for public discussion with the permission of the Faculty of Science of the University of Helsinki, in Auditorium PII, Porthania, Yliopistonkatu 3, Helsinki, on the 19th of August, 2020 at 12 o'clock.

Department of Mathematics and Statistics
University of Helsinki
Helsinki, Finland

Supervisors

Prof. Samuli Siltanen, University of Helsinki, Finland

Prof. Markku Lehtinen, Sodankylä Geophysical Observatory, University of Oulu, Finland

Dr. Olaf Amm, Finnish Meteorological institute, Helsinki, Finland

Dr. Kirsti Kauristie, Finnish Meteorological institute, Helsinki, Finland

Pre-examiners

Prof. Aku Seppänen, University of Eastern Finland

Dr. M Mainul Hoque, Institute for Solar-Terrestrial Physics, Neustrelitz, Germany

Opponent

Prof. Matthew Angling, Spire Global Ltd UK

Custos

Prof. Samuli Siltanen, University of Helsinki, Finland

Contact information

Department of Mathematics and Statistics
P.O. Box 64 (Gustav Hällströmin katu 2)
FI-00014 University of Helsinki
Finland
URL: `http://mathstat.helsinki.fi/`
Telephone: +358 29 419 11

Finnish Meteorological Institute
P.O. Box 503 (Erik Palménin aukio 1)
FI-00101 Helsinki Finland
URL: `https://ilmatieteenlaitos.fi/`
Telephone: +358 29 539 1000

ISSN: 0782-6117
ISBN: 978-952-336-123-2 (paperback)
ISBN: 978-952-336-124-9 (pdf)
`https://doi.org/10.35614/isbn.9789523361249`

ILMATIETEEN LAITOS
METEOROLOGISKA INSTITUTET
FINNISH METEOROLOGICAL INSTITUTE

Published by	**Finnish Meteorological Institute**	Series title, number and report code of publication
	(Erik Palménin aukio 1), P.O. Box 503	Contributions, 173, FMI-CONT-173
	FIN-00101 Helsinki, Finland	Date 17.7.2020

Author	ORCID iD
Johannes Norberg	0000-0003-3155-8894

Title

Bayesian approach to ionospheric imaging with Gaussian Markov random field priors

Abstract

Ionosphere is the partly ionised layer of Earth's atmosphere caused by solar radiation and particle precipitation. The ionisation can start from 60 km and extend up to 1000 km altitude. Often the interest in ionosphere is in the quantity and distribution of the free electrons. The electron density is related to the ionospheric refractive index and thus sufficiently high densities affect the electromagnetic waves propagating in the ionised medium. This is the reason for HF radio signals being able to reflect from the ionosphere allowing broadcast over the horizon, but also an error source in satellite positioning systems.

The ionospheric electron density can be studied e.g. with specific radars and satellite in situ measurements. These instruments can provide very precise observations, however, typically only in the vicinity of the instrument. To make observations in regional and global scales, due to the volume of the domain and price of the aforementioned instruments, indirect satellite measurements and imaging methods are required.

Mathematically ionospheric imaging suffers from two main complications. First, due to very sparse and limited measurement geometry between satellites and receivers, it is an ill-posed inverse problem. The measurements do not have enough information to reconstruct the electron density and thus additional information is required in some form. Second, to obtain sufficient resolution, the resulting numerical model can become computationally infeasible.

In this thesis, the Bayesian statistical background for the ionospheric imaging is presented. The Bayesian approach provides a natural way to account for different sources of information with corresponding uncertainties and to update the estimated ionospheric state as new information becomes available. Most importantly, the Gaussian Markov Random Field (GMRF) priors are introduced for the application of ionospheric imaging. The GMRF approach makes the Bayesian approach computationally feasible by sparse prior precision matrices.

The Bayesian method is indeed practicable and many of the widely used methods in ionospheric imaging revert back to the Bayesian approach. Unfortunately, the approach cannot escape the inherent lack of information provided by the measurement set-up, and similarly to other approaches, it is highly dependent on the additional subjective information required to solve the problem. It is here shown that the use of GMRF provides a genuine improvement for the task as this subjective information can be understood and described probabilistically in a meaningful and physically interpretative way while keeping the computational costs low.

Publishing unit
Space Research and Observation Technologies

Classification (UDC)
519.676, 551.510.413.5

Keywords
Ionosphere, inverse problems, Bayes,
data assimilation, tomography

ISSN and series title
0782-6117
Finnish Meteorological Institute Contributions

ISBN
978-952-336-123-2 (paperback)
978-952-336-124-9 (pdf)

DOI
https://doi.org/10.35614/isbn.9789523361249

Language
English

Pages
85

ILMATIETEEN LAITOS
METEOROLOGISKA INSTITUTET
FINNISH METEOROLOGICAL INSTITUTE

Julkaisija	**Ilmatieteen laitos**	Julkaisun sarja, numero ja raporttikoodi
	(Erik Palménin aukio 1)	Contributions, 173, FMI-CONT-173
	PL 503, 00101 Helsinki	Päiväys 17.7.2020

Tekijä	ORCID iD
Johannes Norberg	0000-0003-3155-8894

Nimeke

Gaussiset Markovin satunnaiskentät bayesiläisessä ionosfäärin kuvantamisessa

Tiivistelmä

Ionosfääri on noin 60–1000 kilometrin korkeudella sijaitseva ilmakehän kerros, jossa kaasuatomien ja -molekyylien elektroneja on päässyt irtoamaan auringon säteilyn ja auringosta peräisin olevien nopeiden hiukkasten vaikutuksesta. Näin syntyneillä ioneilla ja vapailla elektroneilla on sähkö- ja magneettikenttien kanssa vuorovaikuttava sähkövaraus. Ionosfäärillä on siksi merkittävä rooli radioliikenteessä. Se voi mahdollistaa horisontin yli tapahtuvat pitkät radiolähetykset heijastamalla lähetetyn sähkömagneettisen signaalin takaisin maata kohti. Toisaalta ionosfääri vaikuttaa myös sen läpäiseviin korkeampitaajuuksisiin signaaleihin. Esimerkiksi satelliittipaikannuksessa ionosfäärin vaikutus on parhaassakin tapauksessa otettava huomioon, mutta huonoimmassa se voi estää paikannuksen täysin. Näkyvin ja tunnetuin ionosfääriin liittyvä ilmiö lienee revontulet.

Yksi keskeisistä suureista ionosfäärin tutkimuksessa on vapaiden elektronien määrä kuutiometrin tilavuudessa. Käytännössä elektronitiheyden mittaaminen on mahdollista mm. tutkilla, kuten Norjan, Suomen ja Ruotsin alueilla sijaitsevalla EISCAT-tutkajärjestelmällä, sekä raketti- tai satelliittimittauksilla. Mittaukset voivat olla hyvinkin tarkkoja, mutta tietoa saadaan ainoastaan tutkakeilan suunnassa tai mittalaitteen läheisyydestä. Näillä menetelmillä ionosfäärin tutkiminen laajemmalla alueella on siten vaikeaa ja kallista.

Olemassa olevat paikannussatelliitit ja vastaanotinverkot mahdollistavat ionosfäärin elektronitiheyden mittaamisen alueellisessa, ja jopa globaalissa mittakaavassa, ensisijaisen käyttötarkoituksensa sivutuotteena. Satelliittimittausten ajallinen ja paikallinen kattavuus on hyvä, ja kaiken aikaa kasvava, mutta esimerkiksi tarkkoihin tutkamittauksiin verrattuna yksittäisten mittausten tuottama informaatio on huomattavasti vähäisempää.

Tässä väitöstyössä kehitettiin tietokoneohjelmisto ionosfäärin elektronitiheyden kolmiulotteiseen kuvantamiseen. Menetelmä perustuu matemaattisten käänteisongelmien teoriaan ja muistuttaa lääketieteessä käytettyjä viipalekuvausmenetelmiä. Satelliittimittausten puutteellisesta informaatiosta johtuen työssä on keskitytty etenkin siihen, miten ratkaisun löytymistä voidaan auttaa tilastollisesti esitetyllä fysikaalisella ennakkotiedolla.

Erityisesti työssä sovellettiin gaussisiin Markovin satunnaiskenttiin perustuvaa uutta korrelaatiopriori-menetelmää. Menetelmä vähentää merkittävästi tietokonelaskennassa käytettävän muistin tarvetta, mikä lyhentää laskenta-aikaa ja mahdollistaa korkeamman kuvantamisresoluution.

Julkaisijayksikkö
Avaruustutkimus ja havaintoteknologiat

Luokitus (UDK)
519.676, 551.510.413.5

Asiasanat
Ionosfääri, inversio-ongelmat, Bayes, data-assimilaatio, tomografia

ISSN ja avainnimeke
0782-6117
Finnish Meteorological Institute Contributions

ISBN
978-952-336-123-2 (paperback)
978-952-336-124-9 (pdf)

DOI
https://doi.org/10.35614/isbn.9789523361249

Kieli
Englanti

Sivumäärä
85

List of publications

This thesis consists of an introductory part and the following original publications:

I J. Norberg, L. Roininen, J. Vierinen, O. Amm, D. McKay-Bukowski, and M. S. Lehtinen, Ionospheric tomography in Bayesian framework with Gaussian Markov random field priors, *Radio Sci.*, 50(2): 138–152, 2015. ISSN 1944799X. doi: 10.1002/2014RS005431.

II J. Vierinen, J. Norberg, M. S. Lehtinen, O. Amm, L. Roininen, A. Väänänen, P. J. Erickson and D. McKay-Bukowski. Beacon satellite receiver for ionospheric tomography *Radio Sci.*, 49(12):1141-1152, 2014. ISSN 1944799X. doi: 10.1002/2014RS005434.

III J. Norberg, I. I. Virtanen, L. Roininen, J. Vierinen, M. Orispää, K. Kauristie, and M. S. Lehtinen, Bayesian statistical ionospheric tomography improved by incorporating ionosonde measurements, *Atmos. Meas. Tech.*, 9(4): 1859–1869, 2016. ISSN 18678548. doi: 10.5194/amt-9-1859-2016.

IV J. Norberg, J. Vierinen, L. Roininen, M. Orispää, K. Kauristie, W. C. Rideout, A. J. Coster, and M. S. Lehtinen, Gaussian Markov Random Field Priors in Ionospheric 3-D Multi-Instrument Tomography, *IEEE Trans. Geosci. Remote Sens.*, 1–13, 2018. ISSN 1558-0644. doi: 10.1109/TGRS.2018.2847026.

Author's contribution

Publication I: "Ionospheric tomography in Bayesian framework with Gaussian Markov random field prior"

The paper presents the Gaussian Markov random field priors for two-dimensional ionospheric imaging. It is shown how the matrices providing the prior covariance information are built numerically. A simulation study is made by generating samples from prior distributions, from where tomographic measurements are simulated assuming a low Earth orbit satellite overflight and five ground-based receivers. The sampled ionosphere is then reconstructed with the presented algorithm in lower resolution. The results demonstrate the performance of Bayesian inversion in an unrealistic situation where the prior used to simulate the unknown electron density is known exactly and in a more realistic case, where a more rough prior is used. All the numerical simulations and main parts of the writing were carried out by the author.

Publication II:"Beacon satellite receiver for ionospheric tomography"

The paper demonstrates a dual frequency receiver for ground-based measurements of 150 and 400 MHz signals transmitted from low Earth orbit beacon satellites. The paper is a companion paper for Publication I and uses the method presented there for reconstructing the two-dimensional ionospheric electron density from measurements obtained with four new ground-based receivers from a COSMOS 2463 satellite overflight. The results are compared with results from an existing tomographic system in the same region. The deployment of most of the receivers, numerical analysis and writing for the parts concerning tomography were carried out by the author.

Publication III:"Bayesian statistical ionospheric tomography improved by incorporating ionosonde measurements"

The paper presents a study where the developed tomography method is used with different prior models. The compared prior models include the International Reference Ionosphere 2007 model, extrapolation of an European Incoherent Scatter Scientific Association's (EISCAT) ionosonde measurements and a prior assuming a zero electron density. The two-

dimensional results are validated with EISCAT's ultra-high frequency incoherent scatter radar measurements. The numerical analysis and main parts of the writing were carried out by the author. The EISCAT measurements were designed and carried out by the author, while I. I. Virtanen provided reanalysis for the radar measurements.

Publication IV:"Gaussian Markov Random Field Priors in Ionospheric 3-D Multi-Instrument Tomography"

This is the main paper of this thesis. It generalises the method presented in the earlier articles for multi-instrumental three-dimensional ionospheric imaging. The multi-instrument set-up consists of dense ground-based networks of radio receivers for GPS satellite signals, a low Earth orbit satellite receiver network, ionosonde, satellite occultation as well as satellite in situ measurements. The results are validated with EISCAT's ultra- and very-high frequency incoherent scatter radars. The numerical benefits from the sparsity of Gaussian Markov random field priors, inclusion of time propagation and differential bias correction of GPS satellite data are discussed. All the numerical simulations and main parts of the writing were carried out by the author. A. Coster and W. C. Rideout provided the differential bias correction for the GPS data.

Acknowledgements

I would like to pay my special regards to Markku Lehtinen for the original scientific ideas and for taking me in his group already at a very early stage of my studies.

I wish to express my deepest gratitude to the late Olaf Amm who first took me in FMI as his PhD student and got all this started. After Olaf, Kirsti Kauristie took me under her wing, and I would like to thank her for providing me with extraordinary time and space to carry out my endeavours, as well as for the kind support at every turn along the way.

I am thankful to my custos and supervisor at University of Helsinki, Samuli Siltanen, who took me in as a pig in a poke and has provided important support ever since.

I would like to thank Aku Seppänen and Mainul Hoque for the time and effort they put in the pre-examination of my thesis. The supportive feedback from respected professionals encouraged me a lot at the final stages of my work.

My wish was to have one of the true experts in the ionospheric imaging community to act as my opponent. Hence, I am very thankful to Matthew Angling for accepting the invitation.

This study has been carried out in Finnish Meteorological Institute, Helsinki, and Sodankylä Geophysical Observatory, University of Oulu. I wish to thank both organisations. I am especially grateful to Ari-Matti Harri, Jouni Pulliainen and Esa Turunen for their exemplary leadership that has provided an active, science-focused and inspirational working environment.

I cannot thank my colleagues enough. Without Antti Kero, Sebastian Käki, Sari Lasanen, Mikko Orispää, Pentti Posio, Tero Raita, Tomi Teppo, Thomas Ulich, Juha Vierinen and Ilkka Virtanen I would not have been able to complete this research. I am particularly grateful to Lassi Roininen, whose contribution to my study has been essential. In addition to his scientific findings, Lassi's interactive way of working and networks have helped me greatly.

I would like to thank all the collaborators: Anita Aikio, Anthea Coster, Michael Fletcher, Maxime Grandin, Esa Kallio, Thomas Leyser, Derek McKay, Minna Palmroth, William Rideout, Mike Rietveld, Tomas Tallkvist, Heikki Vanhamäki and Dan Whiter. I am looking forward to future collaborative projects.

I would like to acknowledge Baylie Damtie, Björn Gustavsson, Heikki Haario, Marko Laine, Jussi Markkanen, Markku Markkanen, Petteri Piiroinen, Jouni Susiluoto and Simo

Särkkä for the valuable discussions, advice and help they have provided.

I would like to thank Mwaba Hiltunen, Tomi Karppinen, Ulpu Leijala, Kimmo Rautiainen and Miia Salminen for their invaluable peer support.

I am also indebted to Pilvi Ahonen, Riitta Aikio, Dan Anderson, Mikael Frisk, Minna Huuskonen, Sari Jokiniemi, Marina Kurten, Juha Lemmetyinen, Harry Lonka, Kari Mäenpää, Teija Manninen, Marita Mökkönen, Arto Oksanen, Noora Partamies, Kaisa Ryynänen, Mikko Syrjäsuo, Matias Takala, Juho Vehviläinen and Riika Ylitalo for the help they have given me.

I am lucky to have so many good friends, old and new, whom I would like to thank, that the list would be too long and I would still forget someone. I hope that you know who you are.

My biggest gratitude goes first to my beloved godchildren: Lilja, Olavi, Siiri and Tatu, and finally, to my dear family: my grandmother Elma, sisters Taru and Anna, father Kauko, my mother Annukka and little niece Li.

Helsinki, Finland, July 2020
Johannes Norberg

Contents

Chapter 1

Introduction

The ionosphere is a shell of ionisation surrounding the Earth. The ionisation is controlled by solar radiation, particle precipitation, and interactions with the electrically neutral atmosphere. For ionospheric imaging the key plasma parameter is the electron density i.e. the number of free electrons divided by unit volume, often given in scaled units of $\frac{10^{11}}{\mathrm{m}^3}$. The atmospheric electron density is typically horizontally stratified and depends on factors including latitude, season, local time and solar activity. Figure 1.1 presents four vertical incoherent scatter radar measurement profiles of typical daytime ionospheric electron density over Tromsø, Norway. Generally, the electron density maximum takes place around an altitude of 300 km at the so-called F region. Below, around an altitude of 100 km is the E region. In local daytime the E region can be seen as a small enhancement of electron density below the much higher density in the F region. However, at high latitudes during auroral particle precipitation events, especially the E region can have very rapid changes with peak electron densities exceeding that of the F region. The D region takes place at altitudes between 60 and 90 km. The conditions in the D region are strongly coupled with neutral atmospheric processes and the region often has relatively small electron density. The ionosphere extends to around an altitude of 1000 km where it transforms into a plasmasphere with substantially lower electron content. The ionospheric electron density is also often described as *total electron content* (TEC) i.e. the integrated electron density between two locations. Vertical TEC (VTEC) is the TEC integrated along a vertical column. TEC and VTEC are usually given in TEC units $\left(1\ \mathrm{TECU} = \frac{10^{16}}{\mathrm{m}^2}\right)$.

Ionospheric electron density can be observed e.g. with incoherent scatter radars, ionosondes, satellite in situ measurements and remote measurements of the global navigation satellite system (GNSS) and low Earth orbit (LEO) satellite beacons. A two-dimensional simplification of different ionospheric electron density measurements is given in Figure 1.2.

In ionospheric imaging the aim is to reconstruct the two- or three-dimensional electron density from available measurements. The ground-based measurements of GNSS satellite beacon signals is typically the most important data component. The use of the terms

imaging, *tomography* and *data assimilation* is somewhat mixed in the ionospheric literature. The term *imaging* is usually used as a general term to cover the different reconstruction methods. In an optimal case of *tomography*, the unknown would be reconstructed mostly from the available measurements. When operating regionally, especially in two-dimensional cases, the situation in ionospheric imaging is similar to conventional tomographic problems such as medical X-ray tomography, and thus many of the same techniques have been used. On the other hand, in Global three- and four-dimensional situations, the measurements can be extremely sparse and even relatively large areas can be left without any measurements. In these situations some strict background models are required and combined optimally with the available observations. A more illustrative and the most commonly used term in this case is *data assimilation*. Data assimilation and its nomenclature originates mostly from the field of numerical weather prediction.

Even in the best situation, due to limitations in the measurement geometry, the ionospheric imaging problem can be considered a limited angle tomography problem with sparse measurements. This rules out the the generally widely used tomographic algorithms that are based on backprojection. Mathematically the tomographic imaging of ionosphere is an ill-posed inverse problem. In practice this means that the measurements do not contain enough information of the unknown electron densities to give a unique and realistic solution.

Most of the early approaches to ionospheric imaging were based on iterative reconstruction techniques that were developed independently within the fields of image processing and linear algebra. The starting point is an initial guess about the unknown, which is then modified iteratively to correspond with the measurements. The downside is that with incomplete data the result is very dependent on the initial value.

Another approach is provided by so-called classical regularisation methods. With classical regularisation the original problem is modified to as a similar well-posed problem as possible. The problem here is that the interpretation of classical regularisation methods is mostly mathematical: the reason for numerical instability is examined and adjusted. In severely ill-posed problems, as in ionospheric imaging, it can be difficult to interpret the regularisation physically. On the other hand, there can be a lot of physical information available that is difficult to represent accurately with these methods.

In the division used here, the last family of ionospheric imaging methods is provided by the Bayesian approach. In the Bayesian approach, a prior distribution is used to control the set of possible solutions. The prior distribution can often be understood as a probabilistic description of the uncertainty related to the physical quantity of interest. Even though the information in prior distribution and hence the whole approach can be considered subjective, there often exists indisputable physical information that can be used in the construction of the prior. Also, as in ill-posed problems some additional information is required in any case, it is beneficial to know how the information limits the possible solutions. Most of the data assimilation methods used in ionospheric imaging are Bayesian, where physical background models are used in the determination of the prior distribution.

The problems with the Bayesian approach are mostly computational. The numerical computations with proper probability distributions require operations with covariance matrices. Especially in the three-dimensional case the covariance matrices can become excessively large for computation. Hence, one way of seeing the differences within the Bayesian approaches is how the formation and computation of covariance matrices is handled.

In this work Gaussian Markov random field (GMRF) priors are introduced for Bayesian ionospheric imaging. GMRF is a Gaussian random field, but instead of mean and covariance, it is more conveniently defined with its mean and inverse covariance i.e the precision matrix. Following Roininen et al. (2011, 2013), with a suitable parametrisation, the precision matrix of a GMRF can give close approximations for known covariance functions and due to Markov property, the precision matrices are sparse matrices. This reduces the computational costs significantly, making the direct inversion possible for relatively large three-dimensional cases. The use of GMRF then allows the usage of proper prior distributions with physical interpretation, while keeping the computational burden similar to the classical regularisation methods.

The structure of this dissertation summary is the following. Chapter 2 introduces the mathematical background of tomography and the commonly used backprojection methods. In Chapter 3, the measurement model, the resulting linear inverse problem, the classical regularisation method solutions and the iterative solution techniques are presented. The Bayesian approach, and most importantly, the GMRF priors are introduced in Chapter 4. In the original publications, the modelling of the temporal dynamics is discussed only shortly in Publication IV. The generalisation of the method for the spatiotemporal situation is somewhat straightforward, but in a broader context so central that the most used recursive filtering algorithms are presented in Chapter 5. The different ionospheric measurements are then exhibited in Chapter 6. In Chapter 7, a review on the usage and development of the aforementioned imaging methods within the ionospheric research, as well as a description of the numerical method developed within this work is given. Finally, discussion and conclusions are provided in Chapter 8.

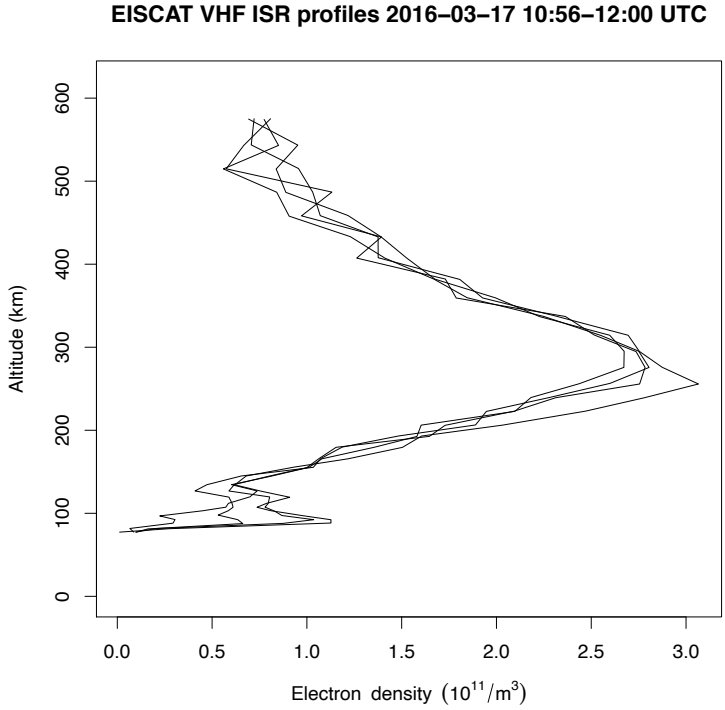

Figure 1.1: Four measurement profiles from European incoherent scatter scientific association's very-high frequency incoherent scatter radar in Tromsø, Norway. The profiles depict the typical vertical structure of daytime ionosphere, with the F-region maximum just below 300 km and the local E-region maximum around 100 km. Local time (UTC + 1 h).

