Helsinki University of Technology Laboratory of Computational Engineering Publications
Teknillisen korkeakoulun Laskennallisen tekniikan laboratorion julkaisuja
Espoo 2008 Report B66

HIERARCHICAL BAYESIAN ASPECTS OF DISTRIBUTED NEUROMAGNETIC SOURCE MODELS

Aapo Nummenmaa

TEKNILLINEN KORKEAKOULU
TEKNISKA HÖGSKOLAN
HELSINKI UNIVERSITY OF TECHNOLOGY
TECHNISCHE UNIVERSITÄT HELSINKI
UNIVERSITE DE TECHNOLOGIE D'HELSINKI

Helsinki University of Technology Laboratory of Computational Engineering Publications
Teknillisen korkeakoulun Laskennallisen tekniikan laboratorion julkaisuja
Espoo 2008 Report B66

HIERARCHICAL BAYESIAN ASPECTS OF DISTRIBUTED NEUROMAGNETIC SOURCE MODELS

Aapo Nummenmaa

Dissertation for the degree of Doctor of Philosophy to be presented with due permission of the Department of Electrical and Communications Engineering, Helsinki University of Technology, for public examination and debate in Auditorium E at Helsinki University of Technology (Espoo, Finland) on the 11th of January, 2008, at 12 noon.

Helsinki University of Technology
Department of Electrical and Communications Engineering
Laboratory of Computational Engineering

Teknillinen korkeakoulu
Sähkö- ja tietoliikennetekniikan osasto
Laskennallisen tekniikan laboratorio

Distribution:
Helsinki University of Technology
Laboratory of Computational Engineering
http://www.lce.hut.fi
P.O. Box 9203
FI-02015 TKK
FINLAND

Tel. +358-9-451 4834
Fax. +358-9-451 4830
http://www.lce.hut.fi

Online in PDF format: http://lib.hut.fi/Diss/2008/isbn9789512291434/

E-mail: Aapo.Nummenmaa@tkk.fi

ISBN 978-951-22-9142-7 (printed)
ISBN 978-951-22-9143-4 (PDF)
ISSN 1455-0474
PicaSet Oy
Helsinki 2008

Abstract

Magnetoencephalography (MEG) enables noninvasive measurements of cerebral activity with excellent temporal resolution, but localising the neural currents generating the extracranial magnetic fields admits no unique solution. By imposing some mathematical constraints on the currents, reasonable solutions to this electromagnetic inverse problem can be obtained.

In this work, we adopt the statistical formulation of the inverse problem in which the constraints are encoded as Bayesian prior probabilities. The prior is combined with a statistical MEG observation model via Bayes' theorem to yield the posterior probability of the unknown parameters, that is the currents, given the MEG data and modeling assumptions. Apart from the currents, the prior probability density may contain further parameters which are subject to uncertainty. These parameters are not related directly to the MEG observations and are called second-level parameters or hyperparameters, giving the model a hierarchical structure.

The thesis considers hierarchical generalisations of the classical Minimum-Norm and Minimum-Current Estimates (MNE and MCE). The MNE and MCE are distributed source reconstruction methods from which the former is known to produce spatially diffuse distributions and the latter more focal. The here studied extensions of the MNE and MCE prior structures allow more general and flexible modeling of distributed sources with properties in between MNE and MCE.

The first two studies included in this thesis involve more theoretical Bayesian analyses on the properties of the hierarchical distributed source models and the resulting inverse estimates. The latter two studies focus on validation of the models with empirical MEG data, practical analyses and interpretation of the inverse estimates.

Tiivistelmä

Magnetoenkefalografia (MEG) mahdollistaa pään ulkopuolelta tapahtuvan aivotoimintojen mittaamisen hyvällä ajallisella tarkkuudella, mutta nämä magneettikentät synnyttävien aivokudoksen sähkövirtojen paikallistaminen vaatii ns. sähkömagneettisen käänteisongelman ratkaisun, joka ei ole yksikäsitteinen. Jos virtakonfiguraatioille asetetaan sopivia matemaattisia rajoitteita, on kuitenkin mahdollista löytää käyttökelpoisia ratkaisuja tähän käänteisongelmaan.

Tässä työssä käänteisongelmaa lähestytään tilastollisesti, ja matemaattiset rajoitteet muotoillaan Bayesilaisittain *a priori* todennäköisyyksinä. Tämä priorijakauma yhdistetään tilastollisen MEG-havaintomallin kanssa, jolloin saadaan Bayesin teoreeman avulla tuntemattomien parametrien eli virtakonfiguraatioiden *a posteriori* -jakauma, joka kertoo eri virtakonfiguraatioden todennäköisyydet, annettuna havaittu data sekä tehdyt mallioletukset. Virtojen lisäksi priorijakaumaan saattaa liittyä muita tuntemattomia suureita, jotka sisältävät epävarmuutta. Nämä parametrit eivät kytkeydy suoraan MEG-mittauksiin, joten ne ovat siis sähkövirtoihin verrattuna seuraavalla mallitasolla. Näitä priorin parametreja kutsutaan hyperparametreiksi, ja mallilla on hierarkinen rakenne.