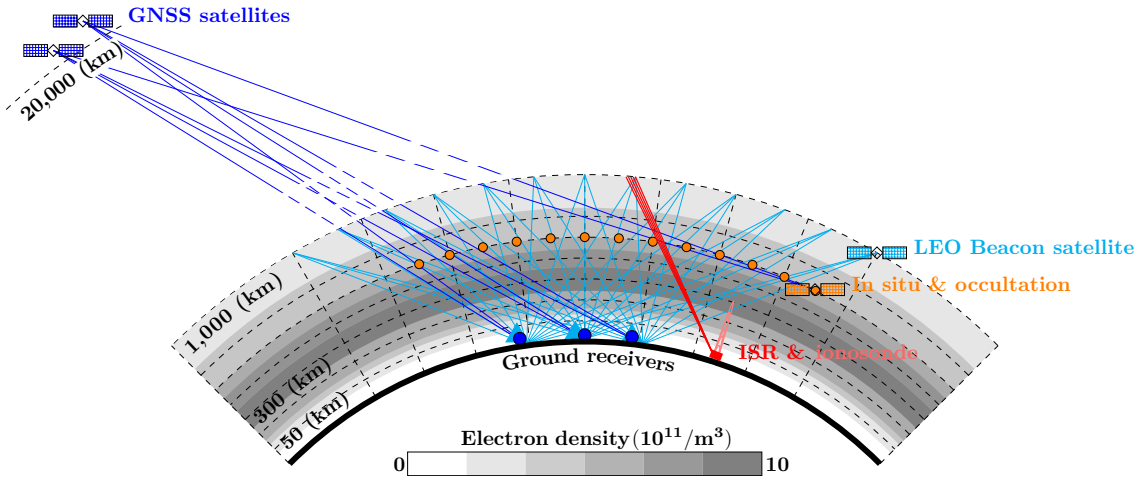

Figure 1.2: Two-dimensional simplification of measurements used in ionospheric imaging.

Chapter 2

Tomography

Tomography refers to cross-sectional imaging of an object from measurements provided by some penetrating waves. Tomographic methods are used in various fields from medicine to geophysics. Good overviews on tomography are provided by Kak and Slaney (1988); Natterer and Wübbeling (2001); Hsieh (2009).

Arguably the simplest and most common type of tomographic set-up is the parallel beam tomography. As the name suggests, several beams are transmitted in parallel on a two-dimensional plane. The beams propagate through the domain and are measured on the opposite side on a plane receiver. In X-ray tomography the measurement would be the attenuation of an X-ray signal during its pass. The set-up of transmitter and receiver planes is then circled around the object, to provide measurements from all directions. Besides the parallel beam tomography, there are various different scanning geometries, the fan beam and cone beam scans being probably the best known alternatives of regular scans.

2.1 Radon transform

Mathematically the situation in parallel beam tomography can be written by describing an unknown image as a function $f : \Psi \to \mathbb{R}$ on a physical domain $\Psi \in \mathbb{R}^2$. A measurement along a signal path of an angle perpendicular to θ and distance s from the origin can then be written generally as a *Radon transform*

$$\mathcal{R}f(\theta, s) = \int_{L(\theta,s)} f(z)\mathrm{d}z = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(z_1, z_2)\delta(z_1\cos\theta + z_2\sin\theta - s)\mathrm{d}z_1\mathrm{d}z_2, \qquad (2.1)$$

where $z = (z_1, z_2) \in \Psi$ and δ is a delta function defining the signal path $L(\theta, s)$ as a line in the image domain. With a fixed θ

$$P_\theta(s) := \mathcal{R}f(\theta, s), \qquad (2.2)$$

where $P_\theta(s)$ is the projection corresponding to parallel beam measurements made in a direction perpendicular to θ.

2.2 Filtered backprojection algorithm (FBP)

The main task in tomography is to reconstruct the unknown image $f(z_1, z_2)$ from the measured projections P_θ. The most straightforward approach is to backproject each measurement over the image domain along the corresponding signal path. When all backprojections are summed, the internal structures will accumulate in the reconstruction. However, the simple backprojection typically produces blurred results. Intuitively the blurring effect can be understood if one assumes a local minimum point with a value of zero. If the Radon transform is non-zero for any of the intersecting lines, in practice the reconstructed value of that point will always be greater than zero.

The blurring can be avoided with the *filtered backprojection* (FBP) algorithm. The FBP is based on the Fourier slice theorem, which states that when a two-dimensional image is projected to a plane with an angle θ, the one-dimensional Fourier transform of that projection corresponds to a radial slice of the two-dimensional Fourier transform of the same angle θ. Hence, the two-dimensional frequency domain of the unknown image can be built slice by slice with the Fourier transformed tomographic projections.

The use of inverse Fourier transform requires an interpolation to a rectangular grid in the frequency domain, or preferably, a change of variables between polar and rectangular coordinates. The change of variables introduces a Jacobian that can be interpreted as a high-pass filter. This is the filter part of FBP and in the frequency domain it is a multiplication operation. The filtering highlights edges and reduces the blurring in the final image.

Especially in most medical applications, where the conditions are well controlled and extensive measurements can be performed, the FBP is the first choice algorithm for its accuracy and relative ease of implementation (Kak and Slaney, 1988). However, the problems with FBP arise especially in the situations where the information provided by the measurements is limited and the Radon transform (2.1) is known only partially with respect to parameter θ or s or both. Including regularising additional information is difficult, hence the approach is severely affected by the incompleteness of data. It is also assumed in FPB that the measurements are precise and thus the measurement errors cannot be modelled explicitly.

2.3 Incomplete data

A common type of incompleteness in tomographic data is referred to as a *limited angle* tomography. As the name suggests, in limited angle tomography the θ angles are available only from a subset of the optimal sphere/half sphere. Examples of such cases are e.g. dental imaging (Hyvönen et al., 2010) and most geophysical tomographic problems such as borehole tomography (Justice et al., 1989).

Another type of incompleteness is the sparseness of the data. *Sparse tomography* can sometimes refer to sparseness of available measurement angles and hence overlap with the

definition of limited angle tomography above. However, it can also express the availability of different possible s and hence the resolution of each projection. The source points s of measurement paths can also be limited in range. With $q \in \mathbb{R}^+$ as a limiting constant, the situation $|s| > q$ is called an *exterior problem* and $|s| < q$ an *interior problem* (Natterer and Wübbeling, 2001).

The aforementioned limitations can be caused e.g. by inherent physical obstructions and inconveniences or medical or economical incentives. As a typical example, in medical tomography the patients exposition to harmful X-rays needs to be kept down, hence the radiation dose is reduced by using sparser angular resolution for measurement directions.

Another limiting factor can be the dynamics of the system. Even if the scanning geometry could provide sufficient accuracy with respect to measurement angles and amount and distribution of measurement paths, the unknown object can experience temporal changes in a shorter time scale than it takes to perform all the measurements (Hahn, 2015).

In satellite tomography, the angles in satellite-to-ground measurements are naturally limited (Figure 1.2). Additionally, for instance Brekke (1997) reports multifold change in electron density within 20 s, whereas GNSS data is typically integrated at least for some minutes for sufficient spatial coverage. Hence, ionospheric tomography can be considered a sparse limited angle tomography problem with relatively high temporal dynamics.

2.4 Discrete model

In practice the measurements (2.1) are made from a finite number of points and angles. Here R projections are assumed from a half sphere as

$$\boldsymbol{\theta} = \left(0, \frac{\pi}{R}, 2\frac{\pi}{R}, \ldots, (R-1)\frac{\pi}{R}\right)^{\mathrm{T}} = (\theta_1, \ldots, \theta_R)^{\mathrm{T}} \in \mathbb{R}^R.$$

For each angle θ, the $P_\theta(s)$ in Equation (2.2) is projected on S points

$$\mathbf{s} = \left(\left(1 - \frac{S+1}{2}\right)\Delta s, \ldots, \left(S - \frac{S-1}{2}\right)\Delta s\right)^{\mathrm{T}} = (s_1, \ldots, s_S) \in \mathbb{R}^S,$$

where Δs is the lateral offset between two adjacent projection points.

Altogether, this discretisation will provide measurements

$$\mathbf{m} = (P_{\theta_1}(s_1), \ldots, P_{\theta_1}(s_S), \ldots, P_{\theta_R}(s_S))^{\mathrm{T}} = (m_1, \ldots, m_j, \ldots, m_M)^{\mathrm{T}} \in \mathbb{R}^M,$$

where $M = RS$. Similarly, the corresponding lines L are denoted as

$$\mathbf{L} = (L(\theta_1, s_1), \ldots, L(\theta_1, s_S), \ldots, L(\theta_R, s_S))^{\mathrm{T}} = (L_1, \ldots, L_j, \ldots, L_M)^{\mathrm{T}} \subset \mathbb{R}^M,$$

For notational clarity, the one-index representation of the right-hand side will be used for all above measurement variables in the sequel.

For numerical modelling, also the unknown function needs to be discretised and understood as an array of unknown values. Typically the function is evaluated on a cartesian grid on the domain Ψ. As the unknown function is typically an image, it is easier to understand and visualise as a matrix, but for the algebraic and notational convenience it is collapsed to a vector and reindexed to one-index represenation

$$\boldsymbol{f} := f(\mathbf{z}) = (f(z_{1,1}), \ldots, f(z_{n,1}), \ldots, f(z_{n,n}))^{\mathrm{T}} \in \mathbb{R}^N,$$

where an $n \times n = N$ discretisation is made at points $\mathbf{z} = (z_{1,1}, \ldots, z_{n,1}, \ldots, z_{n,n})^{\mathrm{T}} \in \mathbb{R}^{N \times 2}$.

Each measurement m_j can then be modelled as an integral in Equation (2.1) and approximated as a Riemann sum

$$m_j = \int_{L_j} f(z) \mathrm{d}z \approx \sum_{i=1}^{N} a_{ji} f_i, \tag{2.3}$$

where $a_{ji} \in \mathbb{R}$ gives the intersection length between the path L_j and pixel i. In matrix form the measurements can then be written as

$$\mathbf{m} \approx \mathbf{A}\boldsymbol{f}, \tag{2.4}$$

where $\mathbf{A} \in \mathbb{R}^{M \times N}$ is a *theory matrix*, where row j is a vector $\mathbf{a}_j = (a_{j1}, \ldots, a_{ji}, \ldots a_{jN}) \in \mathbb{R}^N$.

Generally, the extension to three-dimensional tomography is often carried out by reducing the problem to several two-dimensional problems and reconstructed layer by layer. An alternative approach is to move the two-dimensional scan along the axis of symmetry during the scan. This will result in a three-dimensional helical scan. In cone beam tomography the setup is similar to fan beam, but whereas the fan beam is considered two-dimensional and the corresponding measurement one-dimensional, here the transmitted signal is a three-dimensional cone, received on a plane as a two-dimensional measurement. The Radon transform (2.1) and its inverse apply directly to parallel beam geometry, but alternative formulations for different scans are available and provided e.g. by Natterer and Wübbeling (2001). In ionospheric tomography, in a case where one satellite overflight is measured over a chain of receivers, the problem can be modelled as two-dimensional tomography. As the satellites have different orbits, in a general case where all possible measurements from several satellites are utilised, the measurements take place irregularly in a volume and the problem needs to be modelled in three dimensions.

Three-dimensional discrete model

In a three-dimensional case where the tomographic analysis is carried out directly in $\Psi \in \mathbb{R}^3$, the dimension of the unknown increases to

$$\boldsymbol{f} := f(\mathbf{z}) = (f(z_{1,1,1}), \ldots, f(z_{n,1,1}), \ldots, f(z_{n,n,1}), \ldots, f(z_{n,n,n}))^{\mathrm{T}} \in \mathbb{R}^N,$$

where now

$$\mathbf{z} = (z_{1,1,1}, \ldots, z_{n,1,1}, \ldots, z_{n,n,1}, \ldots, z_{n,n,n})^{\mathrm{T}} \in \mathbb{R}^{N \times 3} \tag{2.5}$$

and $n \times n \times n = N$. The Equations (2.3) and (2.4) remain the same with the corresponding change in dimensions.

As each measurement is here assumed an integral over a line, one measurement intersects only with a small portion of voxels. It is then notable that as the index i in the discrete models run through all unknown voxels, most of the intersection lengths a_i are typically zero and thus the matrix $\mathbf{A}$ is a so-called sparse matrix (see Section 4.3).

Chapter 3

Inverse problem

This section presents the general measurement model used in ionospheric imaging. As will be shown later in Chapter 6, most of the measurements used in the ionosphere imaging are fairly straightforward to linearise. Hence, here the focus is on the linear case. Particular attention is paid to the mathematical interpretation of issues caused by incomplete data and the means to overcome them. The main references for this chapter are Kaipio and Somersalo (2005), Calvetti and Somersalo (2007), Mueller and Siltanen (2012).

3.1 General model

A general forward model with measurement error is here given as

$$\mathbf{m} = \mathcal{A}(f, \boldsymbol{\varepsilon}), \tag{3.1}$$

where $f : \mathbb{R}^d \to \mathbb{R}$, $\mathcal{A}$ is a possibly nonlinear observation operator applied to function f and $\mathbf{m} \in \mathbb{R}^M$ the corresponding measurement vector. All physical measurements suffer from some degree of measurement errors. The error can be related to instrumentation, measurement conditions, or natural variability in the measured phenomena etc. Here a general measurement error $\boldsymbol{\varepsilon}$ is included in the model.

3.2 Linear model

If the observation operator is linear and the measurement error additive, the model can be written as

$$\mathbf{m} = Af + \boldsymbol{\varepsilon}, \tag{3.2}$$

where $f : \mathbb{R}^d \to \mathbb{R}$, A is a linear observation operator applied to function f and the measurement error vector $\boldsymbol{\varepsilon} \in \mathbb{R}^M$ is now additive. Here a zero-mean Gaussian measurement error $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}})$ is assumed. The measurement error and the function f are also assumed statistically independent $\boldsymbol{\varepsilon} \perp f$.

3.2.1 Discretisation

For numerical computations, a discrete model is required. Here a model

$$\mathbf{m} = \mathbf{A}\boldsymbol{f} + \boldsymbol{\varepsilon}, \tag{3.3}$$

is assumed, where similarly to Equation (2.4), the vector $\boldsymbol{f} \in \mathbb{R}^N$ is a discrete approximation of f on lattice $\mathbf{z} \in \mathbb{R}^{N \times d}$ and $\mathbf{A} \in \mathbb{R}^{M \times N}$ a linear transformation matrix that is a discrete approximation of A.

The discrete numerical model is always inaccurate compared to real-world measurements and can induce errors. For discussion on modelling errors see Kaipio and Somersalo (2007).

3.3 Ill-posed problem

Modelling of physical phenomena often results in mathematical problems where the unknown quantities of interest are measured indirectly. The actual measured property is not the primary interest, but physically and mathematically related to it. In the model equations of the previous section, the interest is not in the measurement $\mathbf{m}$, but in the unknown $\boldsymbol{f}$. The task is, given the measurements and the measurement model, solve the unknown $\boldsymbol{f}$, or, in presence of a noise model, estimate $\boldsymbol{f}$. The task is generally known as an *inverse problem*. To understand the origins of the difficulties and inaccuracies that arise with the inverse problems, the concept of a well-posed problem is first recalled.

Following Hadamard, a *well-posed* problem satisfies the following properties:

1. The solution exists.
2. The solution is unique.
3. The solution changes continuously with respect to the data.

If one or several of these properties is violated, the problem is referred to as *ill-posed*. In the linear situation given in Equation (3.3), the first condition is fulfilled if and only if $\mathbf{m} \in \text{Range}(\mathbf{A})$. It can be violated by the approximative nature of matrix $\mathbf{A}$ and the noise model. The second condition is fulfilled if and only if $\text{Ker}(\mathbf{A}) = \{0\}$, which depends on the geometry of the measurements. To see how and when the third condition is violated the *singular value decomposition* (SVD) becomes an essential tool.

SVD of matrix $\mathbf{A}$ can be written as

$$\mathbf{A} = \mathbf{U}\mathbf{D}\mathbf{V}^{\mathrm{T}}, \tag{3.4}$$

where $\mathbf{U} \in \mathbb{R}^{M \times M}$ and $\mathbf{V} \in \mathbb{R}^{N \times N}$ are orthogonal matrices and

$$\mathbf{D} \in \mathbb{R}^{M \times N} = \text{diag}(d_1, \ldots, d_{\min(M,N)}) \tag{3.5}$$

is a non-negative diagonal matrix, where $d_1 \geq d_2 \geq \cdots \geq d_{\min(M,N)}$ are singular values. Generally $\mathbf{D}$ is not a square matrix and even if $N = M$ it can be singular or near singular. This can be concluded from the *condition number*

$$\mathrm{Cond}(\mathbf{A}) = \frac{d_1}{d_{\min(M,N)}}. \tag{3.6}$$

For the third condition of Hadamard, it is required that the condition number of $\mathbf{A}$ is not excessively large (Kaipio and Somersalo, 2005). A matrix with a large condition number is called *ill-conditioned*.

In exact equations of linear algebra one would be concerned with only the first and the second of Hadamard's conditions. The concept of ill-conditioning rises with real world measurements or computational approaches that are contaminated with errors. An ill-conditioned problem is sensitive to errors as even small measurements errors can get amplified to have unrealistically large effects on the numerical solution.

To obtain a unique and stable solution for an inverse problem, some manouvres are required to overcome the ill-posedness. In the following sections, first the classical direct regularisation methods are presented, then the most commonly used iterative reconstruction techniques are described, before going to the Bayesian approach for inverse problems.

3.4 Classical regularisation methods

Approaches to solving the ill-posed problem are often referred to as regularisation methods, stabilisation or prior information. Usually, the procedure can be seen as not solving the original ill-posed problem, but a very similar one that is well-posed.

In the following, the most commonly used regularisation approaches are presented. SVD (3.4) is used here to display the ill-conditioning effect, as well as to demonstrate how the differences between the methods can be reduced to selection of a suitable diagonal matrix to replace the nonexistent inverse of the singular value matrix $\mathbf{D}$ (3.5).

3.4.1 Least squares solution (LS)

For an overdetermined linear inverse problem of the form given in Equation (3.3), often the first attempt to obtain a solution is made with the *least squares* (LS) method

$$\boldsymbol{f}_{\mathrm{LS}} = \underset{\boldsymbol{f}\in\mathbb{R}^N}{\arg\min} \|\mathbf{m} - \mathbf{A}\boldsymbol{f}\|^2 = \mathbf{A}^{\dagger}\mathbf{m} \tag{3.7}$$

where

$$\mathbf{A}^{\dagger} = \left(\mathbf{A}^{\mathrm{T}}\mathbf{A}\right)^{-1}\mathbf{A}^{\mathrm{T}} = \mathbf{V}\mathbf{D}^{\dagger}\mathbf{U}^{\mathrm{T}} \tag{3.8}$$

and

$$\mathbf{D}^{\dagger} = \left(\mathbf{D}^{\mathrm{T}}\mathbf{D}\right)^{-1}\mathbf{D}^{\mathrm{T}} = \mathrm{diag}(1/d_1, \ldots, 1/d_N) \in \mathbb{R}^{N\times M}. \tag{3.9}$$

In many occasions, in linear inverse problems, the matrix $\mathbf{A}$ can have so strong linear dependencies that, despite $M > N$, the problem is effectively underdetermined and ill-conditioned. Often this can be observed from rapidly decreasing singular values. The least squares method is unable to provide a reasonable solution for severely ill-conditioned problems.

3.4.2 Minimum norm solution (MN)

For an underdetermined system the least squares method fails as it cannot select a unique value satisfying the minimisation criteria of Equation (3.7). As the name suggest, the solution with *minimum norm* (MN) is selected from the subspace of all existing least squares solutions as

$$\boldsymbol{f}_{\mathrm{MN}} = \underset{\boldsymbol{f} \in \boldsymbol{f}_{\mathrm{LS}}}{\arg\min} \|\boldsymbol{f}\| = \mathbf{A}^{+}\mathbf{m} \tag{3.10}$$

where $\mathbf{A}^{+} = \mathbf{V}\mathbf{D}^{+}\mathbf{U}^{\mathrm{T}}$ and

$$\mathbf{D}^{+} = \mathrm{diag}(1/d_1, \ldots, 1/d_p, 0, \ldots, 0) \in \mathbb{R}^{N \times M} \tag{3.11}$$

and $p = \max\{i \mid 1 \leq i \leq M, d_i > 0\}$. However, differences of magnitude between the non-zero singular values can also make the MN solutions numerically unstable.

3.4.3 Truncated singular value decomposition (TSVD)

The *truncated singular value decomposition* (TSVD) solution can be obtained as a MN solution for a system where all the singular values of $\mathbf{A}$ that are less than a selected threshold α are set to zero.

$$\boldsymbol{f}_{\mathrm{TSVD}} = \underset{\boldsymbol{f} \in \boldsymbol{f}_{\mathrm{LS}_\alpha^+}}{\arg\min} \|\boldsymbol{f}\| = \mathbf{A}_\alpha^{+}\mathbf{m} \tag{3.12}$$

where

$$\boldsymbol{f}_{\mathrm{LS}_\alpha^+} = \underset{\boldsymbol{f} \in \mathbb{R}^N}{\arg\min} \|\mathbf{m} - \mathbf{A}_\alpha^{+}\boldsymbol{f}\|^2, \quad \mathbf{A}_\alpha^{+} = \mathbf{V}\mathbf{D}_\alpha^{+}\mathbf{U}^{\mathrm{T}} \tag{3.13}$$

and

$$\mathbf{D}_\alpha^{+} = \mathrm{diag}(1/d_1, \ldots, 1/d_{p_\alpha}, 0, \ldots, 0) \in \mathbb{R}^{N \times M} \tag{3.14}$$

with $p_\alpha = \max\{i \mid 1 \leq i \leq \min(N, M), d_i > \alpha\}$.

With LS and MN methods a unique solution can be found, but the solutions can remain ill-conditioned. TSVD stabilises the problem by replacing the $(\min(N, M) - p_\alpha)$ smallest singular values with zeros. Consequently the method ignores the corresponding singular vectors and typically simplifies the structure of the solution. However, there is no unambiguous criterion for selecting an optimal α.

3.4.4 Tikhonov regularisation

Tikhonov regularisation is also known as *Phillips* or *Tikhonov-Phillips regularisation* and *Ridge regression* (Tikhonov and Arsenin, 1977; Phillips, 1962; Hoerl and Kennard, 1970). The method concerns both the residuals and the L^2 norm of the solution. The Tikhonov regularised solution is the minimiser

$$\boldsymbol{f}_{\mathrm{T}} = \underset{\boldsymbol{f}\in\mathbb{R}^N}{\arg\min}\{\|\mathbf{m}-\mathbf{A}\boldsymbol{f}\|^2 + \alpha\|\boldsymbol{f}\|^2\} = \mathbf{A}_\alpha^\dagger\mathbf{m}, \tag{3.15}$$

where

$$\mathbf{A}_\alpha^\dagger = \left(\mathbf{A}^{\mathrm{T}}\mathbf{A}+\alpha\mathbf{I}\right)^{-1}\mathbf{A}^{\mathrm{T}} = \mathbf{V}\mathbf{D}_\alpha^\dagger\mathbf{U}^{\mathrm{T}}, \tag{3.16}$$

$$\mathbf{D}_\alpha^\dagger = \mathrm{diag}\left(\frac{d_1}{d_1^2+\alpha},\ldots,\frac{d_{\min(M,N)}}{d_{\min(M,N)}^2+\alpha}\right) \in \mathbb{R}^{N\times M}. \tag{3.17}$$

Here α is a *regularisation parameter* that controls the balance between the residuals and the norm of the solution.

From the diagonal matrix $\mathbf{D}_\alpha^\dagger$ it is somewhat intuitive to see how Tikhonov regularisation reverts to situations of LS and MN and how α controls the ill-conditioning. Similarly to TSVD, Tikhonov regularisation can provide solutions to ill-conditioned situations where LS and MN methods fail.

The optimal selection of regularisation parameter α is again an ambiguous task, however, different selection criteria are available such as Morozov's discrepancy principle and the L-curve method (Kaipio and Somersalo, 2005; Mueller and Siltanen, 2012).