Väitöskirjassa tutkitaan klassisten miniminormi- ja minimivirtaestimaattien hierarkisia yleistyksiä. Miniminormi- ja minimivirtaestimaatit ovat lähdejakaumamalleihin liittyviä menetelmiä, joista ensimmäinen tuottaa paikallisesti varsin laajalle levineitä ja jälkimmäinen fokaalimpia käänteisongelman ratkaisuja. Näiden menetelmien tässä työssä tutkitut laajennukset mahdollistavat myös yleisempien ja joustavampien, ominaisuuksiltaan miniminormi- ja minimivirtaoletusten väliin sijoittuvien lähdejakaumien mallintamisen.

Kaksi ensimmäistä osatyötä keskittyvät esitettyjen hierarkisten Bayesilaisten lähdejakaumamallien sekä niiden tuottamien käänteisongelman ratkaisujen teoreettiseen tutkimiseen. Kahdessa jälkimmäisessä osatyössä pyritään validoimaan menetelmät käyttäen mitattua MEG dataa, sekä selventämään näiden hierarkisten käänteisongelman ratkaisujen käytännön merkitystä ja tulkintaa.

List of abbreviations

BEM	Boundary-element model, boundary-element method
BOLD	Blood oxygen level dependent
ECD	Equivalent current dipole
EEG	Electroencephalography, electroencephalogram
fMRI	Functional magnetic resonance imaging/image
hMNE	Hierarchical minimum-norm estimate
KL	Kullback-Leibler
MAP	Maximum *a posteriori*
MCE	Minimum-current estimate
MNE	Minimum-norm estimate
MCMC	Markov chain Monte Carlo
MEG	Magnetoencephalography, magnetoencephalogram
MRI	Magnetic resonance imaging/image
NMR	Nuclear magnetic resonance
VB	Variational Bayesian
PT	Parallel tempering
RJMCMC	Reversible jump Markov chain Monte Carlo
SQUID	Superconducting quantum interference device
SS	Slice sampling
SVD	Singular value decomposition

Preface

Laboratory of Computational Engineering at Helsinki University of Technology has been the fertile ground from which the research presented in this thesis has grown during the years 2002-2007. Financially, the process has been fertilised by Suomen Kulttuurirahasto, ComMIT graduate school, and Academy of Finland via the funding granted to our Professors and Centres of Excellence. My sincere thanks to the abovementioned institutions for making the accomplishment of this thesis possible.

I wish to thank Prof. Mikko Sams for giving me the chance to work in this fascinating field under his supervision. Even though I came to start this research with a rather different background, Mikko's faith in the success of this project was solid from the very beginning. It must be admitted that there were times when yours truly was not just as confident. I also wish to thank Dr. Iiro Jääskeläinen, who introduced me to the latest trends in cognitive neuroscience and played a crucial role in launching the project. I express my humble gratitude to Reverend Lampinen and Brother Vehtari for teaching me the secrets of Bayesianism, and saving me from the heretical statistics and eternal damnation. Simple thanks are just not enough for my closest colleague Dr. Toni Auranen, who should be rather awarded a medal for surviving this scientific odyssey in my hazardous company. All other LCE colleagues, with whom I have had the pleasure to collaborate with officially or unofficially, are thanked as well. Outside our laboratory, I thank Prof. Matti Hämäläinen and Dr. Simo Vanni for teaming up with us and bringing their invaluable external expertise to this joint effort.

Prof. Jari Kaipio and Dr. Fa-Hsuan Lin deserve thanks for acting as the preliminary examiners of this thesis, and providing insightful and encouraging comments on the manuscript. During my time at LCE, I have enjoyed the administrative skills of Prof. Kimmo Kaski, Dr. Kaija Virolainen, and Eeva Lampinen, which I wish to acknowledge. Due to fact that many of the former and present LCE graduate students (and their affiliates) have become also good friends, the boundary between work and off duty has been sometimes rather fuzzy. For these rather fuzzy, hilarious moments I thank (in a random order) Dr. Toni Tamminen,

Dr. Ilkka Kalliomäki, Laura Kauhanen, Dr. Antti Kauhanen, Anne-Mari Seppola, Tommi Nykopp, Dr. Sebastian von Alfthan, Pasi Jylänki, Miika Toivanen, Janne Ojanen, Aatu Kaapro, and, I must admit that I might have missed somebody, oh yes, Tommi Repo.

My brothers and their families are thanked for general support over the years. Special thanks go to Lauri and Eero, who have also had the dubious honor to witness many of my masterstrokes, both in good and bad. Beyond all, I thank my mother Anna Raija, who has provided her unconditional love and conditional financial support during all these years. I will never forget. Finally, I wish to dedicate this thesis to the memory of my father Tapio, who was very much with me in the spirit, all the way.

And now this thesis is yours.

Aapo Nummenmaa

List of publications and author's research contributions

This dissertation consists of an overview and the following publications:

I Auranen, T., Nummenmaa, A., Hämäläinen, M. S., Jääskeläinen, I. P., Lampinen, J., Vehtari, A., and Sams, M. (2005). Bayesian analysis of the neuromagnetic inverse problem with ℓ^p-norm priors. *NeuroImage*, 26(3):870–884.