3.4.5 Generalised Tikhonov regularisation

The Tikhonov regularised solution can be generalised to situation where additional constraints are set for the solution

$$\begin{aligned}\boldsymbol{f}_{\mathrm{T}} &= \underset{\boldsymbol{f}\in\mathbb{R}^N}{\arg\min}\{\|\mathbf{A}\boldsymbol{f}-\mathbf{m}\|^2 + \alpha\|\mathbf{L}(\boldsymbol{f}-\bar{\boldsymbol{f}})\|^2\} \\ &= \left(\mathbf{A}^{\mathrm{T}}\mathbf{A}+\alpha\mathbf{L}^{\mathrm{T}}\mathbf{L}\right)^{-1}\left(\mathbf{A}^{\mathrm{T}}\mathbf{m}+\alpha\mathbf{L}^{\mathrm{T}}\mathbf{L}\bar{\boldsymbol{f}}\right)\end{aligned} \tag{3.18}$$

In Equation (3.18) the norm at the right-hand side restricts the solution close to vector $\bar{\boldsymbol{f}}\in\mathbb{R}^N$. Often a difference matrix is selected as $\mathbf{L}\in\mathbb{R}^{N_L\times N}$ to require smoothness for the solution.

The generalised Tikhonov regularisation can also be seen as a solution for a system where the following linear constraints are added to the original equation (3.3). The so-called stacked form is given as

$$\begin{bmatrix}\mathbf{m}\\ \sqrt{\alpha}\mathbf{L}^{\mathrm{T}}\bar{\boldsymbol{f}}\end{bmatrix} = \begin{bmatrix}\mathbf{A}\\ \sqrt{\alpha}\mathbf{L}\end{bmatrix}\boldsymbol{f}_{\mathrm{T}}. \tag{3.19}$$

3.5 Iterative solutions for linear system

The most widely used iterative algorithms in tomograpy are the algebraic reconstruction technique and the EM algorithm. In ionospheric imaging the algebraic reconstruction technique and its derivatives have been used much more frequently and are therefore presented in this chapter. For the EM algorithm see e.g. Natterer and Wübbeling (2001).

Despite the fact that in the field of image reconstruction and tomography, the following iterative techniques have been developed specifically to handle incomplete error contaminated data, they still are general solvers for exact linear systems. Hence, given the error-contaminated measurements $\mathbf{m}$ and the matrix $\mathbf{A}$ in Equation (3.3), these methods actually solve a system

$$\mathbf{m} = \mathbf{A}\boldsymbol{f}_{\boldsymbol{\varepsilon}}. \tag{3.20}$$

However, to simplify the notations, the subindex $\boldsymbol{\varepsilon}$ has been omitted in the remainder of this chapter.

As will be stated in the sections below, the iterative techniques also provide some regularisation for the problem. In practise, the measurements predicted with iterative solutions will never equal the actual measurements with errors, and thus a stopping criterion is needed for the iteration. With incomplete data the methods are then referred to as *truncated iterative methods* (Kaipio and Somersalo, 2005), as the selection of the stopping criterion can be seen as a part of the regularisation scheme.

In the following, the notion of *iteration* refers to the update of $\boldsymbol{f}^{(k)}$ to obtain a new improved approfimation $\boldsymbol{f}^{(k+1)}$. One iteration can consists of other repetitive operations. All approaches require an initial starting value for the unknown, $\boldsymbol{f}^{(0)}$. With incomplete data the solution can be highly dependent on the initial value.

3.5.1 Kaczmarz method

The Kaczmarz method (Kaczmarz, 1937) is a general method for iterative approximative solutions for a system of linear equations, such as Equation 3.20. Besides the original article, an intuitive illustrated description of the method is provided in Kak and Slaney (1988). Another mathematically rigorous treatment is provided by Kaipio and Somersalo (2005).

The intuition of the convergence in the approach is that each measurement i.e. single rows

$$m_j = \mathbf{a}_j^{\mathrm{T}} \boldsymbol{f}, \quad 1 \leq j \leq M$$

define a hyperplane of dimension $\mathbb{R}^{N-1}$. The algorithm starts with an initial guess $\boldsymbol{f}^{(0)}$. The next iteration $\boldsymbol{f}^{(k+1)}$ is obtained by projecting the current solution $\boldsymbol{f}^{(k)}$ on the corresponding hyperplane. The projection for the $(k+1)^{\mathrm{th}}$ iteration can be written as

$$\boldsymbol{f}^{(k+1)} = \boldsymbol{f}^{(k)} + \mathbf{a}_j(\mathbf{a}_j^{\mathrm{T}}\mathbf{a}_j)^{-1}(\mathbf{a}_j^{\mathrm{T}}\boldsymbol{f}^{(k)} - m_j). \tag{3.21}$$

Often a relaxation parameter $0 < \lambda < 2$ is included to control the size of the correction performed at each iteration

$$\boldsymbol{f}^{(k+1)} = \boldsymbol{f}^{(k)} + \lambda \mathbf{a}_j (\mathbf{a}_j^{\mathrm{T}} \mathbf{a}_j)^{-1} (\mathbf{a}_j^{\mathrm{T}} \boldsymbol{f}^{(k)} - m_j). \tag{3.22}$$

For the first M iterations $j = k$, but often more iterations are required for convergence and the procedure is looped over all measurement equations several times, hence $1 \leq k \leq \kappa M$, where $\kappa \in \mathbb{N}$ and $j = k(\mathrm{mod}\ M) + 1$.

Another way to understand this algorithm is to see it as similar to backprojection. In Equations (3.21) and (3.22) the difference between the actual measurement m_j and the simulation of the same measurement from the current iteration $\mathbf{a}_j^{\mathrm{T}} \boldsymbol{f}^{(k)}$ is taken. A backprojection of the difference is then added to corresponding pixels along the ray path.

If a unique solution exists for the linear system, the iterative solution of the Kaczmarz method will converge to it (Tanabe, 1971). In an overdetermined situation $M > N$, if measurement noise is present, the linear system does not have a unique solution as the hyperplanes will not have a unique intersection and the solution will not converge to one point, but will drift between the intersections (Kak and Slaney, 1988). In an underdetermined system $N > M$, where there is again no unique solution available, the algorithm will endogenously provide regularisation as it will converge to the point $\hat{\boldsymbol{f}}$ of possible solutions that minimises $\|\hat{\boldsymbol{f}} - \boldsymbol{f}^{(0)}\|$ i.e. the distance between that point and the given initial value (Tanabe, 1971; Kak and Slaney, 1988).

The Kaczmarz method is primarily an algorithm for solving a linear system, however it is straightforward to include some regularising prior information in it. As said above, the initial value for the unknown already provides one regularisation scheme. In many applications the unknown cannot physically have negative values, if the projection nevertheless produces negative values, the values can be detected and set to zero within the algorithm.

3.5.2 Algebraic reconstruction technique (ART)

The *algebraic reconstruction technique* (ART) was presented in the field of image reconstruction (Gordon et al., 1970). The method is the Kaczmarz method, however some specific features are sometimes included in it.

In the original article Gordon et al. (1970), as well as Kak and Slaney (1988), the weights a_{ij} are not intersection lengths, but they are simply given a value 1 or 0 depending on whether the pixel center is within the signal path with width Δs or not. This has been done to ease the computation as the in/out decision is faster than computing the precise intersection lengths. However, this shortcut is known to often give rise to so-called salt and pepper noise (Kak and Slaney, 1988). Another feature often included in ART is the non-negativity constraint.

3.5.3 Multiplicative algebraic reconstruction technique (MART)

Whereas ART converges to the least squares solution of the linear system, the *multiplicative algebraic reconstruction technique* (MART) (Gordon et al., 1970) is a modification of ART that converges to the maximum entropy solution (Censor, 1983; Raymund et al., 1990). As the name suggests, instead of additional corrections, the unknowns along each raypath are scaled by multiplying as

$$\boldsymbol{f}_i^{(k+1)} = \left(\frac{m_j}{\mathbf{a}_j^{\mathrm{T}} \boldsymbol{f}^{(k)}} \right)^{\lambda_k a_{ji}} \boldsymbol{f}_i^{(k)}, \quad i = 1, \ldots, N \tag{3.23}$$

The update formula is written for a single unknown element as the exponent includes the intersection length between the j^{th} raypath and that specific i^{th} unknown element. The relaxation parameter fulfills $0 \leq \lambda_k \leq 1$ and the initial value for the unknown is given as $\boldsymbol{f}^{(0)} = e^{-1}\mathbf{1}$ (Censor, 1983).

3.5.4 Simultaneous iterative reconstruction tecnique (SIRT)

The update caused by single measurement j in Equation (3.21) can be written as a correction required for the unknown

$$\Delta \boldsymbol{f}^{(k+1),j} = \boldsymbol{f}^{(k+1)} - \boldsymbol{f}^{(k)} = \mathbf{a}_j (\mathbf{a}_j^{\mathrm{T}} \mathbf{a}_j)^{-1} (\mathbf{a}_j^{\mathrm{T}} \boldsymbol{f}^{(k)} - m_j), \quad 1 \leq j \leq M. \tag{3.24}$$

The *simultaneous iterative reconstruction technique* (SIRT) is a modification of ART where the correction (3.24) is computed from each measurement without updating $\boldsymbol{f}$ in between. Only after the corrections are computed for every measurement $j = 1, \ldots, M$, the new iteration is obtained as

$$\boldsymbol{f}_i^{(k+1)} = \boldsymbol{f}_i^{(k)} + \frac{1}{M_i} \sum_j^{M_i} \Delta \boldsymbol{f}_i^{(k+1),j}, \quad i = 1, \ldots, N,$$

where M_i is the number of measurements intersecting the corresponding unknown. The above formula is written for a single unknown element as the number M_i varies for different i. The convergence of SIRT is slower than in ART, but the quality of the reconstructed image can often be better (Kak and Slaney, 1988).

3.5.5 Simultaneous algebraic reconstruction tecnique (SART)

The *simultaneous algebraic reconstruction technique* (SART) (Andersen and Kak, 1984) combines some features of ART and SIRT methods. An important idea in SART is that the reconstruction can be improved with a more accurate modelling of the projections in the forward model. Hence, instead of the pixel approximation, the representation of

the unknown is generalised to a finite set of weighted base functions. In SART specifically bilinear elements are utilised as base functions. The iteration is then carried out non-sequentially with resemblance to SIRT, but in steps of individual projections. In one iteration, the corrections obtained from all measurements in one view angle are combined and used simultaneously in the update. Finally, when the corrections are applied to unknown elements along the ray paths in the projection, a Hamming window function is used to emphasise the corrections made at the middle of the ray and to damp the beginning and the end of the ray.

Chapter 4

Bayesian statistical approach

4.1 Introduction to Bayesian inference

In Bayesian statistical inference, all the variables and parameters are modelled as random variables. The randomness describes the lack of information concerning the realisations of the variables. The conclusions are based on probabilistic statements that are compiled with Bayes' formula

$$p(\boldsymbol{f}|\mathbf{m}) = \frac{p(\mathbf{m}|\boldsymbol{f})p(\boldsymbol{f})}{p(\mathbf{m})}, \tag{4.1}$$

where $p(\boldsymbol{f}|\mathbf{m})$ is the *posterior probability distribution* and $p(\boldsymbol{f})$ the *prior probability distribution* of $\boldsymbol{f}$, and $p(\mathbf{m}|\boldsymbol{f})$ is the *sampling distribution* of $\mathbf{m}$, but can also be seen as the *likelihood function* of $\boldsymbol{f}$ given $\mathbf{m}$. For a fixed $\mathbf{m}$, the marginal distribution $p(\mathbf{m})$ is a constant and independent of $\boldsymbol{f}$, however, it can be difficult to derive from a complicated joint distribution, hence the following *unnormalised posterior distribution* is often considered instead

$$p(\boldsymbol{f}|\mathbf{m}) \propto p(\mathbf{m}|\boldsymbol{f})p(\boldsymbol{f}). \tag{4.2}$$

The prior distribution indicates the most likely state and the related uncertainty of the unknown parameter $\boldsymbol{f}$ before the observations $\mathbf{m}$ are made. The posterior probability distribution is obtained by updating the prior distribution with the likelihood function that connects unknown parameters with the information provided by the observations. The posterior distribution is the solution that combines all the available information on $\boldsymbol{f}$.

As high-dimensional posterior distributions can be difficult to visualise, the distribution is usually characterised with some point and spread estimates. One of the mostly used point estimates is the *maximum a posteriori* (MAP) estimate

$$\boldsymbol{f}_{\mathrm{MAP}} = \underset{\boldsymbol{f}\in\mathbb{R}^N}{\arg\max}\, p(\boldsymbol{f}|\mathbf{m}). \tag{4.3}$$

If the maximiser for the estimator (4.3) exists, it is possible that it is not unique. Another point estimate is the *conditional mean* (CM), which is defined as

$$\boldsymbol{f}_{\mathrm{CM}} = \mathrm{E}\{\boldsymbol{f}|\mathbf{m}\} = \int_{\mathbb{R}^N} \boldsymbol{f}\ p(\boldsymbol{f}|\mathbf{m})\mathrm{d}\boldsymbol{f}. \tag{4.4}$$

Conditional covariance is an estimator for the spread of the posterior distribution. It is defined as

$$\mathrm{cov}(\boldsymbol{f}|\mathbf{m}) = \int_{\mathbb{R}^N} (\boldsymbol{f} - \boldsymbol{f}_{\mathrm{CM}})(\boldsymbol{f} - \boldsymbol{f}_{\mathrm{CM}})^{\mathrm{T}} p(\boldsymbol{f}|\mathbf{m})\mathrm{d}\boldsymbol{f} \in \mathbb{R}^{N\times N}, \tag{4.5}$$

provided that the integral converges. The spread of the posterior distribution describes the remaining uncertainty of the unknown parameter. A typical illustration for the spread is to calculate probability intervals from the posterior covariance estimator.

If the true state of the unknown parameter $\boldsymbol{f}$ is given a non-zero prior probability, as the sample size increases, the posterior distribution is asymptotically independent of the prior distribution and the maximum a posteriori estimate converges to the well-known *maximum likelihood* estimate

$$\boldsymbol{f}_{\mathrm{ML}} = \underset{\boldsymbol{f}\in\mathbb{R}^N}{\arg\max}\ p(\mathbf{m}|\boldsymbol{f}). \tag{4.6}$$

Then again, if the measurements provide only little information on the parameter of interest, the posterior is dominated by the prior.

A connection to ill-posed inverse problems can be seen in situations where the maximum likelihood estimate is not identifiable, but when a prior distribution is included, a proper posterior distribution can be obtained. Especially in highly ill-posed problems, the selection of the prior can then be the most critical phase in the inference and should be done based on expert knowledge on the studied quantity. One of the advantages of the Bayesian approach for inverse problems is that the required stabilisation can be given in a very interpretative manner in terms of physical quantities and related uncertainties.

Before considering the specific linear forward model presented in Section 3.2 the Gaussian model is considered for the general variables $\boldsymbol{f}$ and $\mathbf{m}$. By assuming that $\boldsymbol{f}$ and $\mathbf{m}$ have a joint multivariate Gaussian distribution

$$\begin{bmatrix} \boldsymbol{f} \\ \mathbf{m} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \bar{\boldsymbol{f}} \\ \bar{\mathbf{m}} \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_{\boldsymbol{f}} & \boldsymbol{\Sigma}_{\boldsymbol{f}\boldsymbol{m}} \\ \boldsymbol{\Sigma}_{\boldsymbol{m}\boldsymbol{f}} & \boldsymbol{\Sigma}_{\boldsymbol{m}} \end{bmatrix} \right), \tag{4.7}$$

with Gaussian identities the conditional distribution for $\boldsymbol{f}$ given $\mathbf{m}$ can be written as

$$p(\boldsymbol{f}|\mathbf{m}) \propto \exp\left(-\frac{1}{2}(\boldsymbol{f} - \bar{\boldsymbol{f}}^{(1)}) \left(\boldsymbol{\Sigma}_{\boldsymbol{f}}^{(1)}\right)^{-1} (\boldsymbol{f} - \bar{\boldsymbol{f}}^{(1)})^{\mathrm{T}} \right). \tag{4.8}$$

In the Gaussian system the CM and MAP estimators are the same. With Gaussian identities the MAP and posterior covariance estimators can be written as

$$\bar{\boldsymbol{f}}^{(1)} = \boldsymbol{f}_{\mathrm{MAP}} = \boldsymbol{f}_{\mathrm{CM}} = \bar{\boldsymbol{f}} + \boldsymbol{\Sigma}_{\boldsymbol{fm}}\boldsymbol{\Sigma}_{\boldsymbol{m}}^{-1}\left(\mathbf{m} - \bar{\mathbf{m}}\right) \tag{4.9}$$

and

$$\boldsymbol{\Sigma}_{\boldsymbol{f}}^{(1)} = \mathrm{cov}(\boldsymbol{f}|\mathbf{m}) = \boldsymbol{\Sigma}_{\boldsymbol{f}} - \boldsymbol{\Sigma}_{\boldsymbol{fm}}\boldsymbol{\Sigma}_{\boldsymbol{m}}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{mf}}. \tag{4.10}$$

4.2 Gaussian priors for linear inverse problems

For the likelihood function, the nature of measurements and central limit theorem often justifies the use of Gaussian normal distribution. Then again, the assumption of Gaussian prior distribution is not always the most realistic choice. A downside is that the distribution cannot be easily truncated to consist only of non-negative values. The benefit of the Gaussian prior is that it is a conjugate prior for the Gaussian likelihood, resulting in a Gaussian posterior distribution with the closed form estimators (4.9); (4.10). Hence often, as is the case here, a Gaussian prior distribution is assumed.

The discretisation of the linear forward model (3.2) was briefly discussed in Section 3.2.1. However, the discretisation can be performed at different phases of the solution. Following Tarantola (1987), when modelling an inverse problem, one should first consider whether it is easier to imagine the forward problem acting on a sequence of parameters or on a field. In many geophysical applications, such as ionospheric imaging, it is indeed a natural way to conceptualise the unknown and the prior distribution as a continuous field. In this section Gaussian random fields (Tarantola, 1987; Christakos, 2005; Rasmussen and Williams, 2006) will be utilised.

Definition 4.1. Given a probability space $(\Omega, \mathcal{F}, P)$, a real-valued d-dimensional *spatial random field* (RF) $f(z) := f(z, \omega)$, $z \in \mathbb{R}^d, \omega \in \Omega$ is a family of random variables $\{f(z_1), f(z_2), \dots\}$ at points $z_1, z_2, \dots$, where each random variable is real valued and defined on $(\Omega, \mathcal{F}, P)$.

A spatial RF is a generalisation of a stochastic process. Where a stochastic process is often seen as indexed by points in time, a spatial RF is indexed by d-dimensional Euclidean space, where d is typically two or three.

Definition 4.2. A multivariate *Gaussian random field* (GRF) is a spatial random field where any finite number of random variables have a joint Gaussian distribution.

GRF is then completely specified by its mean and covariance functions and denoted here as

$$f(z) \sim \mathcal{GRF}(\bar{f}(z), K(z, z')), \tag{4.11}$$

where

$$\begin{aligned} \bar{f}(z) =& \mathbb{E}[f(z)] \\ K(z,z') =& \mathbb{E}\left[(f(z)-\bar{f}(z))(f(z')-\bar{f}(z'))^{\mathrm{T}}\right]. \end{aligned} \tag{4.12}$$

The GRF (4.11) is used here as the prior for obtaining all of the following posterior estimators in this chapter. For a finite set of points $\mathbf{z} \in \mathbb{R}^{N\times d}$, as denoted in Equation (2.5), a GRF is simply a multivariate Gaussian normal distribution

$$\boldsymbol{f} \sim \mathcal{N}(\bar{\boldsymbol{f}}, \boldsymbol{\Sigma}_{\boldsymbol{f}}), \tag{4.13}$$

where $\bar{\boldsymbol{f}} \in \mathbb{R}^N$ and $\boldsymbol{\Sigma}_{\boldsymbol{f}} = K(\mathbf{z},\mathbf{z}) \in \mathbb{R}^{N\times N}$.

4.2.1 Continuous Gaussian random field prior

GRFs are closed under linear operations. Hence, when considering a linear problem (3.2) with GRF prior (4.11), the covariance and cross-covariances can be written as

$$\begin{aligned} \mathbb{E}\left[(\mathbf{m}-\bar{\mathbf{m}})(\mathbf{m}-\bar{\mathbf{m}})^{\mathrm{T}}\right] =& \mathbb{E}\left[((A_z f(z)+\boldsymbol{\varepsilon}-A_z\bar{f}(z))(A_{z'}f(z')+\boldsymbol{\varepsilon}-A_{z'}\bar{f}(z'))^{\mathrm{T}}\right] \\ =& \mathbb{E}\left[(A_z(f(z)-\bar{f}(z))(A_{z'}(f(z')-\bar{f}(z')))^{\mathrm{T}}\right] + \mathbb{E}\left[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^{\mathrm{T}}\right] \\ =& A_z K(z,z') A_{z'}^{\mathrm{T}} + \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}} = \mathbf{K}_{AfAf} + \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}, \end{aligned} \tag{4.14}$$

$$\begin{aligned} &\mathbb{E}\left[(\mathbf{m}-\bar{\mathbf{m}})(f(z')-\bar{f}(z'))^{\mathrm{T}}\right] = \mathbb{E}\left[(A_z f(z)-A_z\bar{f}(z))(f(z')-\bar{f}(z'))^{\mathrm{T}}\right] = A_z K(z,z') \\ &\mathbb{E}\left[(f(z)-\bar{f}(z))(\mathbf{m}-\bar{\mathbf{m}})^{\mathrm{T}}\right] = \mathbb{E}\left[(f(z)-\bar{f}(z))(A_{z'}f(z')-A_{z'}\bar{f}(z'))^{\mathrm{T}}\right] = K(z,z')A_{z'}^{\mathrm{T}}, \end{aligned} \tag{4.15}$$

where $\bar{\mathbf{m}} = A_z\bar{f}(z)$, $f \perp \boldsymbol{\varepsilon}$ and the subscript i in operator A_i indicates the parameter in the covariance function that it acts upon. The transpose is defined as $(A_z K(z,z'))^{\mathrm{T}} = K(z,z')A_{z'}^{\mathrm{T}}$.

The interest can now be in arbitrary locations $\mathbf{z} \in \mathbb{R}^{N^*\times d}$, which does not need to be the full lattice and thus $N^* \leq N$. Then, given the measurements $\mathbf{m}$ and a GRF prior (4.11) with mean and covariance as above (4.12), the covariance is a matrix

$$K(\mathbf{z},\mathbf{z}) = \mathbf{K}_{ff} \quad \in \mathbb{R}^{N^*\times N^*}$$

and the covariance matrices between the measurements and the unknowns

$$A_z K(z,\mathbf{z}) = \mathbf{K}_{Aff} = (\mathbf{K}_{fAf})^{\mathrm{T}} = (K(\mathbf{z},z)A_z^{\mathrm{T}})^{\mathrm{T}} \in \mathbb{R}^{M\times N^*}.$$

The posterior distribution for the unknown field at locations $\mathbf{z}$ is then given with Equation (4.8), where

$$\bar{\boldsymbol{f}}^{(1)} = \bar{f}(\mathbf{z}) + \mathbf{K}_{fAf}\left(\mathbf{K}_{AfAf}+\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}\right)^{-1}(\mathbf{m}-\bar{\mathbf{m}}) \ \in \mathbb{R}^{N^*} \tag{4.16}$$

$$\boldsymbol{\Sigma}_{\boldsymbol{f}}^{(1)} = \mathbf{K}_{ff} - \mathbf{K}_{fAf} \left(\mathbf{K}_{AfAf} + \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}\right)^{-1} \mathbf{K}_{Aff} \ \in \mathbb{R}^{N^* \times N^*}. \tag{4.17}$$

If the linear transformations with A_z in Equations (4.14) and (4.15) can be solved analytically without discretising f, the discretisation of the covariance kernel leading to a covariance matrix for the complete discretised domain can be avoided in the MAP estimator. In literature this approach is often referred to as *Kriging* or *Gaussian process* (Rasmussen and Williams, 2006). Especially in these approaches a parameterised covariance kernel is chosen and the parameters are estimated from the data. Gaussian processes with linear operations are discussed by Särkkä (2011) and Minkwitz et al. (2015).