II Nummenmaa, A., Auranen, T., Hämäläinen, M. S., Jääskeläinen, I. P., Lampinen, J., Sams, M., and Vehtari, A. (2007). Hierarchical Bayesian estimates of distributed MEG sources: Theoretical aspects and comparison of variational and MCMC methods. *NeuroImage*, 35(2):669–685.

III Nummenmaa, A., Auranen, T., Hämäläinen, M. S., Jääskeläinen, I. P., Sams, M., Vehtari, A., and Lampinen, J. (2007). Automatic relevance determination based hierarchical Bayesian MEG inversion in practice. *NeuroImage*, 37(3): 876–889.

IV Nummenmaa, A., Auranen, T., Vanni, S., Hämäläinen, M. S., Jääskeläinen, I. P., Lampinen, J., Vehtari, A., and Sams, M. (2007). Sparse MEG inverse solutions via hierarchical Bayesian modeling: evaluation with a parallel fMRI study. *Helsinki University of Technology, Laboratory of Computational Engineering Publications*, Technical Report B65, ISBN 978-951-22-9141-0.

In all articles, the first author had the principal responsibility of writing and preparing the manuscript. The co-authors contributed by revising the text and suggesting modifications. Overall, the co-authors also contributed to the initiation of performed research and background considerations.

In Publication I, I assisted the first author in performing the analyses and participated in the discussions that formed the background to the article. In Publications II–IV, I had the main responsibility in other matters except for the data

simulation, acquisition, and preprocessing which were mainly done by the second author. In Publication IV, the third author made crucial contributions with respect to the experimental design, data preprocessing and visualisation, and interpretation of the results. Publications I–III have been included in the doctoral dissertation of Toni Auranen.

Contents

Chapter 1

Introduction

1.1 Covering note

The organisation of the material presented in the "overview of research" part of the thesis deviates slightly from the canonical form. For this there are a couple of reasons, first of which was the lack of purpose in duplicating what has already been said in the included research articles. The second is that the subject is rather multidisciplinary – the people behind the presented articles are physicists, engineers, psychologists and medical doctors. It has become very clear that writing for such a wide audience is a huge challenge; people with less mathematical background become frustrated with unfamiliar formulae and we, who consider them as friends are always glad to see one.

This first chapter tries to explain and introduce the basic concepts and ideas verbally with minimal mathematical intervention. The second chapter introduces briefly the models and methods, supplied with (at least some) intuitive explanations and afterthoughts. In third chapter the performed studies are outlined on a general level, taking more of a bird's-eye perspective than trying to scratch the surface on few selected spots. The discussion of chapter four concludes the presentation. The auxiliary material presented here is in a sense complementary to the included articles, and hopefully serves as a multidisciplinary bridge between the vast existing literature and the original research articles.

1.2 Overview

Research on nervous systems ranges from molecular biology of individual neurons to cognitive processes behind conscious experience, and is in general termed neuroscience. In one way or another, most disciplines of this enormous field ask the same question "how does the brain work?", only on different levels of the system. While there is seemingly an endless, impassable road from the microscopic

molecular level neural science to the studies of higher mental processes, some of the gaps have been bridged (or at least circumvented) relatively recently.

For instance, it has become possible to develop genetically modified animals to study how specific mutations affect the functioning of the nerve cells, and furthermore, how this is reflected in the animal's behaviour. At the other end of the neuroscience spectrum, emergence of neuroimaging technologies, most prominently functional Magnetic Resonance Imaging (fMRI), has opened the way for studying activity in the intact, living human brain. Even though the fMRI (in its present basic form) provides an indirect measure of neural activity and is somewhat limited in temporal resolution, being based on the dynamics of the cerebral blood flow, it can produce spatial images of brain activations even without any foreign contrast agents injected to the subject.

Besides fMRI, several other methods are available for noninvasive studies of brain function, all of which have their virtues and limitations. The most traditional of these is electroencephalography (EEG) whose magnetic counterpart magnetoencephalography (MEG) is the method central to the present work. EEG and MEG measure respectively the electric potentials and magnetic fields generated by activity in the nervous tissue. As both methods are based on electromagnetic coupling their temporal resolution is excellent, in the millisecond range. In EEG the electrodes are placed on the scalp, and in MEG the superconducting sensors are embedded in liquid helium inside an insulating helmet. Consequently, neither of these methods is capable of producing direct spatial images of the neural activations, as the measurements are obtained from the outside of the head.

The localisation of the neural currents generating the observed extracranial fields is tantamount to solving the electromagnetic inverse problem, which has no unique solution. With a given measurement and conductor (head) geometry, it is possible to arrange currents to configurations in which their electromagnetic fields mutually cancel out causing the MEG and EEG measurements to vanish. Such a current configuration can be added to the "solution" of the inverse problem and the new solution will generate identical MEG and EEG measurements; hence the nonuniqueness.