4.2.2 Discrete multivariate Gaussian prior

Here the Bayesian approach is applied to a discretised linear system of Equation (3.3) on a lattice $\mathbf{z} \in \mathbb{R}^{N \times d}$, such as given in Equation (2.5). Hence, the multivariate normal distribution prior given in Equation (4.13) is used. The covariances and cross-covariances between the variables are then

$$\begin{aligned} \mathbb{E}\left[(\mathbf{m} - \bar{\mathbf{m}})(\mathbf{m} - \bar{\mathbf{m}})^{\mathrm{T}}\right] =& \mathbb{E}\left[(\mathbf{A}\boldsymbol{f} + \boldsymbol{\varepsilon} - \mathbf{A}\bar{\boldsymbol{f}})(\mathbf{A}\boldsymbol{f} + \boldsymbol{\varepsilon} - \mathbf{A}\bar{\boldsymbol{f}})^{\mathrm{T}}\right] \\ =& \mathbb{E}\left[(\mathbf{A}\boldsymbol{f} - \mathbf{A}\bar{\boldsymbol{f}})(\mathbf{A}\boldsymbol{f} - \mathbf{A}\bar{\boldsymbol{f}})^{\mathrm{T}}\right] + \mathbb{E}\left[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^{\mathrm{T}}\right] \\ =& \mathbf{A}\boldsymbol{\Sigma}_{\boldsymbol{f}}\mathbf{A}^{\mathrm{T}} + \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}} \ \in \mathbb{R}^{M \times M} \end{aligned} \tag{4.18}$$

and

$$\begin{aligned} \mathbb{E}\left[(\mathbf{m} - \bar{\mathbf{m}})(\boldsymbol{f} - \bar{\boldsymbol{f}})^{\mathrm{T}}\right] = \mathbb{E}\left[(\mathbf{A}\boldsymbol{f} - \mathbf{A}\bar{\boldsymbol{f}})(\boldsymbol{f} - \bar{\boldsymbol{f}})^{\mathrm{T}}\right] =& \mathbf{A}\boldsymbol{\Sigma}_{\boldsymbol{f}} \ \in \mathbb{R}^{M \times N} \\ \mathbb{E}\left[(\boldsymbol{f} - \bar{\boldsymbol{f}})(\mathbf{m} - \bar{\mathbf{m}})^{\mathrm{T}}\right] = \mathbb{E}\left[(\boldsymbol{f} - \bar{\boldsymbol{f}})(\mathbf{A}\boldsymbol{f} - \mathbf{A}\bar{\boldsymbol{f}})^{\mathrm{T}}\right] =& \boldsymbol{\Sigma}_{\boldsymbol{f}}\mathbf{A}^{\mathrm{T}} \ \in \mathbb{R}^{N \times M}, \end{aligned} \tag{4.19}$$

where $\bar{\mathbf{m}} = \mathbf{A}\bar{\boldsymbol{f}}$. The posterior distribution is then again of the form given in Equation (4.8), with

$$\bar{\boldsymbol{f}}^{(1)} = \bar{\boldsymbol{f}} + \boldsymbol{\Sigma}_{\boldsymbol{f}}\mathbf{A}^{\mathrm{T}} \left(\mathbf{A}\boldsymbol{\Sigma}_{\boldsymbol{f}}\mathbf{A}^{\mathrm{T}} + \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}\right)^{-1} (\mathbf{m} - \bar{\mathbf{m}}) \ \in \mathbb{R}^{N} \tag{4.20}$$

$$\boldsymbol{\Sigma}_{\boldsymbol{f}}^{(1)} = \boldsymbol{\Sigma}_{\boldsymbol{f}} - \boldsymbol{\Sigma}_{\boldsymbol{f}}\mathbf{A}^{\mathrm{T}} \left(\mathbf{A}\boldsymbol{\Sigma}_{\boldsymbol{f}}\mathbf{A}^{\mathrm{T}} + \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}\right)^{-1} \mathbf{A}\boldsymbol{\Sigma}_{\boldsymbol{f}} \ \in \mathbb{R}^{N \times N}. \tag{4.21}$$

Although it is now necessary to form the $N \times N$ prior covariance matrix, the matrix inversions in estimators (4.20) and (4.21) take place in the $M \times M$ measurement space. Hence, if $M \leq N$, the numerical computation is generally easier than in the following model space solution. However, with large N, the prior covariance matrix can become excessively large even for numerical storage.

4.2.3 Model space solution

For the Gaussian linear case, the MAP estimator (4.9) and posterior covariance (4.10) can be derived with well-known Gaussian identities from their joint distribution (4.7). By deriving the posterior distribution directly from the Bayes' formula (4.2) for the linear model (3.3), the quadratic form can also be arranged to provide the following equivalent forms for the estimators

$$\bar{\boldsymbol{f}}^{(1)} = \boldsymbol{\Sigma}_{\boldsymbol{f}}^{(1)} \left(\mathbf{A}^{\mathrm{T}} \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}^{-1} \mathbf{m} + \boldsymbol{\Sigma}_{\boldsymbol{f}}^{-1} \bar{\boldsymbol{f}} \right) \in \mathbb{R}^N \tag{4.22}$$

$$\boldsymbol{\Sigma}_{\boldsymbol{f}}^{(1)} = \left(\mathbf{A}^{\mathrm{T}} \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}^{-1} \mathbf{A} + \boldsymbol{\Sigma}_{\boldsymbol{f}}^{-1} \right)^{-1} \in \mathbb{R}^{N \times N}. \tag{4.23}$$

The above estimators can also be derived from Equations (4.20) and (4.21) with the matrix inversion lemma also known as Woodbury matrix identity or Sherman-Morrison-Woodbury formula (Golub and Van Loan, 2013). Since it is now necessary to invert two $N \times N$ matrices, this would generally be the preferred approach only in situations where $N \ll M$. However, if the prior information can be given directly as inverse covariance the situation can change, as it is demonstrated in the following section.

4.3 Gaussian Markov random field (GMRF) priors

The sparsity of a matrix signifies the large proportion of strict zeros in matrix elements. *Sparse linear system* then refers to a linear system that is so large and sparse that it is beneficial to rethink the standard factorising methods of two-dimensional arrays (Golub and Van Loan, 2013). For a *sparse matrix* it requires significantly less memory to index the non-zero matrix elements as vectors and to use operations designed for such systems.

When solving the posterior estimates given in previous sections, the measurement error covariance $\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}$ is typically assumed as a diagonal matrix. If the theory matrix $\mathbf{A}$ is sparse, as is the case here (2.4), the main concern is the prior covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{f}}$. A proper prior distribution with a non-diagonal covariance structure results in a dense $N \times N$ covariance matrix.

A Gaussian Markov random field (GMRF) is a multivariate Gaussian distribution Satisfying the Markov property. In GMRF, the Markov property indicates that an element conditioned with its neigbouring elements is independent of the rest of the elements in the field. The independence between elements is equivalent to the precision between the elements being zero. Typically GMRFs are used in a situations where the neighbourhood does not include the complete field, hence the precision matrix is characteristically a sparse matrix. A comprehensive introduction to GMRF is provided by Rue and Held (2005).

Definition 4.3. Neighbourhood N_i to $\boldsymbol{f}_i$ is the set $\{\boldsymbol{f}_j, j \in \mathrm{N}_i \mid \|\mathbf{z}_i - \mathbf{z}_j\| \leq r, j \neq i\}$, where radius $r > 0$.

Definition 4.4. A random vector $\boldsymbol{f} \in \mathbb{R}^N$ is called GMRF with respect to neighbourhood N_i with mean $\bar{\boldsymbol{f}}$ and precision matrix $\mathbf{Q}_{\boldsymbol{f}} > 0$ if and only if its density has the form

$$\boldsymbol{f} \sim (2\pi)^{-n/2} |\mathbf{Q}_{\boldsymbol{f}}|^{1/2} \exp\left(-\frac{1}{2} (\boldsymbol{f} - \bar{\boldsymbol{f}})^{\mathrm{T}} \mathbf{Q}_{\boldsymbol{f}} (\boldsymbol{f} - \bar{\boldsymbol{f}}) \right)$$

and

$$[\mathbf{Q}_{\boldsymbol{f}}]_{i,j} \neq 0 \Longleftrightarrow j \in \mathrm{N}_i \quad \forall \quad i \neq j.$$

With GMRF prior, the model space estimators in Equations (4.22) and (4.23) can then be written by replacing $\mathbf{Q}_{\boldsymbol{f}} = \boldsymbol{\Sigma}_{\boldsymbol{f}}^{-1} \in \mathbb{R}^{N \times N}$, resulting in

$$\bar{\boldsymbol{f}}^{(1)} = \boldsymbol{\Sigma}_{\boldsymbol{f}}^{(1)} \left(\mathbf{A}^{\mathrm{T}} \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}^{-1} \mathbf{m} + \mathbf{Q}_{\boldsymbol{f}} \bar{\boldsymbol{f}} \right) \in \mathbb{R}^N \tag{4.24}$$

$$\boldsymbol{\Sigma}_{\boldsymbol{f}}^{(1)} = \left(\mathbf{A}^{\mathrm{T}} \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}^{-1} \mathbf{A} + \mathbf{Q}_{\boldsymbol{f}} \right)^{-1} \in \mathbb{R}^{N \times N}. \tag{4.25}$$

It is not required that the precision matrix $\mathbf{Q}_{\boldsymbol{f}}$ is invertible. A GMRF with a symmetric positive semidefinite precision matrix is called *intrinsic* GMRF, which results in improper prior density for $\boldsymbol{f}$. A useful intrinsic GMRF, and an improper prior distribution, can be constructed by selecting $\mathbf{Q}_{\boldsymbol{f}} = \beta \mathbf{L}^{\mathrm{T}} \mathbf{L}$, where $\mathbf{L}$ is a difference matrix and β a weight parameter. Such a difference prior promotes smoothness, but is invariant to an addition of a constant. If no boundary conditions are added, $\mathrm{Ker}(\mathbf{L}) \neq \{0\}$. Then again, if $\mathrm{Ker}(\mathbf{A}) \cap \mathrm{Ker}(\mathbf{L}) = \{0\}$, the resulting Gaussian posterior distribution defines a proper probability density, with mean and covariance given in Equations (4.24) and (4.25) above. Further, by assuming $\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}} = \sigma^2 \mathbf{I}$ and denoting $\alpha := \sigma^2 \beta$, the resulting MAP estimator (4.24) is the generalised Tikhonov regularised solution given in Equation (3.18).

4.3.1 Correlation priors

The strength of a proper Gaussian prior is in the covariance where the provided information is easily interpretable in a probabilistic and physical sense. The downside is that it results in a full covariance matrix, making the storage and computation problematic when the number of unknowns is large.

GMRF priors are convenient when the ill-posed problem can be stabilised with moderate smoothing. In these cases the use of an intrinsic GMRF is straightforward, the interpretation in a mechanical sense is clear and the sparsity of a difference matrix $\mathbf{L}$ allows computations for much higher dimensional problems than working with a full prior covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{f}}$. Even so, if the problem is severely ill-conditioned and requires stronger stabilisation, the implementation of more strict constraints and boundary conditions in the precision matrix can get complicated, the effect of the possibly overlapping constraints unpredictable, and the physical interpretation difficult.

If sufficient boundary conditions are included into a difference matrix $\mathbf{L}$, the precision $\mathbf{L}^{\mathrm{T}}\mathbf{L} = \mathbf{Q}_{\boldsymbol{f}}$ becomes invertible and $(\mathbf{L}^{\mathrm{T}}\mathbf{L})^{-1} = \mathbf{Q}_{\boldsymbol{f}}^{-1} = \boldsymbol{\Sigma}_{\boldsymbol{f}}$. However, for efficient computation it would be profitable to work with the sparse $\mathbf{L}$ matrix while knowing the covariance structure in $\boldsymbol{\Sigma}_{\boldsymbol{f}}$ without solving it.

So-called *correlation priors* were introduced in Roininen et al. (2011, 2013). Similarly to the GRF case, the starting point for building a correlation prior is the selection of a continuous prior covariance function

$$K(z, z') = \mathrm{Cov}(z, z', \alpha, \boldsymbol{\ell}, \mathbf{c}), \tag{4.26}$$

where the covariance function between points z and z' is parametrised with variance scaling parameter α, correlation length parameters $\boldsymbol{\ell}$ and shape parameters $\mathbf{c}$. In the aforementioned articles it is shown that certain classes of covariance functions can be represented as solutions for systems of stochastic partial differential equations and that these systems can be approximated discretely with combinations of difference matrices. The matrices are formed with differences weighted with $\alpha, \boldsymbol{\ell}$ and $\mathbf{c}$ parameters inherited from the original covariance function, and with a discretisation length parameter h. The solution for the discrete system is a multivariate normal distribution with covariance matrix $(\mathbf{L}_{\mathrm{C}}^{\mathrm{T}}\mathbf{L}_{\mathrm{C}})^{-1}$, where $\mathbf{L}_{\mathrm{C}}$ contains the required weighted difference matrices in a stacked form. The resulting covariance is discretisation independent, which means that the obtained discrete covariance converges to continuous covariance at the discretisation limit. The approach then provides a scheme to write the posterior estimators (4.24) and (4.25) for a known prior covariance function (4.26) with a precision matrix $\mathbf{Q}_{\boldsymbol{f}} = \mathbf{L}_{\mathrm{C}}^{\mathrm{T}}\mathbf{L}_{\mathrm{C}}$.

Chapter 5

Spatiotemporal evolution

The solutions for the linear inverse problem presented in Chapters 3 and 4 have considered individual snapshots, where the unknown is assumed to be static in time and the measurements observed all at once. In this chapter, the state of the system is solved for sequential time steps. The first intuition would be to use any of the presented methods sequentially for different states. Another approach would be to solve the problem for one state and then use the obtained solution, depending on the method, as an initial guess or a prior for the next. Hence, the Bayesian approach provides a very natural way for dealing with temporally dynamic systems. For a more comprehensive treatment of Bayesian filtering and smoothing see e.g. Särkkä (2013).

5.1 Recursive linear estimation

The measurement model in Equation (3.3) is written for the time step l as

$$\mathbf{m}^{(l)} = \mathbf{A}^{(l)}\boldsymbol{f} + \boldsymbol{\varepsilon}^{(l)}, \tag{5.1}$$

where $\boldsymbol{\varepsilon}^{(l)} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}^{(l)})$ and $l = 0, 1, 2, \ldots$ It is assumed that the posterior estimates $\bar{\boldsymbol{f}}^{(l-1)}$ and $\boldsymbol{\Sigma}_{\boldsymbol{f}}^{(l-1)}$ for time $l-1$ are available. The posterior distribution can then be used as a prior for the estimates of the following state l, and the discrete measurement space estimators (4.20) and (4.21) are

$$\bar{\boldsymbol{f}}^{(l)} = \bar{\boldsymbol{f}}^{(l-1)} + \boldsymbol{\Sigma}_{\boldsymbol{f}}^{(l-1)}(\mathbf{A}^{(l)})^{\mathrm{T}} \left(\mathbf{A}^{(l)}\boldsymbol{\Sigma}_{\boldsymbol{f}}^{(l-1)}(\mathbf{A}^{(l)})^{\mathrm{T}} + \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}^{(l)}\right)^{-1} \left(\mathbf{m}^{(l)} - \mathbf{A}^{(l)}\bar{\boldsymbol{f}}^{(l-1)}\right) \tag{5.2}$$

$$\boldsymbol{\Sigma}_{\boldsymbol{f}}^{(l)} = \boldsymbol{\Sigma}_{\boldsymbol{f}}^{(l-1)} - \boldsymbol{\Sigma}_{\boldsymbol{f}}^{(l-1)}(\mathbf{A}^{(l)})^{\mathrm{T}} \left(\mathbf{A}^{(l)}\boldsymbol{\Sigma}_{\boldsymbol{f}}^{(l-1)}(\mathbf{A}^{(l)})^{\mathrm{T}} + \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}^{(l)}\right)^{-1} \mathbf{A}^{(l)}\boldsymbol{\Sigma}_{\boldsymbol{f}}^{(l-1)}. \tag{5.3}$$

Often temporary variables, innovation $\mathbf{S}$ and Kalman gain $\mathbf{G}$, are used to write the above in the following steps

$$\begin{aligned}
\mathbf{S}^{(l)} =& \mathbf{A}^{(l)}\boldsymbol{\Sigma}_{\boldsymbol{f}}^{(l-1)}(\mathbf{A}^{(l)})^{\mathrm{T}} + \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}^{(l)} \\
\mathbf{G}^{(l)} =& \boldsymbol{\Sigma}_{\boldsymbol{f}}^{(l-1)}(\mathbf{A}^{(l)})^{\mathrm{T}}(\mathbf{S}^{(l)})^{-1} \\
\bar{\boldsymbol{f}}^{(l)} =& \bar{\boldsymbol{f}}^{(l-1)} + \mathbf{G}^{(l)}\left(\mathbf{m}^{(l)} - \mathbf{A}^{(l)}\bar{\boldsymbol{f}}^{(l-1)}\right) \\
\boldsymbol{\Sigma}_{\boldsymbol{f}}^{(l)} =& \boldsymbol{\Sigma}_{\boldsymbol{f}}^{(l-1)} - \mathbf{G}^{(l)}\mathbf{S}^{(l)}(\mathbf{G}^{(l)})^{\mathrm{T}}.
\end{aligned} \tag{5.4}$$

The recursive algorithm allows online updating when new information becomes available. However, the algorithm is based on an assumption that the unknown $\boldsymbol{f}$ is constant and each update is accumulating information from the same state. Hence, the l^{th} solution could be obtained also by using all measurements at once.

5.2 Kalman filtering

The more general *Bayesian filtering* is restricted here to a linear Gaussian case, when the algorithm is known more famously as *Kalman filter* (Kalman, 1960). Here the unknown $\boldsymbol{f}^{(l)}$ is assumed to evolve in time with states $l = 1, 2, \ldots$ and that its dynamics can be modelled with a probabilistic state space model. In practise, in comparison with recursive linear filtering above, this means mainly the addition of *linear dynamic model*

$$\boldsymbol{f}^{(l)} = \mathbf{H}^{(l-1)}\boldsymbol{f}^{(l-1)} + \boldsymbol{e}^{(l-1)}, \tag{5.5}$$

where the stochastic dynamics are modelled with a transition matrix $\mathbf{H}^{(l-1)}$ acting on the previous state of the system and with process noise $\boldsymbol{e}^{(l-1)} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{e}}^{(l-1)})$. The measurement model of Equation (5.1) is then written for each state as

$$\mathbf{m}^{(l)} = \mathbf{A}^{(l)}\boldsymbol{f}^{(l)} + \boldsymbol{\varepsilon}^{(l)}. \tag{5.6}$$

Hence, intuitively in Bayesian inference for a dynamical system, the best guess for the present state is not given by the previous posterior distribution, but by their mappings with the transition matrix, resulting in the predictive distribution

$$\boldsymbol{f}^{(l)}|\mathbf{m}^{(1:l-1)} \sim \mathcal{N}(\hat{\boldsymbol{f}}^{(l)}, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{f}}^{(l)}), \tag{5.7}$$

where the mean and covariance are defined below in Equation (5.8).

Now the solution for state l is the posterior distribution where the predictive distribution is used as the prior and the likelihood constructed from the current measurement model and measurements. Traditionally, the estimators are again separated in steps with the new variables:

Prediction step

$$\begin{aligned}\hat{\boldsymbol{f}}^{(l)} =& \mathbf{H}^{(l-1)}\bar{\boldsymbol{f}}^{(l-1)} \\ \hat{\boldsymbol{\Sigma}}_{\boldsymbol{f}}^{(l)} =& \mathbf{H}^{(l-1)}\boldsymbol{\Sigma}_{\boldsymbol{f}}^{(l-1)}(\mathbf{H}^{(l-1)})^{\mathrm{T}} + \boldsymbol{\Sigma}_{\boldsymbol{e}}^{(l-1)}\end{aligned} \tag{5.8}$$

Update step

$$\begin{aligned}\mathbf{v}^{(l)} =& \mathbf{m}^{(l)} - \mathbf{A}^{(l)}\hat{\boldsymbol{f}}^{(l)} \\ \mathbf{S}^{(l)} =& \mathbf{A}^{(l)}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{f}}^{(l)}(\mathbf{A}^{(l)})^{\mathrm{T}} + \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}^{(l)} \\ \mathbf{G}^{(l)} =& \hat{\boldsymbol{\Sigma}}_{\boldsymbol{f}}^{(l)}(\mathbf{A}^{(l)})^{\mathrm{T}}(\mathbf{S}^{(l)})^{-1} \\ \bar{\boldsymbol{f}}^{(l)} =& \hat{\boldsymbol{f}}^{(l)} + \mathbf{G}^{(l)}\mathbf{v}^{(l)} \\ \boldsymbol{\Sigma}_{\boldsymbol{f}}^{(l)} =& \hat{\boldsymbol{\Sigma}}_{\boldsymbol{f}}^{(l)} - \mathbf{G}^{(l)}\mathbf{S}^{(l)}(\mathbf{G}^{(l)})^{\mathrm{T}}.\end{aligned} \tag{5.9}$$

5.3 Kalman smoothing

In Kalman filtering the earlier and current measurements are used to compute the best possible estimate for the current state of the system. However, when applications are not run online, a complete data set $\mathbf{m}^{(l)}$ for each $l = 1, \ldots, T$ might be available. If the interest is not in the last state, but in the whole process, with Bayesian smoothing it is possible to take into account also the future states of the system, while evaluating state l.

The Gaussian linear version of the *Bayesian smoother* is known also as *Rauch-Tung-Striebel smoother* and *Kalman smoother*. In its standard form the smoother algorithm is divided in forward and backward passes. In forward pass, the data is filtered with steps (5.8) and (5.9) and the results are saved. The filtering results are then used in the *backward pass* with steps

$$\begin{aligned}\mathbf{C}^{(l)} =& \boldsymbol{\Sigma}_{\boldsymbol{f}}^{(l)}(\mathbf{H}^{(l)})^{\mathrm{T}}(\hat{\boldsymbol{\Sigma}}_{\boldsymbol{f}}^{(l+1)})^{-1} \\ \tilde{\boldsymbol{f}}^{(l)} =& \bar{\boldsymbol{f}}^{(l)} + \mathbf{C}^{(l)}(\tilde{\boldsymbol{f}}^{(l+1)} - \hat{\boldsymbol{f}}^{(l+1)}) \\ \tilde{\boldsymbol{\Sigma}}_{\boldsymbol{f}}^{(l)} =& \boldsymbol{\Sigma}_{\boldsymbol{f}}^{(l)} + \mathbf{C}^{(l)}(\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{f}}^{(l+1)} - \hat{\boldsymbol{\Sigma}}_{\boldsymbol{f}}^{l+1})(\mathbf{C}^{(l)})^{\mathrm{T}},\end{aligned} \tag{5.10}$$

where $\hat{\boldsymbol{f}}^{(l+1)}$, $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{f}}^{(l+1)}$, $\bar{\boldsymbol{f}}^{(l)}$ and $\boldsymbol{\Sigma}_{\boldsymbol{f}}^{(l)}$ are the predicted and filtered solutions from the forward pass. The backward pass is started from the state T, with $\tilde{\boldsymbol{f}}^{(T)} = \bar{\boldsymbol{f}}^{(T)}$ and $\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{f}}^{(T)} = \boldsymbol{\Sigma}_{\boldsymbol{f}}^{(T)}$.

5.4 Ensemble Kalman filter (EnKF)

When increasing the spatial resolution in a three-dimensional model the number of unknown variables N increases rapidly and the $N \times N$ dimensional covariance matrices become

infeasible to handle in the above filtering and smoothing algorithms. In *ensemble Kalman filter* (EnKF) (Evensen, 1994, 2003, 2009) the maintenance of large covariance matrices is eased by not solving the updated posterior covariance in Equation (5.9) directly. Instead, samples from each posterior distribution are simulated and the covariance information is carried within the sample. The EnKF belongs to a wider category of particle filters and it was developed mainly for nonlinear problems. Here the main idea is presented in a linear setting.