By imposing suitable constraints on the solutions, reasonable reconstructions of the neural sources based on EEG and MEG can be obtained. In this work, we will take the route of encoding these constraints as Bayesian *a priori* probabilities on the space of possible current configurations. The Bayesian formulation casts the inverse problem to the realm of statistical inference. The prior probabilities are combined with the likelihood which quantifies how likely different current configurations would generate the obtained data. The resulting posterior probability distribution tells how probable each of the candidate solutions is, given the data and our modeling assumptions, and it is our "solution" to the inverse problem. In principle, the Bayesian framework provides a coherent way to incorporate any kind of prior information, anatomical or physiological, to the electromagnetic

inverse problem.

While looking perfectly foolproof theoretically, the Bayesian approach has its own obstacles. Calculation of the posterior distribution in practice is a complicated mathematical task, analytically intractable for all but the simplest models. Several numerical approximation methods have been developed for this matter – from these stochastic Markov chain Monte Carlo (MCMC) and deterministic Variational Bayesian (VB) methods are studied in this thesis. Furthermore, the inference is always conditioned on some more or less arbitrary modeling assumptions, and how sensitive the inferences are in this respect is a crucial question from the viewpoint of interpretation of the results.

The established researcher with years of experience in practical MEG analysis will now ask what else is new. In less laconic terms, what are the motives for the studies presented here, and how are the methods different from the standards routinely used in hundreds of MEG and EEG laboratories. Firstly, the increasing computational resources and methodological development in statistical modeling enable the application of more complex (and hopefully also more realistic) models to the electromagnetic inverse problem. The more complex models also necessitate more careful validation and theoretical inspection, and above all, demonstration of practical utility and usability before getting foothold among the experimental scientists. Secondly, the increase in data collection rate puts a heavy load on developing software and efficient data analysis procedures which can be automatised and handle large subject populations, and for which the analyses can be easily replicated if necessary.

Here we have studied hierarchical Bayesian generalisations of models presently used in practical MEG analysis. To obtain a realistic view on the properties of the models and methods, we have used both empirical data and simulations for validation. We have tried to be objective also in pinpointing the possible weaknesses and potential problems in the methods, for these are really the forces that drive the methodological development forward.

1.3 General references

As the studies involve several measurement techniques, statistical models and computational methods, the space for introductory material per topic is limited. Fortunately, excellent textbooks and review articles exist for reference. The exception is the last section which contains a survey of the recent Bayesian MEG inverse literature and is the only topic to which the author has contributed scientifically. Therefore, I here list the "textbook-level" general references with the purpose of increasing the readability of the parsimonious presentation, and in a sense to acknowledge the sources from which the author has absorbed much of this chapter's material over the years. Explicit references are provided for those

matters which are considered to be nonelementary or more difficult to find in the literature.

The definite reference for general neuroscience is (Kandel et al., 2000). Purves et al. (2001) give a condensed presentation covering approximately the same topics. Modeling of neural systems on various levels is presented by Dayan and Abbot (2001), which may help the mathematically oriented minds to get a more concrete grip of the subject. MEG is thoroughly exposed by Hämäläinen et al. (1993) and various aspects of MRI and fMRI by Jezzard et al. (2001) and Huettel et al. (2004). Foundations of Bayesian statistics are developed in detail by Bernardo and Smith (2000) and practical matters of Bayesian data analysis along with numerical sampling methods by Gelman et al. (2003). Introduction to Variational Bayesian methods can be found from (Lappalainen and Miskin, 2000) and (Opper and Saad, 2001, especially chapters 10 and 11). Kaipio and Somersalo (2005) provide an up-to-date view on inverse problems, with special emphasis on statistical and computational methods and practical examples. Finally, the more established MEG and EEG inverse methods are reviewed by Baillet et al. (2001).

1.4 Electromagnetic measures of brain activity

1.4.1 Origins of the MEG signal

Main cellular constituents of the brain are neurons and glial cells; neurons are the information processing units while glia mainly provide metabolic and structural support. Neurons typically consist of a soma (cell body), dendrites (receive and integrate information from other neurons), and an axon (sends information to other neurons). Broadly speaking, information is transferred within a neuron electrically and between neurons chemically.

The neuron is surrounded by a membrane, across which an electric potential is maintained. If the magnitude of this potential exceeds a certain threshold near the beginning of the axon, the neuron "fires" an action potential which travels along the axon like a localised wave. The action potential eventually reaches a terminal of the axon localised near a dendrite or soma of another neuron. The junction is called a synapse between the two neurons and they are termed presynaptic and postsynaptic neurons in this respect.

The action potential triggers a release of a neurotransmitter substance to the synaptic space between the neurons, which conveys the information about the action potential to the postsynaptic neuron chemically. The neurotransmitter binds to the membrane of the postsynaptic neuron and alters its biochemical properties. As a consequence, the membrane voltage is changed, a process which is called a postsynaptic potential. The postsynaptic potential causes an electric field, and a flow of current into (or out of) the cell (depending on the sign of the potential). This current flows along the dendrite decaying exponentially. Thus, looking at a

distance, the postsynaptic potential looks like a small current dipole with a source and a sink. When these dipolar currents suitably sum up, the resulting electromagnetic field can be measured even from the surface or outside of the head. The intracellular dipolar currents are called primary currents. Because current cannot accumulate anywhere, secondary or volume currents flow in the extracellular space to complete the loop.

For a cohesive summation of the dipoles caused by individual postsynaptic potentials, these must be temporally synchronised and they must be oriented in a parallel fashion. The cortex has indeed a laminar structure, and the main dendrites of the neurons are predominantly oriented perpendicular to the cortical surface. In rough numbers, tens of square millimeters of cortex must be activated to elicit a current dipole of 10 nAm, which corresponds to field strengths typically observed in MEG. So, the MEG measurement reflects the "summed" or "average" synaptic activity of a rather large population of neurons.

Why then, we do not see the action potentials themselves with MEG? The traveling wave potential, under suitable circumstances, can be modeled with two oppositely oriented dipoles, that is a quadrupole. As a function of distance r, the quadrupole field decays as $1/r^3$ whereas dipole field decays as $1/r^2$. Hence, at the slower timescales of the postsynaptic potentials, the dipole field dominates at large distances. Whether the action potentials manifest themselves in the high frequency content of extracranial brain signals is under debate.

1.4.2 The MEG device

The neuromagnetic fields are extremely small in comparison with the earth's magnetic field; approximately by a factor of 10^{-9}. Consequently, the MEG measurement system involves several components of high technological sophistication. The detection of the tiny fields is based on Superconducting QUantum Interference Device (SQUID) sensors. To maintain the SQUIDs in superconducting state, they are embedded in liquid helium. The SQUIDs are coupled to the external magnetic field via superconducting flux transformers of different types. Gradiometers, for instance, are sensitive to spatial changes in the magnetic field but insensitive to a static homogeneous field, thereby suppressing effects of spatially slowly varying external disturbances. In the modern multichannel magnetometer devices, the whole head is covered with sensors arranged inside a helmet-shaped dewar.

The sensors must be also shielded from external interference, and the MEG device is usually located inside a room constructed from several layers of aluminum and μ-metal to obtain an effective shielding over wide range of frequencies. Inevitably, the recorded MEG signal will always contain some noise: the human heart produces a magnetic signal, the thermal noise of the dewar, geomagnetic distortions and so forth. Eye movements, blinks and movements of magnetic objects attached to the subject can produce artefactual components to the recorded

signal. Some of the noise sources can be removed afterwards or online by signal processing techniques, but good quality of the recorded "raw" data can never be overemphasised.

1.4.3 Differences between EEG and MEG

Even though EEG measures in theory exactly the same postsynaptic currents as MEG, there are differences in practice. As discussed in the context of the forward problem (see below), MEG is not very sensitive to dipoles with certain orientations, whereas EEG in principle is sensitive to all primary current components. The skull is a good electric insulator but magnetically transparent, and consequently, creating a good conductor model necessary for solving the forward problem (see below) is more difficult for EEG than MEG. Instrumentation of EEG is much less expensive as it is based on relatively simple electric potential measurements. In what follows we will consider MEG solely; however, all inverse methods and models discussed apply in principle also to EEG. It is possible to measure EEG simultaneously with MEG and theoretically this would give the best results as they provide complementary measures of the same neural currents.

1.4.4 The forward problem

Calculating the external magnetic fields generated by a primary current dipole located inside the volume conductor (*i.e.*, the head) is called the forward problem of MEG. While solving the forward problem is straightforward in principle, via quasistatic Maxwell's equations, analytical solutions exist only for special cases and otherwise numerical strategies must be employed. For the "spherical head" model analytical solution is available, along with some important consequences. The most important of these is that primary current components oriented along the radius of the sphere do not generate magnetic fields outside the head. Thus, MEG is sensitive mostly[1] to the currents located in the walls of the cortical fissures. More realistic conductor models can be obtained with aid of additional anatomical information. In this study, we use a boundary-element model (BEM) in which bounding surfaces between different tissue types are mathematically constructed, based on anatomical MRIs. For MEG, the boundary of the inner skull is typically sufficient, assuming the inside a homogeneous conductor and the outside a perfect insulator.

Whatever numerical strategy is adopted for the solution of the forward problem, the Maxwell's equations are linear and the principle of superposition applies. That is, if we know the magnetic fields of dipoles with unit magnitude at the locations of interest, fields of the current configurations with all other amplitudes can

[1] Exactly spherical heads are rather rare.

be obtained as linear combinations of these. So, if we divide the volume of interest to N grid points, and we know the locations of the M sensors with respect to the head, the solution of the forward problem at time t can be written as a matrix equation:

$$\boldsymbol{B}(t) = \boldsymbol{G}\boldsymbol{J}(t). \tag{1.1}$$

The matrix $\boldsymbol{G}$ is called the *gain matrix*, and its element $\boldsymbol{G}_{ij}$ is equal to the magnetic field recorded with sensor i, if a current dipole of unit magnitude and fixed orientation is placed at grid point j. The forward equation (1.1) is merely a manifestation of the superposition principle: the vector of obtained MEG measurements $\boldsymbol{B}(t)$ depends linearly on the vector of the current amplitudes $\boldsymbol{J}(t)$ through the gain matrix $\boldsymbol{G}$. The gain matrix thus contains all information about the geometry of the sensor array and the conductivity properties of the head, and it is the output of the forward computation. The gain matrix is usually assumed to stay constant over time, and thus is computed only once. Physically this means that the positions of the sensors with respect to the head should remain constant. Minimising subject's head movement during the MEG recording is thus very important.