First an initial prior *ensemble*

$$\mathbf{F}^{(0)} = [\boldsymbol{f}_1^{(0)}, \dots \boldsymbol{f}_{N_{\text{ens}}}^{(0)}] \ \in \mathbb{R}^{N \times N_{\text{ens}}} \tag{5.11}$$

is generated, where the number of ensemble members $N_{\text{ens}} \ll N$.

At step l an *ensemble of observations*

$$\mathbf{M}^{(l)} = [\mathbf{m}_1^{(l)}, \dots \mathbf{m}_{N_{\text{ens}}}^{(l)}] \ \in \mathbb{R}^{M \times N_{\text{ens}}}, \tag{5.12}$$

is simulated as $\mathbf{m}_i^{(l)} = \mathbf{m}^{(l)} + \boldsymbol{\varepsilon}_i$, where $\boldsymbol{\varepsilon}_i \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}})$. The *predicted ensemble* is then obtained as

EnKF prediction step

$$\hat{\mathbf{F}}^{(l)} = \mathbf{H}^{(l-1)}\mathbf{F}^{(l-1)} \ \in \mathbb{R}^{N \times N_{\text{ens}}}. \tag{5.13}$$

The predicted ensemble mean is then

$$\hat{\boldsymbol{f}}_{\text{ens}}^{(l)} = \frac{1}{N_{\text{ens}}} \hat{\mathbf{F}}^{(l)} \mathbf{1}_{N_{\text{ens}} \times 1} \ \in \mathbb{R}^N \tag{5.14}$$

and the corresponding sample covariance

$$\begin{aligned} \hat{\boldsymbol{\Sigma}}_{\boldsymbol{f}_{\text{ens}}}^{(l)} =& \frac{1}{N_{\text{ens}} - 1} \left(\hat{\mathbf{F}}^{(l)} - \hat{\boldsymbol{f}}_{\text{ens}}^{(l)} \mathbf{1}_{1 \times N_{\text{ens}}} \right) \left(\hat{\mathbf{F}}^{(l)} - \hat{\boldsymbol{f}}_{\text{ens}}^{(l)} \mathbf{1}_{1 \times N_{\text{ens}}} \right)^{\mathrm{T}} \\ =& \boldsymbol{\Gamma}_{\boldsymbol{f}_{\text{ens}}}^{(l)} (\boldsymbol{\Gamma}_{\boldsymbol{f}_{\text{ens}}}^{(l)})^{\mathrm{T}}, \end{aligned} \tag{5.15}$$

where $\boldsymbol{\Gamma}_{\boldsymbol{f}_{\text{ens}}}^{(l)} \in \mathbb{R}^{N \times N_{\text{ens}}}$. The *posterior ensemble* matrix $\mathbf{F}^{(l)}$ is then obtained with EnKF update step.

EnKF update step

$$\begin{aligned} \mathbf{V}^{(l)} =& \mathbf{M}^{(l)} - \mathbf{A}^{(l)} \hat{\mathbf{F}}^{(l)} \\ \mathbf{S}^{(l)} =& \left(\mathbf{A}^{(l)} \boldsymbol{\Gamma}_{\boldsymbol{f}_{\text{ens}}}^{(l)} \right) \left((\boldsymbol{\Gamma}_{\boldsymbol{f}_{\text{ens}}}^{(l)})^{\mathrm{T}} (\mathbf{A}^{(l)})^{\mathrm{T}} \right) + \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}^{(l)} \\ \mathbf{G}^{(l)} =& \boldsymbol{\Gamma}_{\boldsymbol{f}_{\text{ens}}}^{(l)} \left((\boldsymbol{\Gamma}_{\boldsymbol{f}_{\text{ens}}}^{(l)})^{\mathrm{T}} (\mathbf{A}^{(l)})^{\mathrm{T}} \right) (\mathbf{S}^{(l)})^{-1} \\ \mathbf{F}^{(l)} =& \hat{\mathbf{F}}^{(l)} + \mathbf{G}^{(l)} \mathbf{V}^{(l)}, \end{aligned} \tag{5.16}$$

wherein the order of operations can be selected such that no $N \times N$ matrices are formed at any stage. The ensemble carries correct error statistics for deriving the ensemble mean and covariance, as with increasing ensemble size, the linear EnKF solution converges exactly to the Kalman filter solution (Evensen, 2009).

Chapter 6

Ionospheric measurements

In this chapter the physical background of the mostly used measurements in ionospheric imaging are presented. The chapter begins with an introduction to general electromagnetic wave propagation as it provides the basis for the modelling of radio measurements of satellite-transmitted signals and ionosonde measurements. As the ground-based measurements of *very-high frequency* (VHF, 30–300 MHz) and *ultra-high frequency* (UHF, 300–3000 MHz) satellite signals are the most important data component and the understanding of the related measurement errors and biases is essential, the background of these measurements will be examined in more detail. Ionosonde, incoherent scatter radar and satellite in situ measurements provide precise information on ionospheric electron density; however, as the spatial coverage of these measurements is local-scale, they can be used only as additional or validation data and are reviewed here more briefly.

6.1 Electromagnetic wave propagation

An electromagnetic wave propagating along axis $z \in \mathbb{R}^+$ in temporally and spatially homogeneous medium can be represented as

$$\psi(t, z) = E_0 \cos(\omega t - kz) = E_0 \cos\left(\omega t - \frac{\omega}{c} n z\right), \tag{6.1}$$

where E_0 is the peak amplitude, t is the time, ω is the angular frequency, $k = \omega/v$ is the wavenumber and c and v are the velocities of an electromagnetic wave in vacuum and in the medium correspondingly. Finally the refractive index n describes the velocity of the electromagnetic wave in thee medium. It is defined as

$$n = \frac{c}{v}. \tag{6.2}$$

6.1.1 Ionospheric refractive index

The complex *refractive index* of magnetised plasma containing free electrons is given by the Appleton-Lassen or the Appleton-Hartree formula (Budden, 1961)

$$n^2 = 1 - \frac{X}{1 - \mathrm{i}Z + \frac{\frac{1}{2}Y^2 \sin^2\theta}{(1-X-\mathrm{i}Z)} \pm \sqrt{\frac{\frac{1}{4}Y^4 \sin^4\theta}{(1-X-\mathrm{i}Z)^2} - Y^2\cos^2\theta}}, \tag{6.3}$$

where $\mathrm{i} = \sqrt{-1}$ is the imaginary number, $X = \frac{\omega_\mathrm{p}^2}{\omega^2}$, $Y = \frac{\omega_H}{\omega}$, $Z = \frac{\nu}{\omega}$, ω_p is the angular plasma frequency, ν is the electron collision frequency, θ is the angle between wave normal and inclination of magnetic field and ω_H is the electron gyrofrequency.

Plasma frequency

The angular *plasma frequency* can be written as

$$\omega_\mathrm{p} = \sqrt{\frac{N_e e^2}{\epsilon_0 m}}, \tag{6.4}$$

where N_e is the electron density, e is the electron charge, m is the electron mass and ϵ_0 is the permittivity of free space. Useful conversions between electron density and (temporal) plasma frequency $\mathrm{f_p} = \frac{\omega_\mathrm{p}}{2\pi}$ can then be obtained by inserting the natural constants to Equation (6.4) and solving

$$\begin{aligned} \mathrm{f_p} &\approx 8.98 \times \sqrt{N_e} \quad \text{(Hz)} \quad \text{and} \\ N_e &\approx 0.012 \times \mathrm{f_p^2} \quad \text{(1/m}^3\text{)}. \end{aligned} \tag{6.5}$$

Collision frequency

The complex part in the Appleton-Lassen formula (6.3) is related to absorption. The absorption results when charged particles that oscillate along with the electromagnetic wave collide with other, mainly neutral particles. The collisions then decrease the energy of the radiation. At mid latitudes, the effective collision frequency is less than 10^4 Hz (Fehmers, 1996). The maximum collision frequency measured in an example case in Tromsø 1991 is around 10^6 Hz at an altitude of 100 km, from where it decreases rapidly as the altitude increases (Brekke, 1997). As can be seen in the Appleton-Lassen formula (6.3), part Z containing collisions decreases with increasing signal frequency. When the frequency of the propagating electromagnetic wave is greater than about 1 MHz the electron collisions can be neglected and Z can be approximated with zero (Budden, 1961; Davies, 1965).

Gyrofrequency and magnetic field

Gyrofrequency can be written as

$$\omega_H = \frac{B_0|e|}{m}, \tag{6.6}$$

where B_0 is the magnetic field strength. It is the angular frequency of a charged particle, here electron, circling around a uniform magnetic field in a direction perpendicular to the field. In a case of magnetised plasma, the plusminus sign in Appleton-Lassen formula (6.3) indicates how the propagating wave is split in two modes. The mode with the "+" sign is called *ordinary* and the mode with the "−" sign *extraordinary* component.

A general typical value for gyrofrequency is around 1 MHz (Budden, 1961) and 1.5 MHz (Parkinson et al., 1996). Also Y in Appleton-Lassen formula (6.3), that contains the gyrofrequency, decreases with increasing signal frequency.

6.1.2 Group refractive index

As can be seen in Appleton-Lassen formula (6.3), the ionospheric refraction depends on the frequency of the propagating signal, i.e. the ionosphere is a *dispersive* medium. When considering a modulated signal the velocities of the signal carrier phase and modulation envelope will differ due to the dispersion. Following Davies (1965, 1990), to demonstrate this effect, two electromagnetic waves propagating along one-dimensional axis $z \in \mathbb{R}^+$ are considered

$$\begin{aligned} \psi_1(t,z) =& E_0 \cos(\omega t - kz) \\ \psi_2(t,z) =& E_0 \cos((\omega + \Delta\omega)t - (k + \Delta k)z), \end{aligned} \tag{6.7}$$

where $\Delta\omega$ and Δk are the small differences in angular frequency and wavenumber between signals ψ_1 and ψ_2. An amplitude-modulated signal can be created by summation and then writing with trigonometric identities as

$$\begin{aligned} \psi_1(t,z) + \psi_2(t,z) =& 2E_0 \cos\left(\frac{\Delta\omega}{2}t - \frac{\Delta k}{2}z\right) \cos\left[\left(\omega + \frac{\Delta\omega}{2}\right)t - \left(k + \frac{\Delta k}{2}\right)z\right] \\ \approx& 2E_0 \cos\left(\frac{\Delta\omega}{2}t - \frac{\Delta k}{2}z\right) \cos(\omega t - kz), \end{aligned} \tag{6.8}$$

where the first cosine term represents the modulation envelope and the second the carrier phase. The modulation envelope then propagates with *group velocity*

$$v_\mathrm{g} = \lim_{\Delta k \to 0} \frac{\Delta\omega}{\Delta k} = \frac{\mathrm{d}\omega}{\mathrm{d}k}. \tag{6.9}$$

The *group refractive index* can then be given with the definition in Equation (6.2) as

$$n_\mathrm{g} = \frac{c}{v_\mathrm{g}} = c\frac{\mathrm{d}k}{\mathrm{d}\omega} = \frac{\mathrm{d}}{\mathrm{d}\omega}(ck) = \frac{\mathrm{d}}{\mathrm{d}\omega}\left(c\frac{\omega}{v}\right) = \frac{\mathrm{d}}{\mathrm{d}\omega}\left(\frac{c}{v}\omega\right) = \frac{\mathrm{d}}{\mathrm{d}\omega}(n(\omega)\omega) = n(\omega) + \omega\frac{\mathrm{d}n(\omega)}{\mathrm{d}\omega}. \tag{6.10}$$

6.1.3 Tropospheric refractive index

While the electromagnetic wave approaches Earth's surface, in neutral atmosphere, below the ionosphere, the number of free electrons decreases to zero and the ionospheric refractive index approaches one. However, due to dry gases and water vapour, the refractive index in the troposphere differs from the free space. The accumulated phase difference caused by the tropospheric refraction can be significant, but typically less than that of the ionosphere. The changes in the total contribution of the troposphere are also within $\pm 10\%$ even in longer time periods, whereas the ionosphere can have large rapid changes (Klobuchar, 1996). More detailed studies on tropospheric parameters are provided by Bernhardt et al. (2000); Rüeger (2002). According to Wells et al. (1986) the troposphere is nondispersive for frequencies below 30 GHz. Kaplan and Hegarty (2006) state that the limit is 15 GHz. Below these limits, the tropospheric refractive index is then independent of the frequency, and the group and phase velocities are equal. Hence, it is enough to denote the *tropospheric refractive index* here as

$$n_{\mathrm{tr}} = 1 + \Delta n_{\mathrm{tr}}, \tag{6.11}$$

that is the refractive index of vacuum perturbed with the tropspheric contribution Δn_{tr}.

6.2 Radio measurements of satellite transmissions

As of today, there exist several global navigation satellite systems (GNSS), such as GPS, GLONASS, GALILEO and BEIDOU that can be used for ionospheric observations. The different GNSSs operate in UHF frequencies ranging from GALILEO's lowest frequency of 1176.45 MHz to the highest GLONASS frequency of 1605.375 MHz. In ionospheric studies the most used satellite system has been GPS with the main frequencies of 1575.42 (L1) and 1227.60 MHz (L2). The satellite orbit altitudes used by different GNSS are for GLONASS 19,100 km, for GPS 20,180 km, for BeiDou 21,528 km and for GALILEO 23,222 km (Kaplan and Hegarty, 2006). Besides GNSS systems, low Earth orbit (LEO) beacon satellites have also been used frequently in atmospheric studies. LEO satellite beacons operate typically with dual or tri-band VHF and UHF frequencies of 150, 400 and 1067 MHz. LEO refers to orbital altitudes less than 1,500 km (Bernhardt et al., 2000; Yamamoto, 2008; Vierinen et al., 2014).

6.2.1 Refractive indices for VHF and UHF signals

The principle of ionospheric observations with radio measurements of satellite beacon signals is based on the connection between the frequency-dependent refractive index and electron density, given with the Appleton-Lassen formula (6.3). However, in its original form given above, the connection is nonlinear and complex. For an efficient measurement model a linear equation without imaginary part is sought for.

As the UHF and VHF frequencies are much greater than 1 MHz, the electron collisions can be neglected and Z can be approximated with zero (Budden, 1961; Davies, 1965). This removes the imaginary part from the Appleton-Lassen formula (6.3). A gyrofrequency of 1.5 MHz results in $Y = 0.01$ at 150 MHz and decreasing with increasing signal frequency. Hence, Y in the Equation (6.3) can also be approximated with zero.

When electron collisions and gyrofrequencies are both omitted, the Appleton-Lassen formula (6.3) simplifies to

$$n^2 = 1 - X = 1 - \frac{\omega_\mathrm{p}^2}{\omega^2}. \tag{6.12}$$

The relation is indeed more simple, however, still nonlinear. Therefore, as $\omega \gg \omega_\mathrm{p}$, the refractive index can be approximated with first order Taylor polynomial at $\frac{\omega_\mathrm{p}}{\omega} = 0$ with

$$n \approx 1 - \frac{1}{2}\left(\frac{\omega_\mathrm{p}}{\omega}\right)^2. \tag{6.13}$$

Inserting $\omega_\mathrm{p} = \sqrt{\frac{N_e e^2}{\epsilon_0 m_e}}$ (rad/s) then results in

$$n \approx 1 - \frac{N_e e^2}{2\epsilon_0 m_e \omega^2}. \tag{6.14}$$

The simplifying and linearising assumptions in Approximation (6.14) give rise to an error less than 1% with 150 MHz frequency, decreasing with higher frequency (Fehmers, 1996).

With Equation (6.10), the group refractive index for VHF and UHF signals can then be derived as

$$n_\mathrm{g} = n(\omega) + \omega\frac{\mathrm{d}n(\omega)}{\mathrm{d}\omega} \approx 1 - \frac{N_e e^2}{2\epsilon_0 m_e \omega^2} + 2\frac{N_e e^2}{2\epsilon_0 m_e \omega^2} = 1 + \frac{N_e e^2}{2\epsilon_0 m_e \omega^2}. \tag{6.15}$$

6.2.2 Wave propagation of VHF and UHF signals

Spatially inhomogeneous medium wave propagation in Equation (6.1) can be written as

$$\psi(t, L) = E_0 \cos\left(\omega t - \frac{\omega}{c}\int_0^L n(z)\mathrm{d}z\right), \tag{6.16}$$

where, for the sake of convention, the integral is defined from receiver at $z = 0$ to satellite at distance $z = L$.

It is here enough to concentrate on the integral part inside the cosine function (6.16) that is the effect of the medium for the propagating signal. For the carrier phase it is typically given in radians

$$\phi(L) = \frac{\omega}{c}\int_0^L n(z)\mathrm{d}z \tag{6.17}$$

and for the modulation envelope in metres

$$\rho(L) = \int_0^L n_{\mathrm{g}}(z)\mathrm{d}z. \tag{6.18}$$

When taking into account the nondispersive tropospheric contribution (6.11), the phase (6.17) can be written for the different intervals as

$$\phi(L) = \frac{\omega}{c}\left(\int_0^{L_0} n_{\mathrm{tr}}(z)\mathrm{d}z + \int_{L_0}^{L} n(z)\mathrm{d}z\right), \tag{6.19}$$

where the first integral is defined along the signal path from ground receiver to distance L_0 where the ionospheric refraction becomes significant, and the second integral from this altitude to the upper boundary of the ionosphere L. Inserting the refractive index (6.14) results in radians as

$$\begin{aligned}\phi(L) =& \frac{\omega}{c}\left[\int_0^{L_0}(1+\Delta n_{\mathrm{tr}})\mathrm{d}z + \int_{L_0}^{L}\left(1-\frac{N_e(z)e^2}{2\epsilon_0 m_e\omega^2}\right)\mathrm{d}z\right]\\ =& \frac{\omega}{c}\left[L + \int_0^{L_0}\Delta n_{\mathrm{tr}}\mathrm{d}z - \int_{L_0}^{L}\frac{N_e(z)e^2}{2\epsilon_0 m_e\omega^2}\mathrm{d}z\right]\\ =& \frac{\omega}{c}L + \frac{\omega}{c}T(L) - \frac{\alpha}{c\omega}TEC(L).\end{aligned} \tag{6.20}$$

Similarly for the group delay, combining Equations (6.18), (6.11) and (6.15) results in metres as

$$\rho(L) = L + T(L) + \frac{\alpha}{\omega^2}TEC(L). \tag{6.21}$$

In both Equations (6.20) and (6.21), L is the range between the transmitter and receiver, $T(L) = \int_0^{L_0}\Delta n_{\mathrm{tr}}\mathrm{d}z$ is the tropospheric contribution, $\alpha = \frac{e^2}{2\epsilon_0 m_e}$ a combination of constants and

$$TEC(L) = \int_{L_0}^{L} N_e(z)\mathrm{d}z \tag{6.22}$$

is the slant *total electron content* (TEC). The last term, which includes TEC, is positive when considering phase velocity and negative when group velocity is considered. In GNSS literature temporal frequency is often used instead of angular. The coefficient in the ionospheric part is then $\frac{\alpha}{\omega} = \frac{\alpha}{4\pi^2\mathrm{f}^2} \approx \frac{40.3}{\mathrm{f}^2}$.

6.2.3 Observables

The two main measurement types, made from satellite beacon signals are the *pseudorange* observable based on the group delay given in Equation (6.21) and the *carrier phase* observable based on the phase advancement given in Equation (6.20). For other ionospheric

effects on satellite signals, such as Doppler shift and Faraday rotation and their use as measurements see e.g. Klobuchar (1985, 1996).

In traditional use of satellite positioning, the main interest in an individual measurement is in the range between a user with an unknown position and a satellite with a known position. In such range measurement the ionosphere is a source of error. In ionospheric measurements the location of the receiver and hence the range to the satellite is known and the primary interest is in the unknown TEC.

The satellite orbital altitudes used by different GNSSs are around 20,000 km. In Equation (6.21), the contribution of a relatively high TEC of 100 TECU in pseudorange extends from 15 m at 1605.375 MHz to 1,800 m at 150 MHz. The total contribution of troposphere to pseudorange in satellite beacon frequencies is approximately between 2.4–25 m (Kaplan and Hegarty, 2006). Hence, for the navigation, with the aid of atmospheric models the range estimation can be carried out with sufficient accuracy even from a single measurement. However, when the interest is in the ionosphere, it is evident from the numbers above that the situation is much worse.

Equations (6.20) and (6.21) describe an ideal measurement taking into account only the physical composition of the atmosphere. Unfortunately, in real life, the measurements also suffer from several other nuisances. Below, the most significant errors and biases for the TEC measurements are included in the models of both observables. The error sources omitted here include antenna-phase center variations, earth tides, ocean loading, and for phase measurement, the phase windup effect (Håkansson et al., 2017). Taking the additional errors and biases into account consolidates the intuition why individual measurements as such are useless for most of the applications considering TEC.

To overcome this problem, at least partially, measurements with two different frequencies ω_1 and ω_2 are used. As part of the errors and biases are dispersive and part nondispersive, i.e. coherent, the coherent errors can then be canceled by combining the measurements, leaving only the frequency-dependent part.

Pseudorange observable

The transmitted GNSS satellite signals are modulated with different pseudorandom noise codes depending on the satellite system and the frequency at issue. When received, the signal is aligned with a reference signal and the modulated signals are compared. Due to the pseudorandomness the maximum correlation between the received and the replica code is achieved only when the signals are aligned perfectly. The amount the replica code needs to be shifted for maximising the correlation provides the travelling time of the signal. When multiplied with velocity c, the range between the receiver and the satellites is obtained (Wells et al., 1986; Kaplan and Hegarty, 2006). In GNSS positioning terminology the measurement of group delay is known as pseudorange, as it includes several biases. By including the most significant bias terms in Equation (6.21), the *pseudorange* can be

written as

$$\begin{aligned}\rho_\omega(L,t) =& c\left(\tau_{rec} + \tau_{sat} + b_{\rho,rec,\omega,code} + b_{\rho,sat,\omega,code}\right) + M_{\rho,sat,rec,\omega,code}(t) \\ &+ L + T(L) + \frac{\alpha}{\omega^2} TEC(L) + \varepsilon_{\rho,sat,rec,\omega,code}(t),\end{aligned} \tag{6.23}$$

where the new parameters τ are the receiver and satellite clock errors, parameters b are the receiver and satellite hardware biases, M is the multipath error and ε is the measurement error due to thermal noise etc. The dependencies of different parameters to specific observations are given with subindex variables $\omega, \rho, rec, sat, code$ referring in corresponding order to frequency, observable type, receiver and satellite names and the measured code. For example, $b_{\rho,\mathrm{PRN02,L1,C/A}}$ is the satellite bias for pseudorange measurements that uses GPS satellite PRN02 and L1 frequency with coarse/acquisition code modulation.

A rule of thumb for the measurement precision is 1% of the period between two code epochs (Wells et al., 1986). For GPS codes, this results in a precision of 1 ns for the P-code and 10 ns for the C/A-code. Converted to TEC measurements the P-code precisions are then 1.9 TECU for L1 and 1.1 TECU for L2. For C/A-code the precision in L1 is 19 TECU.