It is noteworthy that the measured fields depend linearly only on the amplitudes of the dipoles, not on their positions or orientations. If we constrain the possible locations of the source dipoles to the cortical sheet, we can also utilise the anatomical information in constraining the orientations of the dipoles perpendicular to this surface. If anatomical information is not available, unit dipoles for the three dimensions of the cartesian space can be considered for each grid point location.

1.4.5 The inverse problem

Suppose now that we are given a set of observed MEG measurements, $\boldsymbol{B}_{obs}(t), t = 1, \ldots, T$. The inverse problem is to solve the matching currents $\boldsymbol{J}_{sol}(t), t = 1, \ldots, T$, so that these together satisfy equation (1.1). As already mentioned, the inverse problem admits no unique solution. Before showing how this stems from equation (1.1), let us consider the basic taxonomy of the inverse solution strategies; there are two main approaches to the neuromagnetic inverse problem. For simplicity, we consider a single timepoint of the MEG data.

The multidipole modeling approach tries to explain the observed fields with few equivalent current dipoles (ECDs). Now, one dipole is characterised by position (three coordinates), direction (two coordinates) and amplitude (one coordinate), that is altogether six parameters. Recalling that the observed fields depend nonlinearly on the position and orientation variables of the dipole, finding the ECD that best explains the measured fields is a nonlinear optimisation problem in six-dimensional parameter space. For a single dipole this can be efficiently solved,

but with increasing number of dipoles the optimisation becomes very hard due to the combinatorial explosion of local optima and the increasing dimensionality of the parameter space. One problem is also that the number of the ECDs needed is not known beforehand.

In any case, a modern multichannel MEG device provides hundreds of channels, say 306, while one dipole contains six parameters. So, as long as we are fitting less than fifty dipoles to the data, there are more equations (channels) than unknowns (dipole parameters), and in this respect the multidipole fitting is considered an overdetermined inverse problem. With more equations than unknowns, it is possible that a "perfect" solution does not exist. If in reality the data were generated by two dipoles, and we are trying to find one ECD, it will explain only part of the observed data, no matter how hard we optimise the dipole parameters.

On the other hand, forming a discrete grid to cover the space of possible sources, we must only estimate the amplitude of the current at each location, which seems more of a linear problem. To cover each potentially activated brain location with reasonable accuracy, several thousands of grid points are needed. In the case of unknown dipole orientation, three amplitude parameters (or more properly, current components) are associated with each gridpoint, leading to the number of unknowns being more than tenfold in comparison with the number of equations. This approach is termed distributed source modeling, as a distribution of the source current throughout the brain is searched. Because now the number of unknowns is vast in comparison with the number of equations, the distributed source estimation is an underdetermined inverse problem.

Now we are in position to look at the basic properties of the distributed inverse problem under study in this thesis. For simplicity, let us assume that we know the orientations of the dipoles at each of the N grid points and that we still consider a single timepoint (index t is then dropped from the notation). Now we have M MEG channels and the gain matrix $\boldsymbol{G}$ is of dimension $M \times N$, where $N \gg M$. The rows of the gain matrix, denoted by $\boldsymbol{\mathcal{L}}_i$, are called lead fields, and there is one for each sensor. The lead field tells how each unit dipole of the grid shows up in that sensor – it is actually a distributed current configuration reflecting how sensitive the sensor is to each grid location. All MEG measurements in accord with our forward model can be explained by a linear combination of the lead field current configurations. Not all MEG channels are necessarily linearly independent; the number of linearly independent lead fields is equal to the dimension of the range space of the gain matrix $\boldsymbol{G}$, and we denote it by $\dim(\boldsymbol{\mathcal{L}})$; this may actually be smaller than M. Let us then define the space of magnetically silent current configurations:

$$\boldsymbol{\mathcal{N}} = \{\boldsymbol{J} | \boldsymbol{G}\boldsymbol{J} = \boldsymbol{0}\}\,. \tag{1.2}$$

Mathematical term for the space $\boldsymbol{\mathcal{N}}$ is the null space of gain matrix $\boldsymbol{G}$, let us call its dimension $\dim(\boldsymbol{\mathcal{N}})$.

Now, an arbitrary current $\boldsymbol{J}$ can be decomposed into components which either 1) produce observable MEG fields or 2) are magnetically silent. The dimension of the subspace for components of 1) is $\dim(\boldsymbol{\mathcal{L}})$ and for components of 2) it is $\dim(\boldsymbol{\mathcal{N}})$. Because the dimension of all possible source configurations is N, we have the following:[2]

$$\dim(\boldsymbol{\mathcal{L}}) + \dim(\boldsymbol{\mathcal{N}}) = N. \tag{1.3}$$

As the number of linearly independent lead fields is always smaller or equal to the number of MEG channels M, we get

$$\dim(\boldsymbol{\mathcal{N}}) \geq N - M. \tag{1.4}$$

In concrete numbers, assuming a grid of 3000 source points and 306 MEG channels yields an estimate for the dimension of the magnetically silent current configurations: $\dim(\boldsymbol{\mathcal{N}}) \geq 2694$. It turns out that a vast majority of all possible current configurations belong to the magnetically silent subspace! Any such current, or a linear combination of these, can always be added to our "inverse solution", and the corresponding MEG measurements remain the same. Even though it seems awkward that the currents can twist and turn in this overwhelmingly high-dimensional space at will, the real moral of the story is that with 306 measurements there are 306 things to learn from the data. Thus, in order to obtain sensible inverse estimates, we must limit the space of possible solutions in some suitable manner. The manner in which this is done, is from a mathematical viewpoint largely a matter of taste, but of course the physiological plausibility of the solutions plays a crucial role.

The simplest way to obtain a solution which is well-defined and unique, would be to formulate it as a linear combination of the lead fields, and find the coefficients which best match the observed MEG data. Because there is one lead field for each sensor and these generate the space of possible measurements, we would get a unique representation of the solution in this basis, if the lead fields are linearly independent. A slightly modified version of this procedure, including effects of measurement noise, is widely used in practical MEG analysis. The lead fields are, however, spatially rather diffuse as the distances from the sources to the sensors are typically centimeters. In other words, a MEG sensor picks up signal over a large area of cortex. Hence, estimates of this type also easily spread over several distinct anatomical structures, even if it would be physiologically more plausible to assume that the MEG response is generated by a spatially well-defined area. In search of ways to obtain additional anatomical and physiological information, we turn to the magnetic resonance imaging.

[2] This is the basic rank-nullity theorem for matrices, which can be found, for instance, from (Harville, 1999, p. 585).

1.5 Magnetic resonance imaging

1.5.1 Physical principles

Atomic nuclei with odd mass number have a nonzero intrinsic quantum-mechanical angular momentum called spin. The angular momentum gives the nucleus a magnetic moment, which is proportional to the spin, the constant of proportionality being called the gyromagnetic ratio. For sake of concreteness, let us consider a hydrogen nucleus, that is a proton. Hydrogen is the most abundant chemical element and constituent of water, and hence plays a central role in many magnetic resonance studies.

When placed in an external homogeneous static magnetic field, quantum theory dictates that the component of the spin magnetic moment along the axis of the external field has two possible states, either parallel or anti-parallel. These states have different energies, the parallel being the ground state, and the energy difference increases linearly with respect to the strength of the magnetic field. The most efficient excitation of the spins from the ground state to the antiparallel state, by electromagnetic radiation, happens when the energy of the quanta of this radiation exactly matches the difference of the energy states. This resonance frequency is also called Larmor frequency and it is equal to the product of the gyromagnetic ratio and the external field strength (divided by 2π). The phenomenon of Nuclear Magnetic Resonance (NMR) thus leads to a resonance frequency peak in the spectrum of electromagnetic radiation emitted by the substance when the excited spin(s) return to the ground state.

Suppose now that we have a population of protons inside a sample volume. In room temperature and small external static fields, roughly half of the protons will be spontaneously excited to the antiparallel state of higher energy, because the energy difference of the two states is small in comparison with the thermal energy. However, if the external field is so strong that the spontaneous transition becomes sufficiently improbable, there will be a macroscopic net magnetisation, because more protons will stay in the ground state. The direction of the net magnetisation vector is along the static magnetic field in equilibrium. Now, if a suitable time-varying electromagnetic field is applied to the sample at the resonant frequency, the magnetic moment vector will be knocked out of its equilibrium position, eventually returning back. This process can be detected from the electromagnetic field which the moving magnetisation vector generates. The resonant frequency for hydrogen is in the radiofrequency band, about 42 MHz/Tesla and therefore the excitation pulses are referred to as rf-pulses.

1.5.2 Image formation

Even though the basic physical phenomenon of NMR was experimentally studied already in the mid 40's, the first NMR-based images were obtained in the early 70's. How the NMR signal could be manipulated to provide spatial information and how to form images from the resulting data are indeed nontrivial problems; the Nobel Prize in Medicine 2003 was awarded partly for this work. The encoding is actually done by superimposing spatially varying magnetic fields on top of the large homogeneous field. These gradient fields have the effect that the Larmor frequency will be altered for different spatial locations in the object. By applying the rf-pulses and gradient fields in suitable temporal sequences, information from different spatial frequencies of the object can be acquired, and the resulting data comprises the Fourier transformation of the object. The direct spatial image is then obtained computationally by the inverse Fourier transform.

1.5.3 Contrast mechanisms

The 3D imaging object such as the human brain is divided into volume elements or voxels, the voxels in a 2D plane or "slice" corresponding to the basic images from which 3D object is reconstructed. For each voxel we can imagine an individual macroscopic magnetisation vector, which in equilibrium points to the direction of the static main field. When the longitudal magnetisation vector is tipped from the equilibrium position with a suitable rf-pulse, it will have also a component perpendicular to the main magnetic field. This component is called transversal magnetisation vector.