Differential group delay

When pseudorange measurements with two different angular frequencies ω_1 and ω_2 are available, the coherent part consisting of the range (with possible errors), clock errors and tropospheric error is canceled out in subtraction, resulting in

$$\begin{aligned}\Delta\rho(L,t) =& \rho_{\omega_2}(L,t) - \rho_{\omega_1}(L,t) \\ =& \alpha\left(\frac{1}{\omega_2^2} - \frac{1}{\omega_1^2}\right) TEC(L) \\ &+ cb_{\rho,rec,\omega_2,} - cb_{\rho,rec,\omega_1,} + cb_{\rho,sat,\omega_2} - cb_{\rho,sat,\omega_1} \\ &+ M_{\omega_2,\rho,2}(t) - M_{\omega_1,\rho,1}(t) + \varepsilon_{\omega_2,\rho}(t) - \varepsilon_{\omega_1,\rho}(t).\end{aligned} \tag{6.24}$$

The TEC can be then solved as

$$\begin{aligned}\frac{1}{\alpha}\left(\frac{\omega_1^2\omega_2^2}{\omega_1^2-\omega_2^2}\right)\Delta\rho(L,t) =& TEC(L) + \frac{1}{\alpha}\left(\frac{\omega_1^2\omega_2^2}{\omega_1^2-\omega_2^2}\right)(cb_{\rho,rec,\omega_2,} - cb_{\rho,rec,\omega_1,}) \\ &+ \frac{1}{\alpha}\left(\frac{\omega_1^2\omega_2^2}{\omega_1^2-\omega_2^2}\right)(cb_{\rho,sat,\omega_2} - cb_{\rho,sat,\omega_1}) \\ &+ \frac{1}{\alpha}\left(\frac{\omega_1^2\omega_2^2}{\omega_1^2-\omega_2^2}\right)(M_{\omega_2,\rho,2}(t) - M_{\omega_1,\rho,1}(t)) \\ &+ \frac{1}{\alpha}\left(\frac{\omega_1^2\omega_2^2}{\omega_1^2-\omega_2^2}\right)(\varepsilon_{\omega_2,\rho}(t) - \varepsilon_{\omega_1,\rho}(t)) \\ =& TEC(L) + DCB_{\rho,rec,\omega_1,\omega_2} + DCB_{\rho,sat,\omega_1,\omega_2} \\ &+ M_{\rho,\omega_1,\omega_2}(t) + \varepsilon_{\rho,\omega_1,\omega_2}(t),\end{aligned} \tag{6.25}$$

where the scaling of the differential error and bias terms converts them into TEC units. As it is convenient to use TEC units from here on, the terms are renamed on the last line. Variable DCB stands for *differential code bias* (DCB) and it is unknown. For the more precise P-code differential group measurement a 2 ns precision results in a 5.7 TECU precision. The possible multipath errors in the observations are in the scale of 10 m (Wells et al., 1986) resulting in almost 100 in TECU. Hence, in practice, the differential group delay TEC can be considered as an absolute measurement up to DCB, but contaminated with relatively large measurement noise.

Carrier phase observable

The *Carrier phase*, also known as *carrier beat phase*, *phase difference* and *phase advancement* measurement or observable, is used similarly for LEO beacon and GNSS differential carrier phase measurements. In the carrier phase measurement the difference is taken between the incoming signal phase and a constant reference frequency generated in the receiver. The measurement is the phase difference. When measuring the phase, the initial number of full phase difference cycles cannot be detected. Hence, a new bias term, *phase ambiguity* γ, needs to be included in the model. If the signal is lost during the measurement a new phase ambiguity bias term needs to be added into the model (Wells et al., 1986; Vierinen et al., 2014).

By including the phase ambiguity, with other additional bias and error parameters, in Equation (6.20), the *carrier phase observable* can be written as

$$\begin{aligned}\phi(t) =& \omega\left(\tau_{rec} + \tau_{sat} + b_{\phi,rec,\omega} + b_{\phi,sat,\omega}\right) + \frac{\omega}{c}\left(L + T(L)\right) \\ &- \frac{\alpha}{c\omega}TEC(L) + M_{\phi,sat,rec,\omega}(t) + \gamma_{sat,rec,\omega} + \varepsilon_{\phi,sat,rec,\omega}(t),\end{aligned} \tag{6.26}$$

where similarly to pseudorange measurements the clock and hardware bias terms τ and b are converted from seconds and range and tropospheric bias from metres to radians.

For a differential carrier phase measurement in GPS L1 frequency the precision rule of 1% from the wavelength results in a range precision of 2 mm (Wells et al., 1986). The corresponding TEC measurement precision is approximately 0.02 TECU. The TEC measurement precision improves with lower frequencies.

Differential carrier phase measurement

As the carrier phase observable is measured in radians, measurements in two frequencies need to be scaled to same frequency before the subtraction.

$$\begin{aligned}
\Delta\phi =& \phi_{\omega_2} - \frac{\omega_2}{\omega_1}\phi_{\omega_1} \\
=& \omega_2 \left(\tau_{rec} + \tau_{sat} + b_{\phi,rec,\omega_2} + b_{\phi,sat,\omega_2}\right) + \frac{\omega_2}{c}\left(L + T(L)\right) \\
& - \frac{\alpha}{c\omega_2} TEC(L) + M_{\phi,sat,rec,\omega_2}(t) + \gamma_{sat,rec,\omega_2} + \varepsilon_{\phi,sat,rec,\omega_2}(t) \\
& - \frac{\omega_2}{\omega_1}\Big[\omega_1 \left(\tau_{rec} + \tau_{sat} + b_{\phi,rec,\omega_1} + b_{\phi,sat,\omega_1}\right) + \frac{\omega_1}{c}\left(L + T(L)\right) \\
& + \frac{\alpha}{c\omega_1} TEC(L) + M_{\phi,sat,rec,\omega_1}(t) + \gamma_{sat,rec,\omega_1} + \varepsilon_{\phi,sat,rec,\omega_1}(t)\Big] \\
=& \frac{\alpha}{c}\left(\frac{\omega_2^2 - \omega_1^2}{\omega_1^2\omega_2}\right) TEC(L) \\
& + \omega_2 \left(b_{\phi,rec,\omega_2} - b_{\phi,rec,\omega_1}\right) + \omega_2 \left(b_{\phi,sat,\omega_2} - b_{\phi,sat,\omega_1}\right) + M_{\phi,sat,rec,\omega_2}(t) - \frac{\omega_2}{\omega_1} M_{\phi,sat,rec,\omega_1}(t) \\
& + \gamma_{sat,rec,\omega_2} - \frac{\omega_2}{\omega_1}\gamma_{sat,rec,\omega_1} + \varepsilon_{\phi,sat,rec,\omega_2}(t) - \frac{\omega_2}{\omega_1}\varepsilon_{\phi,sat,rec,\omega_1}(t)
\end{aligned} \tag{6.27}$$

Converting the measured difference to TEC units results in

$$\begin{aligned}
\frac{c}{\alpha}\left(\frac{\omega_1^2\omega_2}{\omega_2^2 - \omega_1^2}\right)\Delta\phi =& TEC(L) + \frac{c}{\alpha}\left(\frac{\omega_1^2\omega_2}{\omega_2^2 - \omega_1^2}\right)\left(\omega_2 \left(b_{\phi,rec,\omega_2} - b_{\phi,rec,\omega_1}\right) - \omega_2 \left(b_{\phi,sat,\omega_2} - b_{\phi,sat,\omega_1}\right)\right) \\
& + \frac{c}{\alpha}\left(\frac{\omega_1^2\omega_2}{\omega_2^2 - \omega_1^2}\right)\left(M_{\phi,sat,rec,\omega_2}(t) - \frac{\omega_2}{\omega_1} M_{\phi,sat,rec,\omega_1}(t)\right) \\
& + \frac{c}{\alpha}\left(\frac{\omega_1^2\omega_2}{\omega_2^2 - \omega_1^2}\right)\left(\gamma_{sat,rec,\omega_2} - \frac{\omega_2}{\omega_1}\gamma_{sat,rec,\omega_1}\right) \\
& + \frac{c}{\alpha}\left(\frac{\omega_1^2\omega_2}{\omega_2^2 - \omega_1^2}\right)\left(\varepsilon_{\phi,sat,rec,\omega_2}(t) - \frac{\omega_2}{\omega_1}\varepsilon_{\phi,sat,rec,\omega_1}(t)\right) \\
=& TEC(L) + IFB_{\phi,rec,\omega_1,\omega_2} + IFB_{\phi,sat,\omega_1,\omega_2} + M_{\phi,sat,rec,\omega_1,\omega_2}(t) \\
& + \gamma^*_{\phi,sat,rec,\omega_1,\omega_2} + \varepsilon^*_{\phi,sat,rec,\omega_1,\omega_2}(t),
\end{aligned} \tag{6.28}$$

where the scaled differential bias terms are in TEC units and renamed at the last line as *interfrequency bias* (IFB) denoted here with variable IFB, and the naming for the rest of the parameters is self-explanatory.

The phase ambiguity γ^* is an unknown constant for a continuous measurement between a satellite-receiver pair. The IFBs can be assumed constant for each individual receiver and satellite when using the same frequency pairs. Hence, the IFBs can be included in the phase ambiguity parameter as

$$\gamma_{\phi,sat,rec,\omega_1,\omega_2} := IFB_{\phi,rec,\omega_1,\omega_2} + IFB_{\phi,sat,\omega_1,\omega_2} + \gamma^*_{\phi,sat,rec,\omega_1,\omega_2}.$$

The multipath error is on a centimeter scale (Wells et al., 1986) and can be included in the measurement error term

$$\varepsilon_{\phi,sat,rec,\omega_1,\omega_2}(t) := M_{\phi,sat,rec,\omega_1,\omega_2}(t) + \varepsilon^*_{\phi,sat,rec,\omega_1,\omega_2}(t).$$

The differential carrier phase TEC measurement can then be written as

$$\frac{c}{\alpha}\left(\frac{\omega_1^2\omega_2}{\omega_2^2-\omega_1^2}\right)\Delta\phi = TEC(L) + \gamma_{\phi,sat,rec,\omega_1,\omega_2} + \varepsilon_{\phi,sat,rec,\omega_1,\omega_2}(t). \tag{6.29}$$

The precision of differential phase measurement at GPS frequencies is approximately 0.03 TECU and differential phase measurements with much lower LEO beacon frequencies are even more precise. However, as $\gamma_{\phi,sat,rec,\omega_1,\omega_2}$ is not known, the measurement remains relative.

6.2.4 Carrier phase leveling

The differential code measurement is absolute up to DCB, but has a low precision. The phase differential measurement is relative due to the unknown phase ambiguity; however the measurement precision is high.

To achieve the absolute scale for the more accurate carrier phase measurement, the differential carrier phase is fitted to the differential group delay measurement. The fitting is done by calculating the offset σ_{offset} between the two measurements using the high-elevation parts as the low-elevation measurements are more prone to multipath errors (Klobuchar, 1996; Horvath and Crozier, 2007). The curve fitting results in real TEC measurement, which is absolute with respect to phase ambiguity and has the accuracy of the phase measurements, but still contains hardware biases. The final measurement can be written as

$$\frac{c}{\alpha}\left(\frac{\omega_1^2\omega_2}{\omega_2^2-\omega_1^2}\right)\Delta\phi + \sigma_{\text{offset}} = TEC(L) + DCB_{\rho,rec,\omega_1,\omega_2} + DCB_{\rho,sat,\omega_1,\omega_2} + \varepsilon_{\phi,sat,rec,\omega_1,\omega_2}(t). \tag{6.30}$$

Differential code bias (DCB)

The reason why differential phase measurements with DCB, instead of phase ambiguity, are preferred is that the phase ambiguity remains the same only for each individual satellite-receiver lock. The DCBs are also unknown, however, it can often be assumed that each individual signal transmitted from a GNSS satellite has a DCB that is independent of the receiver. Correspondingly, it can be assumed that each receiver has a DCB for each different signal, independent of the transmitting satellite. In comparison to TEC, it can be assumed that the changes in DCBs are slow. However, the receiver DCB depend on e.g. temperature (Coster et al., 2013) and daily variations of over 8 TECU are reported (Dyrud et al., 2008). In the GLONASS system, each satellite transmits slightly different frequencies and thus each satellite-receiver combination has its unique DCBs for different signals between the pair. Detailed reviews about GNSS biases are provided by Dyrud et al. (2008); Håkansson et al. (2017). For bias calibration see e.g. Dyrud et al. (2008); Vierinen et al. (2016).

6.2.5 LEO beacon satellite measurement model

The ground-based measurement of LEO dual-frequency beacon signals is a differential carrier phase measurement as given in Section 6.2.3. The TEC measurements can then be written as a linear model given in Equation (3.3)

$$\mathbf{m}_{\mathrm{LEO}} \approx \mathbf{A}_{\mathrm{LEO}} \boldsymbol{f} + \mathbf{B}_{\gamma} \boldsymbol{\gamma} + \boldsymbol{\varepsilon}_{\mathrm{LEO}}, \tag{6.31}$$

where the vector $\mathbf{m}_{\mathrm{LEO}}$ consists of individual relative slant TEC measurements given in (6.29) and correspondingly $\boldsymbol{\varepsilon}_{\mathrm{LEO}}$ the measurement errors $\varepsilon_{\phi,sat,rec,\omega_1,\omega_2}$. Similarly to Equation (2.3), the rows of matrix $\mathbf{A}_{\mathrm{LEO}}$ are discrete approximations for the integral operators of slant TEC (6.22) for all the measured signal paths operating on the unknown electron density values $\boldsymbol{f} \in \mathbb{R}^N$ in the discretised three-dimensional domain. The vector $\boldsymbol{\gamma}$ consists of unknown phase ambiguity constants $\gamma_{\phi,sat,rec,\omega_1,\omega_2}$ and it needs to be taken into account as an additional unknown. As the phase ambiguity remains the same during each continuous observation, several individual measurements in the above model share a common

$\gamma_{\phi,sat,rec,\omega_1,\omega_2}$ parameter.

$$\mathbf{B}_\gamma = \begin{bmatrix} 1 & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & 0 & 0 & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 1 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & 0 \\ 0 & 0 & 0 & \dots & 1 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \dots & 1 \end{bmatrix} \tag{6.32}$$

is a design matrix of zeros and ones, which picks the correct ambiguity parameter for each measurement.

6.2.6 GNSS satellite measurement model

TEC measurements of GNSS satellite signals are based on the levelled carrier phase measurement given in Equation (6.30). In the form of Equation (3.3), a vector of GNSS measurements can then be modelled as

$$\mathbf{m}_{\text{GNSS}} \approx \mathbf{A}_{\text{GNSS}}\boldsymbol{f} + \mathbf{B}_{\text{rec}}\boldsymbol{b}_{\text{rec}} + \mathbf{B}_{\text{GNSS}}\boldsymbol{b}_{\text{GNSS}} + \boldsymbol{\varepsilon}_{\text{GNSS}}, \tag{6.33}$$

where again the vector $\mathbf{m}_{\text{GNSS}}$ consists of individual GNSS TEC measurements (6.30). The measurement error vector $\boldsymbol{\varepsilon}_{\text{GNSS}}$ consists of error terms $\varepsilon_{\phi,sat,rec,\omega_1,\omega_2}$. Similarly to LEO measurements, the matrix $\mathbf{A}_{\text{GNSS}}$ is the discretisation of the integrals in slant TEC (6.22) acting on the discretised electron density values $\boldsymbol{f}$, as shown in Equation (2.3) for the two-dimensional measurements. The vectors $\boldsymbol{b}_{\text{rec}}$ and $\boldsymbol{b}_{\text{sat}}$ consists of DCBs of the measurements. $\mathbf{B}_{\text{rec}}$ and $\mathbf{B}_{\text{sat}}$ are design matrices, similar to $\mathbf{B}_\gamma$ in Equation (6.32), picking the correct DCB for each measurement.

The altitudes of GNSS satellites are around 20,000 km, thus besides the ionosphere, most of the signal propagation takes place in the plasmasphere above. The electron density in plasmasphere is generally much lower than in the ionosphere; however, due to the long ray paths, the resulting contribution to electron content can be significant. In Lunt et al. (1999) it is reported that during solar minimum the plasmaspheric contribution over Europe is typically a few TEC units. At night, especially in winter it can constitute 50% or even more in GNSS measurements. The contribution decreases at higher latitudes and the proportional contribution decreases towards solar maximum.

If the whole domain spanned by the receivers and satellites should be modelled similarly, the resulting grid size can become unnecessarily high-dimensional. One technique to reduce

dimensions is by using an irregular grid, where voxel sizes increase towards the boundaries, particularly at high altitudes. Another scheme is to extend the grid only over the ionosphere and use plasmaspheric models (see e.g. Jakowski and Hoque (2018) and references therein) for the exceeding parts of the measurements. The plasmaspheric model can be introduced into the measurement model in (6.22), as

$$TEC(L) = \int_{L_0}^{L_{\text{iono}}} N_e(z)\mathrm{d}z + \int_{L_{\text{iono}}}^{L} N_e(z)\mathrm{d}z \approx \int_{L_0}^{L_{\text{iono}}} N_e(z)\mathrm{d}z + \int_{L_{\text{iono}}}^{L} N_{e,\text{pmodel}}(z), \quad (6.34)$$

where L_{iono} is the upper boundary altitude of the reconstruction grid and $N_{e,\text{pmodel}}$ a plasmaspheric model. A straightforward approach is to assume a uniform but unknown plasmaspheric electron density, resulting in

$$TEC(L) \approx \int_{L_0}^{L_{\text{iono}}} N_e(z)\mathrm{d}z + (L - L_{\text{iono}})N_{e,\text{punif}}, \quad (6.35)$$

where $N_{e,\text{punif}}$ the uniform plasmaspheric electron density constant. The TEC in (6.34) or (6.35) is then plugged into equation (6.30). The selection of (6.35) introduces an additional unknown in the final measurement model (6.33).

In ionospheric studies, the majority of GNSS TEC measurements are typically made with ground-based receivers, with fixed and known locations. However, in satellite *radio occultation* (RO) the GNSS measurement is carried out with a LEO satellite onboard receiver (Hajj et al., 1994). The main difference between the two is thus that in RO also the receiver is in motion. The equations above can be used for modelling both ground-based and RO measurements.

6.3 Ionosonde measurements

In 1924 Breit and Tuve (1926) proved the existence of an ionised layer in Earth's atmosphere by receiving ionospheric echoes from a transmitted *high-frequency* (HF, 3–30 MHz) signal. The seminal study laid the foundation for ionospheric soundings and ionosondes.

An ionosonde is practically a radar transmitting HF pulses and measuring the time it takes for a pulse to travel back and forth to the reflection altitude in the ionosphere. The reflection occurs when the refractive index reaches zero. For the ordinary mode that is the altitude where the plasma frequency matches the frequency of the propagating wave, while a signal with a higher frequency than the current maximum plasma frequency will penetrate trough the ionosphere. Hence, the range of the transmitted frequencies should cover the current plasma frequency. Usual ionospheric peak electron densities range from 10^{10} to 10^{12} $\frac{1}{\text{m}^3}$ (Klobuchar, 1985), with conversions given in Equation (6.5) the corresponding plasma frequency range from 0.9 to 9 MHz. A typical ionosonde covers the frequencies from 0.5 to 20 MHz.

When the signal frequency in use is close to the plasma frequency the earlier assumptions regarding collision and gyro frequencies are not valid. After the lowest frequencies, when the signal frequency is higher than 2 MHz (Klobuchar, 1985), the collisions can again be neglected, however, even at the highest frequencies used in ionosondes, the presence of the magnetic field and hence gyrofrequency needs to be taken into account in the refractive index given with Appleton-Lassen formula (6.3).

The pulses transmitted from an ionosonde travel at group velocity (6.9). Before the reflection occurs, the propagating wave is slowed down by the ionisation below the reflection altitude. Hence, deriving the altitude from the signal travel time assuming that the pulses propagate with the speed of light will result in so-called *virtual height*

$$h' = \frac{c}{2}\Delta t. \tag{6.36}$$

The modelling for the roundtrip time can be carried out more accurately by taking the group index into account as

$$\Delta t = \frac{2}{c}\int_0^h n_{\mathrm{g}}(z)\mathrm{d}z, \tag{6.37}$$

where the integral is defined along a line from ionosonde location at $z = 0$ to $z = h$, where h is the *real height* i.e. the actual reflection height. When the pulse frequency and plasma frequency up to current altitude are included, the virtual height can be written as

$$h'(\omega) = \int_0^{h(\omega)} n_{\mathrm{g}}(\omega_{\mathrm{p}}(z), \omega)\mathrm{d}z. \tag{6.38}$$

The solution for $h(\omega)$ in Equation (6.38) is a nonlinear problem and cannot be solved within the linear framework provided in Chapter 4. In ionospheric imaging, usually real height profiles pre-analysed with some specific scaling algorithm, such as NhPC (Huang and Reinisch, 1996) within Automatic Real Time Ionogram Scaler with True Height (ARTIST) (Reinisch and Huang, 1983), POLynomial ANalysis (POLAN) (Titheridge, 1985), Autoscala (Pezzopane and Scotto, 2004) and NeXtYZ (Zabotin et al., 2006) are used.

An analysed real height profile is a vector of reflection altitudes for corresponding pulse frequencies. Typically it is assumed that the reflections take place directly above the instrument location. This results in a measurement model

$$h_{\mathrm{ionos}}(\omega) = h(\omega) + \varepsilon_{\mathrm{ionos}}(\omega), \tag{6.39}$$

where $\varepsilon_{\mathrm{ionos}}$ are errors in altitude that are dependent on each other within each analysed profile. When converting the plasma frequencies and measurement errors to electron density and approximating them in model grid points, ionosonde measurements can be modelled in the form of Equation (3.3) as

$$\mathbf{m}_{\mathrm{ionos}} \approx \mathbf{A}_{\mathrm{ionos}}\boldsymbol{f} + \boldsymbol{\varepsilon}_{\mathrm{ionos}}, \tag{6.40}$$

where $\mathbf{A}_{\text{ionos}}$ is a simple design matrix picking the column of $\boldsymbol{f}$ closest to the instrument location up to the highest reflection altitude.

6.4 Incoherent scatter radar measurements

The principle of a basic radar is to transmit pulsed or continuous electromagnetic waves and to receive the signal reflected from a hard target, such as an aeroplane, a ship or a speeding car. Based on the travel time and the doppler shift of the signal the measurement typically gives the distance to the target as well as the speed in the radial direction from the radar.

The incoherent scatter radar (ISR) theory was first proposed by Gordon (1958) to investigate Earth's ionosphere. Intuitively the incoherent scatter can be understood as a number of small scatterers distributed randomly in a volume. However, instead of an actual reflection, the free electrons in the ionosphere will accelerate when illuminated with the incident electric field of the radar signal. As a result, the electrons start to re-radiate as Hertzian dipoles in the corresponding frequency. The physical phenomenon is called Thomson scattering. The movement of the electrons in ionospheric plasma is not completely free, but it is dominated by the significantly more massive positive ions. Thus, even though the incoherent backscatter is from the electrons, the measurement will include a Doppler effect originating from the ion velocities.

In contrast to a single hard target, the ionosphere is a continuous medium and the scattering will take place at several altitudes. If a plain continuous sine wave signal is transmitted, it is impossible to say from which distance the received signal is scattering. To overcome this so-called range ambiguity, some transmission modulation is required in the transmitted signal.

Due to spatial and temporal fluctuations in plasma, the measured backscattered field can be considered as a Gaussian random variable with zero mean. Hence, it is more informative to estimate the covariance of the measurements. The estimated autocovariance function can be represented also with its Fourier transform pair i.e. power spectral density. Based on the power spectral density, plasma parameters such as electron density, ion and electron temperature, ion mass ratio and ion velocity can be obtained.

For ionospheric imaging, the most important plasma parameter is the electron density. An individual electron density measurement can be modelled as

$$m_{\text{IS}} = N_e(z) + \varepsilon_{\text{IS}}, \tag{6.41}$$

where $z \in \mathbb{R}^3$ is the measurement location along the radar beam and ε_{IS} the corresponding measurement error. Similarly to ionosonde measurements, the incoherent scatter radar measurements can be interpolated to model grid points and modelled as direct measurements of unknown electron densities in the form of Equation (3.3) as

$$\mathbf{m}_{\text{IS}} \approx \mathbf{A}_{\text{IS}}\boldsymbol{f} + \boldsymbol{\varepsilon}_{\text{IS}}, \tag{6.42}$$

where the design matrix $\mathbf{A}_{\mathrm{IS}}$ selects values of $\boldsymbol{f}$ related to corresponding measurements.

6.5 Langmuir probe in situ measurements

Langmuir probe is named after Irving Langmuir, who pioneered the method at General Electric in the 1920s. It is one of the most straightforward ways of measuring plasma (Klobuchar, 1985). However, it is an in situ measurement performed inside the medium, hence, ionospheric plasma measurements require a vehicle with an access to the ionosphere, such as a satellite or a rocket.

Langmuir probe measurements are based on detection of electric current between two conducting surfaces interacting with the medium. Typically the current is measured between a plane, a cylindrical, or a spherical shaped electrode and the satellite surface. The measurements are carried out by changing the probe potential in small steps. The sweep over a range of realistic potentials produces a voltage-current curve. Plasma parameters, such as electron density, electric potential of plasma and electron temperature, can then be determined from the curve (Klobuchar, 1985; Hargreaves, 1992; Chen, 2003).

Similarly to ionosonde and incoherent scatter radar measurements, an individual electron density measurement provided my Langmuir probe is modelled here as

$$m_{\mathrm{LP}} = N_e(z) + \varepsilon_{\mathrm{LP}}, \tag{6.43}$$

where $z \in \mathbb{R}^3$ is the location of the probe. For a discretised system (3.3), a vector of measurements is modelled again as

$$\mathbf{m}_{\mathrm{LP}} \approx \mathbf{A}_{\mathrm{LP}}\boldsymbol{f} + \boldsymbol{\varepsilon}_{\mathrm{LP}}, \tag{6.44}$$

where the design matrix $\mathbf{A}_{\mathrm{LP}}$ selects values of $\boldsymbol{f}$ related to the probe location at the time of the measurement.

Chapter 7

Development of methodology in ionospheric imaging

The use of tomographic methods for ionospheric imaging was first suggested by Austen et al. (1986) and later published in Austen et al. (1988). The article presented a two-dimensional simulation study assuming LEO satellite measurements from a chain of receiver stations. Iterative ART and SIRT algorithms were used with a Chapman profile (Chapman, 1931) as an initial guess.

The first electron density reconstructions with real observations from LEO satellite transmissions were presented by Andreeva (1990). The relative nature of phase measurements was taken into account by solving the phase ambiguity within the inversion that was carried out with ART. Pryse and Kersley (1992) used independent EISCAT incoherent scatter radar data to validate reconstruction results obtained with a setup of two receivers and the SIRT algorithm.

Another early simulation study was carried out with MART by Raymund et al. (1990). In the aforementioned study the limitations of ionospheric measurements and necessity of prior information was acknowledged. The limitations of the satellite measurement geometry and the resulting ill-posedness were studied more explicitly later by Yeh and Raymund (1991) and Raymund et al. (1994b). Studies on resolution limits due to geometric limitations and effects on station spacing were later carried out by Na et al. (1995) and Sutton and Na (1995). Saksman et al. (1997) showed that due to restricted measurement geometry an infinite amount of ionospheric electron density functions can be defined that are invisible to such a set-up.

Work presented in Fremouw et al. (1992) had several major contributions for the field. It discussed the use of stochastic inversion presented in Tarantola and Valette (1982), used earlier mostly in geophysics. However, the inversion was carried out with weighted damped least squares (Menke, 1989), which is analogous to generalised Tikhonov regularisation. The approach also utilised basis functions, namely Fourier basis functions in the horizontal

direction and empirical orthogonal functions (EOF) in the vertical. The EOFs were based on model ionospheres.

The use of ionosonde measurements in ionospheric imaging was speculated on in Kersley et al. (1993). Raymund et al. (1994a) used scaled ionograms in a simulation study. Later Heaton et al. (1995) used ionosonde as prior information.

The inclusion of GNSS measurements to ionospheric imaging was already suggested by Yunck et al. (1988). The far-sighted speculations also acknowledged the lack of vertical information provided by ground-based satellite measurements and considered possibilities of satellite radio occultation (RO) measurements. A two-dimensional simulation study with GPS-to-LEO RO measurements was carried out by Hajj et al. (1994), where the TSVD approach was used to regularise the inversion. Singular values were also used to study the effect of improved measurement geometry. The first experimental results with GPS measurements were carried out by Rius et al. (1997). The approach used a three-dimensional spatial domain with four vertical layers and Kalman filter for the temporal dimension.

The use of basis functions in three dimensions was presented by Howe et al. (1998). The functions were constructed by combining spherical harmonics in the horizontal and EOF in the vertical direction. The simulation study considered GPS and LEO measurements and solved the GPS DCB within the procedure. The use of EOFs for three-dimensional tomography was continued later with Multi-Instrument Data Analysis System (MIDAS) by Mitchell and Spencer (2003); Bust et al. (2007); Chartier et al. (2012); Bruno et al. (2019).

In global-scale four-dimensional ionospheric imaging the amount and quality of measurements vary spatially, and in voxel-based approaches, the role of realistic physical prior information becomes even more pivotal. Due to the significant role of the prior distribution, these approaches are often referred to as data assimilation methods, and the prior distribution is then more commonly known as *background model*. The data assimilation methods and nomenclature originate from meteorology, oceanography and geophysics (Tarantola, 1987; Menke, 1989; Daley, 1991; Daley and Barker, 2000). In ionospheric imaging, most of the applied methods are variations of Kalman filter and can use any measurements that can be modelled as linearised functions of electron density. However, the large number of unknown parameters give rise to computational issues with the Kalman filter approach. The size of the covariance matrices in model space is $N \times N$ and computational complexity for school-book matrix multiplication and inversion grows as $\mathcal{O}(N^3)$. Hence, in the following methods there are practically two main differences: First, the selected background ionospheric model that can be anything from a simple climatology to complicated parametrised physical models. Second, how the algorithm handles the covariance matrices that are too large to fit in the computer memory.

Bust et al. (2004, 2007) derived the Ionospheric Data Assimilation Three-Dimensional (IDA3D) algorithm from Three-Dimensional Variational Data Assimilation (3DVAR) (Daley, 1991). The 3DVAR is a general approach allowing a nonlinear forward model, however,

when assuming a linear forward model and Gaussian measurement error and background distributions, the approach reverts to the Kalman filter. IDA3D uses ground-based GPS and LEO satellite, satellite RO, satellite in situ and ionosonde measurements. Several ionospheric models have been used as a backgound model, including International Reference Ionosphere (IRI) (Bilitza et al., 1993; Bilitza, 2001) and Parameterized Ionosphere Model (PIM) (Daniell et al., 1995). The background information is fed into the Kalman filter in the prediction step that is a mixture of earlier time step and the background model (Bust and Mitchell, 2008). According to Bust et al. (2004), IDA3D does not solve the posterior covariance, but only its diagonal i.e. the posterior variance.

Angling and Cannon (2004) presented an approach very similar to IDA3D, later entitled Electron Density Assimilative Model (EDAM) (Angling and Khattatov, 2006; Angling, 2008). EDAM uses PIM or IRI model as the prior mean and updates it with satellite RO and ground-based satellite measurements. EDAM is also a version of a Kalman filter, where a persistence model with exponential delay is used as a dynamic model. Only the diagonal of the posterior covariance matrix is solved and parametric correlations are given for non-diagonal entries for the following time step. Prior covariance matrix elements with the distance exceeding a predefined value are discarded resulting in a sparser covariance matrix. The method is also capable of solving DCBs within the tomographic analysis (Angling, 2008). Angling and Jackson-Booth (2011) added virtual height ionosonde measurements to EDAM in a nonlinear setting.

A method called GPS Ionospheric Inversion (GPSII) was presented by Fridman et al. (2006, 2009). It uses GPS and LEO satellite, GPS-LEO satellite RO, radio altimeter VTEC, LEO in situ, and ionosonde measurements. It uses a nonlinear model to obtain nonnegative solutions and the actual solution is obtained with the Newton-Kontorovich method. Within iterations, the GPSII uses a combination of generalized Tikhonov regularisation and Bayesian approach with a discrete multivariate Gaussian prior: The method uses a prior covariance matrix but it is also a weighted with a regularisation parameter. The Kalman filter-type error covariance matrix propagation is not used. Both IRI 2000 and PIM have been used as a background model. It is stated that a factorised (separable) representation of the covariance matrix has a substantial positive effect on the memory requirements and computation speed. GPSII solves DCB within the inversion.

A method utilising the Gaussian random field prior/Kriging/Gaussian process was introduced by Minkwitz et al. (2015). In this approach the covariance is given as a three-dimensional function that can be integrated according to measurement geometry to obtain the covariance for the TEC measurements. The covariance function parameters are then estimated from the measurement data. For the actual inversion, the covariance needs to be evaluated then only between the reconstructed locations and the TEC measurements. Hence, the reconstruction could be carried out e.g. for an individual two-dimensional plane inside the actual three-dimensional domain. In Minkwitz et al. (2016) the approach was extended to a four-dimensional case by adding the temporal dimension to the covariance function.

An approach called the Global Assimilation of Ionospheric Measurements (GAIM) model (Schunk et al., 2004; Scherliess et al., 2004; Gardner et al., 2014; Scherliess et al., 2017) has been developed by Utah State University. There exist different versions of the approach that use a reduced-state Kalman filter and ensemble Kalman filter. Different background models have been used from more simple ionospheric models (Schunk et al., 2004) to a physical Ionosphere-Plasmasphere Model that utilises ionospheric drivers such as neutral densities and winds, magnetospheric and equatorial electric fields, and auroral precipitation (Scherliess et al., 2004).

The similarly named Global Assimilative Ionospheric Model (GAIM) has been developed by the University of Southern California and the Jet Propulsion Laboratory (USC/JPL) (Rosen et al., 2001; Hajj et al., 2004; Wang et al., 2004). GAIM USC/JPL, again, has a simpler version utilising a Kalman filter, where covariance matrix elements corresponding to distances over a preset value are discarded. A more complicated version uses the 4DVAR approach. 4DVAR (Courtier et al., 1994) is a general variational approach for data assimilation, which, in a case of linear measurement and dynamical models, reverts to Kalman smoother (Carrassi et al., 2018). The 4DVAR version of GAIM USC/JPL incorporates several ionospheric drivers that are estimated along the electron density within a range of time.

More recently, Elvidge and Angling (2019) presented a method called advanced ensemble electron density assimilation system (AENeAS). Similarly to different GAIM models, AENeAS is a physics-based data assimilation model and it seeks to predict the ionospheric state. AENeAS uses Thermosphere Ionosphere Electrodynamics General Circulation Model (TIE-GCM) (Qian et al., 2014) and NeQuick (Nava et al., 2008) as background models and local ensemble transform Kalman filter (LETKF) (Hunt et al., 2007). LETKF is a version of EnKF where the assimilation is performed only for local regions, which further reduces the state space of the model.

An early review on ionospheric imaging methods is provided by Raymund (1994) and a comparison of methods by Raymund (1995). Another introduction to ionospheric imaging and its early methods is given by Fehmers (1996). A book on ionospheric imaging with focus on iterative methods is provided by Kunitsyn et al. (2003). Bust and Mitchell (2008) provide a comprehensive review article where most of the present methods are already discussed.

7.1 TomoScand

TomoScand is a system for ionospheric imaging generated during the 2010s in the Finnish meteorological institute and Sodankylä geophysical observatory, University of Oulu. At this point it has been used mostly regionally over Northern Europe. It utilises measurements of GPS, GLONASS, GALILEO and LEO satellite signals, incoherent scatter radars, ionosondes, satellite in situ probings and GNSS-to-LEO RO. The GNSS DCBs can be

estimated within the system.

Similarly to methods such as IDA3D, EDAM etc. presented above, the TomoScand algorithm uses a simplified Kalman filter without solving the posterior covariance, or solving only its diagonal. Currently the dynamical model (5.8) in use is a persistence model with transition $\mathbf{H}^{(l-1)} = \lambda \mathbf{I}$ and attenuation $0 \leq \lambda \leq 1$. Alternatively, some ionospheric model can be used in the prediction step or directly as a prior mean.

The essential difference to aforementioned similar techniques is in the construction of the prior covariance. In paticular, TomoScand uses GMRF correlation priors, presented in Section 4.3, for representing the prior distribution. TomoScand then relies heavily on sparse matrix implementations. Currently MUltifrontal Massively Parallel sparse direct Solver (MUMPS) (Amestoy et al., 2001, 2019) with an R interface RMUMPS *(https://github.com/morispaa/rmumps)* is used for solving the high-dimensional linear problem in parallel. The main steps of TomoScand are given in Algorithm 1. An example visualisation of TomoScand reconstruction from GNSS measurements is given in Figure 7.1.

Algorithm 1 TomoScand analysis

1. Set the spatial and temporal domain.
   - Grid (lat, long, alt) (geographic coordinates, possibly irregular).
   - Start and end time (UTC).
2. Read data
   - Read measurements (GNSS, LEO, ionosonde, in situ, occultation,...).
   - Read GNSS satellite DCBs (Section 6.2.4).
   - Data quality control and filtering.
3. Form measurement models
   - Formation of matrices $\mathbf{A}_{\mathrm{GNSS}}, \mathbf{A}_{\mathrm{LEO}}, \ldots$ corresponding with the measurements in use (Chapter 6).
   - Measurement error estimation, if not provided with data.
4. Form prior distribution (i.e. background model) (Chapter 4).
   - Set prior mean for the unknown electron density (i.e. background mean).
     - Prediction step (5.8)
   - Set prior covariance (i.e. background error covariance).
     - Set standard deviation/variance mask for the unknown electron density. Previous posterior variance can be used (see Step 6 below).
     - Set correlation lengths (in all 3 coordinate directions).
     - Form matrix $\mathbf{L}_{\mathrm{C}}$ (Section 4.3.1).
   - Set prior distributions for:
     - GNSS DCBs
     - LEO phase ambiguity
     - Uniform plasmaspheric contribution
5. Solve MAP estimate (4.24) for the sparse linear system (RMUMPS).
6. Optionally solve the posterior covariance, its diagonal, or parts of it (RMUMPS).
7. Save and plot results.
8. Set start and end time for the next step and start again from item 2 above.

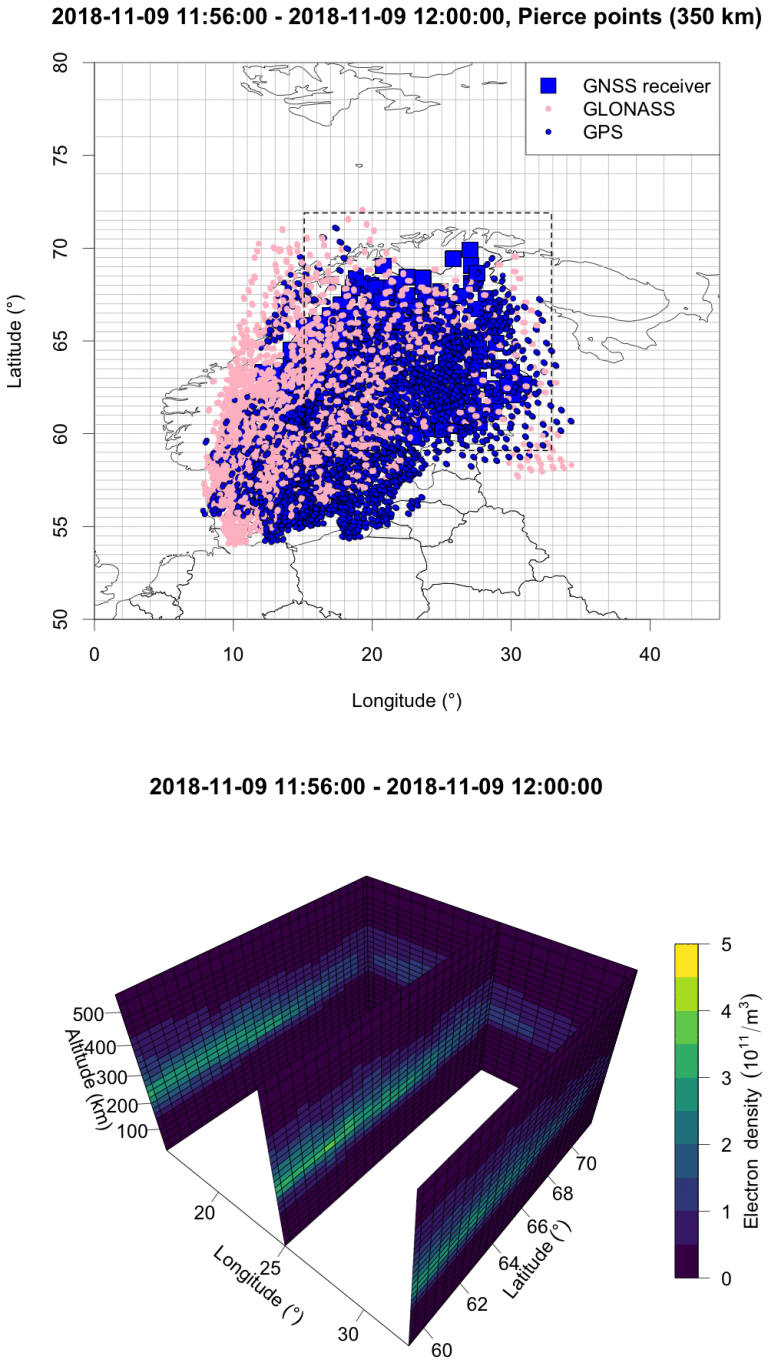

Figure 7.1: Example output from TomoScand analysis. Top: Tomographic domain, with irregular reconstruction grid and locations of the GPS and GLONASS satellite pierce points at an altitude of 350 km within a 2-min interval. Bottom: Three-dimensional reconstruction of ionospheric electron density.

Chapter 8

Discussion and conclusions

In this thesis an algorithm for four-dimensional multi-instrument ionospheric electron density imaging is developed. The algorithm uses a Bayesian approach for obtaining the most probable state of the ionospheric electron density, by updating the prior distribution i.e. the existing information of the ionospheric state with a set of new measurements. When used sequentially, the method is generally known as Kalman filter. In contrast to other Bayesian approaches used for ionospheric imaging, the prior distribution is essentially given as a Gaussian Markov Random Field correlation prior. The approach allows determining the prior covariance in an intuitive manner with a parametric function. However, for the numerical computations the covariance information is represented with a sparse precision matrix. Thus, instead of forming the $N \times N$ covariance matrix, approximately the same information is given with a precision matrix where the number of non-zero elements grows only as $\mathcal{O}(N)$. The precision matrix is also quick to construct and can easily be modified for different covariance structures. Effectively the same information can also be given for different discretisations of the domain as well as for irregular grids, as long as the discretisation lengths remain substantially shorter than the corresponding correlation lengths.

Unfortunately the sparsity cannot contribute to further steps when using full Kalman filter. Despite the initial sparsity, the resulting posterior covariance would again be a full and dense matrix that cannot be solved for high-dimensional problems. A solution for the diagonal of the posterior covariance i.e. posterior variance is possible. In this respect the approach is on par with the earlier methods.

To take into account the covariance from one time step to another, methods such as ensemble Kalman filter should be considered. However, on a regional scale, the dynamical transitions in ionospheric electron density can be substantial even in short timescales. Hence, even if the previous posterior covariance was solved, the role of the poorly known process noise covariance at the prediction step can be significant. Especially, when considering ionospheric imaging in an operational manner, how much can be achieved by putting much effort into advancing the covariance temporally is a relevant question. On the other

hand, for understanding the uncertainty related to any ionospheric electron density reconstructions, the examination of posterior covariances is essential.

For numerical computations, an R (R Core Team, 2017) implementation of the algorithm, called TomoScand, was written. Similarly to other imaging methods it can use any ionospheric electron density model as its background. However, on a regional scale the ionospheric electron density models can sometimes be severely flawed. Even in a case with dense receiver networks, the data assimilation can struggle if the background information is critically misleading. Hence, recently only a persistence model with attenuation coefficient has been used. Instead, emphasis is placed on the modelling of the prior/background error covariance that controls the uncertainty associated with the unknown electron density distribution. The approach improves the performance in regional imaging of the high-latitude ionospheric dynamics, providing an extension for the local measurements such as ISR and ionosonde measurements.

References

P. R. Amestoy, I. S. Duff, J. Koster, and J.-Y. J.-Y. L'Excellent. A Fully Asynchronous Multifrontal Solver Using Distributed Dynamic Scheduling. *SIAM Journal on Matrix Analysis and Applications*, 23(1):15–41, 2001. ISSN 0895-4798. doi: 10.1137/S0895479899358194.

P. R. Amestoy, A. Buttari, J. Y. L'Excellent, and T. Mary. Performance and scalability of the block low-rank multifrontal factorization on multicore architectures. *ACM Transactions on Mathematical Software*, 45(1):1–26, 2019. ISSN 15577295. doi: 10.1145/3242094.

A. H. Andersen and A. C. Kak. Simultaneous Algebraic Reconstruction Technique (SART): A Superior Implementation of the ART Algorithm. *Ultrasonic imaging*, 6:81–94, 1984. ISSN 15596915. doi: 10.1145/2387358.2387363.

E. S. Andreeva. Radio tomographic reconstruction of ionization dip in the plasma near the Earth. *J. Exp. Theor. Phys. Lett.*, 52:142–148, 1990.

M. J. Angling. Annales Geophysicae First assimilations of COSMIC radio occultation data into the Electron Density Assimilative Model (EDAM). *Annales Geophysicae*, pages 353–359, 2008.

M. J. Angling and P. S. Cannon. Assimilation of radio occultation measurements into background ionospheric models. *Radio Science*, 39:1–11, 2004. doi: 10.1029/2002RS002819.

M. J. Angling and N. K. Jackson-Booth. A short note on the assimilation of collocated and concurrent GPS and ionosonde data into the Electron Density Assimilative Model. *Radio Science*, 46(4):1–7, 2011. ISSN 00486604. doi: 10.1029/2010RS004566.

M. J. Angling and B. Khattatov. Comparative study of two assimilative models of the ionosphere. *Radio Science*, 41(April):1–11, 2006. doi: 10.1029/2005RS003372.

J. Austen, S. Franke, C. Liu, and K. Yeh. Application of computerized tomography techniques to ionospheric research. In A. Tauriainen, editor, *International Beacon Satellite Symposium on Radio Beacon Contribution to the Study of Ionization and Dynamics of*

the Ionosphere and to Corrections to Geodesy and Technical Workshop, Proceedings. Part 1 (A87-50101 22-46)., pages 25–35., Oulu, Finland, 1986. University of Oulu.

J. R. Austen, S. J. Franke, and C. H. Liu. Ionospheric imaging using computerized tomography Jeffrey. *Radio Science*, 23(3):299–307, 1988.

P. A. Bernhardt, C. A. Selcher, S. Basu, G. S. Bust, and S. C. Reising. Atmospheric studies with the tri-band beacon instrument on the COSMIC constellation. *Terrestrial, Atmospheric and Oceanic Sciences*, 11(1):291–312, 2000. ISSN 10170839.

D. Bilitza. International Reference Ionosphere 2000. *Radio Science*, 36(2):261–275, 2001. doi: 10.1029/2000RS002432.

D. Bilitza, K. Rawer, L. Bossy, and T. Gulyaeva. International reference ionosphere - past, present, and future: II. Plasma temperatures, ion composition and ion drift. *Advances in Space Research*, 13(3):15–23, 1993. ISSN 02731177. doi: 10.1016/0273-1177(93)90241-3.

G. Breit and M. A. Tuve. A test of the existence of the conducting layer. *Physical Review*, 28(3):554–575, 1926. ISSN 0031899X. doi: 10.1103/PhysRev.28.554.

A. Brekke. *Physics of the Upper Polar Atmosphere*. Wiley-Praxis Series in Atmospheric Physics. Wiley, 1997. ISBN 9780471960188.

J. Bruno, C. N. Mitchell, K. H. Bolmgren, and B. A. Witvliet. A realistic simulation framework to evaluate ionospheric tomography. *Advances in Space Research*, 2019. ISSN 02731177. doi: 10.1016/j.asr.2019.11.015. URL `https://linkinghub.elsevier.com/retrieve/pii/S0273117719308233`.

K. G. Budden. *Radio Waves in the Ionosphere: The Mathematical Theory of the Reflection of Radio Waves from Stratified Ionised Layers*. Cambridge University Press, Cambridge, 1st edition, 1961. ISBN 9780521043632.

G. S. Bust and C. N. Mitchell. History, current state, and future directions of ionospheric imaging. *Rev. Geophys.*, 46, 2008. doi: 10.1029/2006RG000212.

G. S. Bust, T. W. Garner, T. L. Gaussiran II, and T. L. Gaussiran. Ionospheric Data Assimilation Three-Dimensional (IDA3D): A global, multisensor, electron density specification algorithm. *J. Geophys. Res.*, 109(A11):1–14, 2004. ISSN 21699402. doi: 10.1029/2003JA010234.

G. S. Bust, G. Crowley, T. W. Garner, T. L. Gaussiran II, R. W. Meggs, C. N. Mitchell, P. S. J. Spencer, P. Yin, and B. Zapfe. Four-dimensional GPS imaging of space weather storms. *Space Weather*, 5(2), 2007. ISSN 15427390. doi: 10.1029/2006SW000237.

D. Calvetti and E. Somersalo. *Introduction to Bayesian Scientific Computing: Ten Lectures on Subjective Computing (Surveys and Tutorials in the Applied Mathematical Sciences)*. Springer-Verlag, Berlin, Heidelberg, 2007. ISBN 0387733930.

A. Carrassi, M. Bocquet, L. Bertino, and G. Evensen. Data assimilation in the geosciences: An overview of methods, issues, and perspectives. *Wiley Interdisciplinary Reviews: Climate Change*, 9(5):1–79, 2018. ISSN 17577799. doi: 10.1002/wcc.535.

Y. Censor. Finite Series-Expansion Reconstruction Methods. 71(3), 1983.

S. Chapman. The absorption and dissociative or ionizing effect of monochromatic radiation in an atmosphere on a rotating earth part {I} and part {II}. Grazing incidence. *Proceedings of the Physical Society*, 43(5):483–501, sep 1931. doi: 10.1088/0959-5309/43/5/302.

A. T. Chartier, C. N. Mitchell, and D. R. Jackson. A 12 year comparison of MIDAS and IRI 2007 ionospheric Total Electron Content. *Advances in Space Research*, 49(9):1348–1355, may 2012. ISSN 02731177. doi: 10.1016/j.asr.2012.02.014.

F. F. Chen. Lecture Notes on Langmuir Probe Diagnostics. *IEEE-ICOPS meeting*, pages 1–40, 2003.

G. Christakos. *Random Field Models in Earth Sciences*. Dover Publications, Mineola N.Y., 2005. ISBN 0-486-43872-4.

A. J. Coster, J. Williams, A. Weatherwax, W. C. Rideout, and D. Herne. Accuracy of GPS total electron content: GPS receiver bias temperature dependence. *Radio Science*, 48(2):190–196, 2013. ISSN 00486604. doi: 10.1002/rds.20011.

P. Courtier, J. Thépaut, and A. Hollingsworth. A strategy for operational implementation of 4D-Var, using an incremental approach. *Quarterly Journal of the Royal Meteorological Society*, 120(519):1367–1387, 1994. ISSN 00359009. doi: 10.1256/smsqj.51911.

R. Daley. *Atmospheric Data Analysis*. Cambridge Atmospheric and Space Science Series. Cambridge University Press, 1991. ISBN 9780521382151.

R. Daley and E. Barker. *NAVDAS Source Book. NRL Atmospheric Variational Data Assimilation System*. 2000.

R. E. Daniell, L. D. Brown, D. N. Anderson, M. W. Fox, P. H. Doherty, D. T. Decker, J. J. Sojka, and R. W. Schunk. Parameterized ionospheric model: A global ionospheric parameterization based on first principles models. *Radio Science*, 30(5):1499–1510, 1995. ISSN 1944799X. doi: 10.1029/95RS01826.

K. Davies. *Ionospheric Radio Propagation*. National Bureau of Standards Monograph, 1965.

K. Davies. *Ionospheric Radio*. IEE electromagnetic waves series. Peregrinus on behalf of the Institution of Electrical Engineers, London, 1990. ISBN 9780863411861.

L. Dyrud, A. Jovancevic, A. Brown, D. Wilson, and S. Ganguly. Ionospheric measurement with GPS: Receiver techniques and methods. *Radio Science*, 43(6):1–11, 2008. ISSN 00486604. doi: 10.1029/2007RS003770.

S. Elvidge and M. J. Angling. Using the local ensemble Transform Kalman Filter for upper atmospheric modelling. *Journal of Space Weather and Space Climate*, 9:A30, 2019. ISSN 2115-7251. doi: 10.1051/swsc/2019018.

G. Evensen. Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics. *Journal of Geophysical Research*, 99: 143–162, 1994.

G. Evensen. The Ensemble Kalman Filter: Theoretical formulation and practical implementation. *Ocean Dynamics*, 53(4):343–367, 2003. ISSN 16167341. doi: 10.1007/s10236-003-0036-9.

G. Evensen. *Data Assimilation, the Ensemble Kalman Filter*. Springer-Verlag, Dordrecht Heidelberg London New York, 2nd edition, 2009. ISBN 9783642037108. doi: 10.1007/978-3-642-03711-5.

G. C. Fehmers. *Tomography of the Ionosphere*. 1996. ISBN 9789038604381.

E. J. Fremouw, J. A. Secan, and B. M. Howe. Application of stochastic inverse theory to ionospheric tomography. *Radio Science*, 27(5):721–732, 1992. ISSN 1944799X. doi: 10.1029/92RS00515.

S. V. Fridman, L. J. Nickisch, M. Aiello, and M. Hausman. Real-time reconstruction of the three-dimensional ionosphere using data from a network of GPS receivers. *Radio Science*, 41(5):1–7, 2006. ISSN 00486604. doi: 10.1029/2005RS003341.

S. V. Fridman, L. J. Nickisch, and M. Hausman. Personal-computer-based system for real-time reconstruction of the three-dimensional ionosphere using data from diverse sources. *Radio Science*, 44(3):1–12, 2009. ISSN 00486604. doi: 10.1029/2008RS004040.

L. C. Gardner, R. W. Schunk, L. Scherliess, J. J. Sojka, and L. Zhu. Global assimilation of ionospheric measurements-gauss markov model: Improved specifications with multiple data types. *Space Weather*, 12(12):675–688, 2014. ISSN 15427390. doi: 10.1002/2014SW001104.

G. H. Golub and C. F. Van Loan. *Matrix Computations*. Johns Hopkins Studies in the Mathematical Sciences. Johns Hopkins University Press, fourth edi edition, 2013. ISBN 9781421407944.

R. Gordon, R. Bender, and G. T. Herman. Algebraic Reconstruction Techniques (ART) for three-dimensional electron microscopy and X-ray photography. *Journal of Theoretical Biology*, 29(3):471–481, 1970. ISSN 10958541. doi: 10.1016/0022-5193(70)90109-8.

W. E. Gordon. Incoherent Scattering of Radio Waves by Free Electrons with Applications to Space Exploration by Radar. *Proceedings of the IRE*, 46(11):1824–1829, 1958. ISSN 00968390. doi: 10.1109/JRPROC.1958.286852.

B. N. Hahn. Dynamic linear inverse problems with moderate movements of the object: Ill-posedness and regularization. *Inverse Problems and Imaging*, 9(2):395–413, 2015. ISSN 19308345. doi: 10.3934/ipi.2015.9.395.

G. A. Hajj, R. Ibañez-Meier, E. R. Kursinski, and L. J. Romans. Studying the ionosphere with the global positioning system. *International Journal of Imaging Systems & Technology*, 5(2):174–187, 1994. ISSN 00486604. doi: 10.1002/ima.1850050214.

G. A. Hajj, B. D. Wilson, C. Wang, X. Pi, and I. G. Rosen. Data assimilation of ground GPS total electron content into a physics-based ionospheric model by use of the Kalman filter. *Radio Science*, 39(1):1–17, 2004. ISSN 0048-6604. doi: 10.1029/2002rs002859.

M. Håkansson, A. B. Jensen, M. Horemuz, and G. Hedling. Review of code and phase biases in multi-GNSS positioning. *GPS Solutions*, 21(3):849–860, 2017. ISSN 15211886. doi: 10.1007/s10291-016-0572-7.

J. K. Hargreaves. *The solar-terrestrial environment: An introduction to geospace - the science of the terrestrial upper atmosphere, ionosphere and magnetosphere.* Cambridge University Press, Cambridge, UK; New York, USA; Melbourne, Australia, 1992. ISBN 0521327482.

J. A. T. Heaton, S. E. Pryse, and L. Kersley. Improved background representation, ionosonde input and independent verification in experimental ionospheric tomography. *Ann. Geophys.*, 13(1297-1302):1297–1302, 1995.

A. E. Hoerl and R. W. Kennard. Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics*, 12(1):55–67, 1970. ISSN 15372723. doi: 10.1080/00401706.1970.10488634.

I. Horvath and S. Crozier. Software developed for obtaining GPS-derived total electron content values. *Radio Science*, 42(2), 2007. ISSN 00486604. doi: 10.1029/2006RS003452.

B. M. Howe, K. Runciman, and J. A. Secan. Tomography of the ionosphere: Four-dimensional simulations. *Radio Science*, 33(1):109–128, 1998. ISSN 00486604. doi: 10.1029/97RS02615.

J. Hsieh. *Computed Tomography: Principles, Design, Artifacts, and Recent Advances*. SPIE, John Wiley & Sons, Bellingham, 2 edition, 2009. ISBN 0819444251. doi: 10.1117/3.817303.

X. Huang and B. W. Reinisch. Vertical electron density profiles from the digisonde network. *Advances in Space Research*, 18(6):121–129, 1996. ISSN 02731177. doi: 10.1016/0273-1177(95)00912-4.

B. R. Hunt, E. J. Kostelich, and I. Szunyogh. Efficient data assimilation for spatiotemporal chaos: A local ensemble transform Kalman filter. *Physica D: Nonlinear Phenomena*, 230 (1-2):112–126, 2007. ISSN 01672789. doi: 10.1016/j.physd.2006.11.008.

N. Hyvönen, M. Kalke, M. Lassas, H. Setälä, and S. Siltanen. Three-dimensional dental X-ray imaging by combination of panoramic and projection data. *Inverse Problems and Imaging*, 4(2):257–271, 2010. ISSN 19308337. doi: 10.3934/ipi.2010.4.257.

N. Jakowski and M. M. Hoque. A new electron density model of the plasmasphere for operational applications and services. 2018.

J. H. Justice, A. Vassiliou, S. Singh, J. D. Logel, P. A. Hansen, B. R. Hall, P. R. Hutt, and J. J. Solanki. Acoustic tomography for monitoring enhanced oil recovery. 1989.

S. Kaczmarz. Angenäherte Auflösung von Systemen linearer Gleichungen (english translation by Stockman, J.). *Bulletin International de l'Académie Polonaise des Sciences et des Lettres. Classe des Sciences Mathématiques et Naturelles. Série A, Sciences Mathématiques.*, 35:355 – 7, 1937.

J. Kaipio and E. Somersalo. *Statistical and Computational Inverse Problems*. Applied Mathematical Sciences. Springer, New York, 2005. ISBN 9780387271323.

J. Kaipio and E. Somersalo. Statistical inverse problems: Discretization, model reduction and inverse crimes. *Journal of Computational and Applied Mathematics*, 198(2):493–504, 2007. ISSN 03770427. doi: 10.1016/j.cam.2005.09.027.

A. C. Kak and M. Slaney. *Principles of Computerized Tomographic Imaging*. IEEE Press, New York, electronic edition, 1988. ISBN 978-0-89871-494-4. doi: 10.1118/1.1455742.

R. E. Kalman. A new approach to linear filtering and prediction problems. *Journal of Fluids Engineering, Transactions of the ASME*, 82(1):35–45, 1960. ISSN 1528901X. doi: 10.1115/1.3662552.

E. D. Kaplan and C. J. Hegarty. *Understanding GPS - Principles and Applications*. Artech House, Inc., Norwood, 2nd edition, 2006. ISBN 1-58053-894-0. doi: 10.1016/S1364-6826(97)83337-8.

L. Kersley, J. A. T. Heaton, S. E. Pryse, and T. D. Raymund. Experimental ionospheric tomography with ionosonde input and EISCAT verification. *Annales Geophysicae*, 11: 1064–1074, 1993.

J. A. Klobuchar. Ionospheric radio wave propagation. In A. S. Jursa, editor, *HANDBOOK OF GEOPHYSICS AND THE SPACE ENVIRONMENT*, chapter 10, pages 1–111. AIR FORCE GEOPHYSICS LABORATORY, AIR FORCE SYSTEMS COMMAND, UNITED STATES AIR FORCE, 4 edition, 1985.

J. A. Klobuchar. Ionospheric effects on GPS positioning. In B. W. Parkinson and J. J. Spilker, editors, *Global Positioning System : Theory and Applications, vol. 1*, chapter 12, pages 485–515. American Institute of Aeronautics and Astronautics, Inc., 1996. ISBN 978-1-60086-638-8. doi: https://doi.org/10.2514/4.866388.

V. E. Kunitsyn, E. S. Andreeva, S. J. Franke, and K. C. Yeh. Tomographic investigations of temporal variations of the ionospheric electron density and the implied fluxes. *Geophysical Research Letters*, 30(16):1851, 2003. ISSN 0094-8276. doi: 10.1029/2003gl016908.

N. Lunt, L. Kersley, and G. J. Bailey. The influence of the protonosphere on GPS observations: model simulations. *Radio Science*, 34(3):725–732, 1999. ISSN 00486604. doi: 10.1029/1999RS900002.

W. Menke. *Geophysical Data Analysis: Discrete Inverse Theory*. Academic Press, San Diego, revised ed edition, 1989. ISBN 0080507328.

D. Minkwitz, K. G. Van Den Boogaart, T. Gerzen, and M. Hoque. Tomography of the ionospheric electron density with geostatistical inversion. *Annales Geophysicae*, 33:1071–1079, 2015. doi: 10.5194/angeo-33-1071-2015.

D. Minkwitz, K. G. Van Den Boogaart, T. Gerzen, M. Hoque, and M. Hernández-pajares. Ionospheric tomography by gradient-enhanced kriging with STEC measurements and ionosonde characteristics. *Annales Geophysicae*, 34:999–1010, 2016. doi: 10.5194/angeo-34-999-2016.

C. N. Mitchell and P. S. J. Spencer. A three-dimensional time-dependent algorithm for ionospheric imaging using GPS. *Annals of Geophysics*, 46(4):687–696, 2003. ISSN 15935213. doi: 10.4401/ag-4373.

J. L. Mueller and S. Siltanen. *Linear and Nonlinear Inverse Problems with Practical Applications*. Society for Industrial and Applied Mathematics, Philadelphia, PA, 2012. doi: 10.1137/1.9781611972344.

H. Na, B. Hall, and E. Sutton. Ground staion spacing effects in ionospheric tomography. *Annales Geophysicae*, 13:1288–1296, 1995.

F. Natterer and F. Wübbeling. *Mathematical Methods in Image Reconstruction*. SIAM, Philadelphia, 2001. ISBN 0898716225. doi: 10.1118/1.1455744.

B. Nava, P. Coısson, and S. M. Radicella. Atmospheric and Solar-Terrestrial Physics A new version of the NeQuick ionosphere electron density model. 70:1856–1862, 2008. doi: 10.1016/j.jastp.2008.01.015.

B. W. Parkinson, J. J. Spilker Jr., M. Aparicio, P. Brodie, L. Doyle, J. Rajan, P. Torrione, A. J. Van Dierendonck, P. Axelrad, S. G. Francisco, J. A. Klobuchar, M. S. Braasch, J. F. Zumberge, W. I. Bertiger, F. Van Graas, N. Ashby, L. Kruczynski, and F. D. Natali. *Global Positioning System: Theory and Applications*, volume 1. American Institute of Aeronautics and Astronautics, Inc., Washington, 1996. ISBN 156347106X.

M. Pezzopane and C. Scotto. Software for the automatic scaling of critical frequency f0F2 and MUF(3000)F2 from ionograms applied at the Ionospheric Observatory of Gibilmanna. *Annals of Geophysics*, 47(6):1783–1790, 2004. ISSN 15935213. doi: 10.4401/ag-3375.

D. L. Phillips. A Technique for the Numerical Solution of Certain Integral Equations of the First Kind. *J. ACM*, 9(1):84–97, jan 1962. ISSN 0004-5411. doi: 10.1145/321105.321114.

S. E. Pryse and L. Kersley. A preliminary experimental test of ionospheric tomography. *Journal of Atmospheric and Terrestrial Physics*, 54(7-8):1007–1012, 1992. ISSN 00219169. doi: 10.1016/0021-9169(92)90067-U.

L. Qian, A. G. Burns, B. A. Emery, B. Foster, G. Lu, A. Maute, A. D. Richmond, R. G. Roble, S. C. Solomon, and W. Wang. TIE-GCM: A community model of the coupled thermosphere/ ionosphere system. In J. Huba, R. Schunk, and G. Khazanov, editors, *Modeling the ionosphere-thermosphere system*, pages 73–84. 2014. doi: 10.1029/2012GM001297.

R Core Team. R: A Language and Environment for Statistical Computing, 2017. URL `https://www.r-project.org/`.

C. E. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, Cambridge, 2006. ISBN 026218253X. doi: 10.1142/S0129065704001899.

T. D. Raymund. Ionospheric tomography algorithms. *International Journal of Imaging Systems and Technology*, 5(2):75–85, 1994. ISSN 10981098. doi: 10.1002/ima.1850050204.

T. D. Raymund. Comparison of several ionospheric tomography algorithms. *Annales Geophysicae*, 13:1254–1262, 1995.

T. D. Raymund, J. R. Austen, S. J. Franke, C. H. Liu, J. A. Klobuchar, and J. Stalker. Application of computerized tomography to the investigation of ionospheric structures. *Radio Science*, 25(5):771–789, 1990. ISSN 1944799X. doi: 10.1029/RS025i005p00771.

T. D. Raymund, Y. Bresler, D. N. Anderson, and R. E. Daniell. Model-assisted ionospheric tomography: A new algorithm. *Radio Science*, 29(6):1493–1512, 1994a.

T. D. Raymund, S. J. Franke, and K. C. Yeh. Ionospheric tomography: its limitations and reconstruction methods. *Journal of Atmospheric and Terrestrial Physics*, 56(5), 1994b. ISSN 00219169. doi: 10.1016/0021-9169(94)90104-X.

B. W. Reinisch and X. Huang. Automatic calculation of electron density profiles from digital ionograms: 3. Processing of bottomside ionograms. *Radio Science*, 18(3):477–492, 1983. ISSN 1944799X. doi: 10.1029/RS018i003p00477.

A. Rius, G. Ruffini, and L. Cucurull. Improving the vertical resolution of ionospheric tomography with GPS occultations. *Geophysical Research Letters*, 24(15):2291–2294, 1997. ISSN 0094-8276. doi: 10.1029/97gl52283.

L. Roininen, M. S. Lehtinen, S. Lasanen, M. Orispää, and M. Markkanen. Correlation priors. *Inverse Probl. and Imag.*, 5(1):167–184, 2011. doi: 10.3934/ipi.2011.5.167.

L. Roininen, P. Piiroinen, and M. S. Lehtinen. Constructing continuous stationary covariances as limits of the second-order stochastic difference equations. *Inverse Problems and Imaging*, 7(2):611–647, may 2013. ISSN 19308337. doi: 10.3934/ipi.2013.7.611.

I. G. Rosen, C. Wang, G. A. Hajj, X. Pi, and B. Wilson. An adjoint method based approach to data assimilation for a distributed parameter model for the ionosphere. *Proceedings of the IEEE Conference on Decision and Control*, 5:4406–4408, 2001. ISSN 01912216. doi: 10.1109/CDC.2001.980895.

H. Rue and L. Held. *Gaussian Markov Random Fields: Theory And Applications (Monographs on Statistics and Applied Probability)*. Chapman & Hall/CRC, 2005. ISBN 1584884320.

J. M. Rüeger. Refractive Index Formulae for Radio Waves. *Integration of Techniques and Corrections to Achieve Accurate Engineering*, pages 1–13, 2002.

E. Saksman, T. Nygrén, and M. Markkanen. Ionospheric structures invisible in satellite radiotomography. *Radio Science*, 32(2):605–616, 1997. ISSN 00486604.

S. Särkkä. Linear operators and stochastic partial differential equations in Gaussian process regression. In *Artificial Neural Networks and Machine Learning - ICANN 2011 - 21st International Conference on Artificial Neural Networks, Espoo, Finland, June 14-17, 2011, Proceedings, Part II*, pages 1–8, 2011. doi: 10.1007/978-3-642-21738-8-20.

S. Särkkä. *Bayesian filtering and smoothing*. Cambridge University Press, Cambridge, UK, 2013. ISBN 978-1-107-61928-9. doi: 10.1017/CBO9781139344203.

L. Scherliess, R. W. Schunk, J. J. Sojka, and D. C. Thompson. Development of a physics-based reduced state Kalman filter for the ionosphere. *Radio Science*, 39(1):1–12, 2004. ISSN 0048-6604. doi: 10.1029/2002rs002797.

L. Scherliess, R. W. Schunk, L. C. Gardner, J. V. Eccles, L. Zhu, and J. J. Sojka. The USU-GAIM-FP data assimilation model for ionospheric specifications and forecasts. In *2017 32nd General Assembly and Scientific Symposium of the International Union of Radio Science, URSI GASS 2017*, volume 2017-Janua, pages 1–4, 2017. ISBN 9789082598704. doi: 10.23919/URSIGASS.2017.8104978.

R. W. Schunk, L. Scherliess, J. J. Sojka, D. C. Thompson, D. N. Anderson, M. Codrescu, C. Minter, T. J. Fuller-Rowell, R. A. Heelis, M. Hairston, and B. M. Howe. Global Assimilation of Ionospheric Measurements (GAIM). *Radio Science*, 39(1):1–11, 2004. ISSN 0048-6604. doi: 10.1029/2002rs002794.

E. Sutton and H. Na. Comparison of geometries for ionospheric tomography. 30(1):115–125, 1995.

K. Tanabe. Projection method for solving a singular system of linear equations and its applications. *Numerische Mathematik*, 17(3):203–214, 1971. ISSN 0029599X. doi: 10.1007/BF01436376.

A. Tarantola. *Methods for Data Fitting and Parameter Estimation*. Inverse Problem Theory. Elsevier Science Publishers B. V., Netherlands, 1987.

A. Tarantola and B. Valette. Generalized nonlinear inverse problems solved using the least squares criterion. *Reviews of Geophysics*, 20(2):219, 1982.

A. N. Tikhonov and V. Y. Arsenin. *Solutions of ill-posed problems*. V. H. Winston & Sons, Washington, D.C.: John Wiley & Sons, New York, 1977.

J. E. Titheridge. Ionogram analysis with the generalised program POLAN. Technical report, WORLD DATA CENTER A for Solar-Terrestrial Physics, Auckland, New Zealand, 1985.

J. Vierinen, J. Norberg, M. S. Lehtinen, O. Amm, L. Roininen, A. Väänänen, P. J. Erickson, and D. McKay-Bukowski. Beacon satellite receiver for ionospheric tomography. *Radio Science*, 49(12):1141–1152, 2014. ISSN 1944799X. doi: 10.1002/2014RS005434.

J. Vierinen, A. J. Coster, W. C. Rideout, P. J. Erickson, and J. Norberg. Statistical framework for estimating GNSS bias. *Atmospheric Measurement Techniques*, 9(3):1303–1312, 2016. ISSN 18678548. doi: 10.5194/amt-9-1303-2016.

C. Wang, G. A. Hajj, X. Pi, I. G. Rosen, and S. Wing. Development of the Global Assimilative Ionospheric Model. *Radio Science*, 39(1):1–11, 2004. ISSN 00486604. doi: 10.1029/2002RS002854.

D. Wells, N. Beck, D. Delikaraoglou, A. Kleusberg, E. J. Krakiwsky, G. Lachapelle, R. B. Langley, M. Nakiboglu, K.-P. Schwarz, J. M. Tranquilla, and P. Vaníček. *Guide to GPS Positioning*. CANADIAN GPS ASSOCIATES, New Brunswick, Canada, 1986. ISBN 0920114733.

M. Yamamoto. Digital beacon receiver for ionospheric TEC measurement. *Earth, Planets and Space*, 60(3):21–24, 2008. ISSN 1343-8832. doi: 10.1186/BF03353137.

K. C. Yeh and T. D. Raymund. Limitations of ionospheric imaging by tomography. *Radio Science*, 26(6):1361–1380, 1991. ISSN 1944799X. doi: 10.1029/91RS01873.

T. P. Yunck, G. F. Lindal, and C.-H. Liu. The Role of GPS in Precise Earth Observation. In *IEEE PLANS '88.,Position Location and Navigation Symposium, Record. 'Navigation into the 21st Century'*, pages 251–258, Orlando, FL, USA, USA, 1988. IEEE. doi: 10.1109/PLANS.1988.195491.

N. A. Zabotin, J. W. Wright, and G. A. Zhbankov. NeXtYZ: Three-dimensional electron density inversion for dynasonde ionograms. *Radio Science*, 41(6):1–12, dec 2006. ISSN 00486604. doi: 10.1029/2005RS003352.