How these components of the magnetisation return to the equilibrium state is phenomenologically described by two time constants, (longitudal) T_1 and (transversal) T_2. The constant T_1 stems from how quickly the microscopic spins (protons) return back to the ground state due to the interactions with their local environment. The value of T_2 depends upon how much the spins interact with each other. The upshot is that for different substances such as biological tissues, these values are different and thus reflected in the contrast values of the MRIs obtained, for instance, from the brain.

Along with the basic proton density, the T_1 and T_2 contrasts are always present in the MRIs but with manipulation of the pulse sequence used in the image collection a particular contrast can be enhanced for visualisation of specific tissues. For instance, T_1-weighted images portray the spatial distribution of T_1 values in the brain, and voxels with short T_1 (containing, *e.g.*, axons) will appear bright, and those voxels with long (containing, *e.g.*, cerebrospinal fluid) will appear dark. There are many other contrast mechanisms; dynamic contrasts which can give information about diffusion or perfusion of water in the brain and so forth.

1.5.4 Functional MRI

In the early 90's, the possibility of obtaining MRIs which somehow reflect the activity of the brain was demonstrated. The T_2 relaxation constant is affected by small local disturbances to the external magnetic field – in presence of such a disturbance, the transversal magnetisation will decay faster and the contrast is called T_2^*. The hypothesis was, that an activation of some local area of the brain is accompanied by increased regional blood flow to meet the metabolic demands.

The first fMRIs were obtained by using a magnetically susceptible contrast agent which, after injection to the bloodstream, would flow in larger proportions to the activated area altering its T_2^*-value. It was already well known that depending whether oxygen is bound to haemoglobin or not, it has different magnetic susceptibility. Soon after the fMRI studies with an external contrast agent, the first Blood Oxygen Level Dependent (BOLD) fMRIs were obtained. The dependence of the BOLD fMRI signal only on the endogenous balance of oxyhaemoglobin vs. deoxyhaemoglobin, made the technique highly noninvasive and popular. A crucial preliminary stage for making fMRI feasible was development of echo-planar imaging, a MRI protocol which allows a rapid acquisition of the images. This work was also in part awarded in the 2003 Nobel Prize.

The BOLD contrast does not directly measure the electric activity of the neurons, but relies on changes in regional cerebral blood flow and oxygen consumption, which is generally termed haemodynamics. The fMRI-haemodynamic response following stimulus-triggered activity lasts from ten to twenty seconds, and shows significant variability over brain regions and between subjects. BOLD signal can be generated by large blood vessels which drain the activated area, but may still be located relatively far away. Despite its limitations, fMRI provides direct spatial information about brain activity in a noninvasive way, and as such is an invaluable method for experimental brain research.

1.6 Bayesian data analysis

1.6.1 Basics of inference

Suppose we have a set of observed data $\mathcal{D}$, which could be a functional or anatomical magnetic resonance image, a vector of MEG fields at a timepoint, or something completely different. In the dataset there is usually a component we are interested in, the "signal", reflecting the underlying phenomenon which the data collection procedure was designed to reveal. There is also some variability in the data that we consider unrelated to the phenomenon of interest, and is here called "noise". To distinguish signal from noise, we must make assumptions about the process which generates the data.

In this presentation we focus on parametric generative models, that is, we set

up a mathematical model which quantifies how the data are generated as a function of a set of some unknown parameters, denoted by vector $\boldsymbol{\theta}$. In the Bayesian approach, all variables, data and parameters, are considered to be random or stochastic by nature. This merely expresses the philosophical or practical attitude that observables or unknown parameters are always subject to some uncertainty. Of course, the models can contain constants or variables, which are considered to be known for all practical purposes. The data variables are singled out only in that they are fixed to their observed values in the process of inference. The goal of the Bayesian statistical inference is to obtain the conditional probabilities of different parameter values given the observed data. One could then ask questions such as what is the expected value of this parameter given the data, how much uncertainty there is in the expected value and so forth.

The first task is to set up a law which quantifies the probability of obtaining a set of data $\mathcal{D}$ if the parameter value was $\boldsymbol{\theta}$. This conditional probability is denoted by $p(\mathcal{D}|\boldsymbol{\theta}, \mathcal{M})$, and when looked as a function of the parameters it is called the likelihood. We note that constructing the likelihood is quite similar in nature to the forward problem, which asks "if we know the currents inside the conductor, what are the MEG measurements it generates". The symbol $\mathcal{M}$ is introduced to remind that the whole probability model is also subject to the specific assumptions that we make about the data generation process. Because everything is conditioned on $\mathcal{M}$, it is a redundant symbol and as such often omitted, but we keep it throughout the presentation for its pedagogic value.

Based on our knowledge of the phenomenon under study, we might have some idea about parameter values which are more probable, perhaps by related studies, physical constraints or otherwise. We call the mathematical formulation of this information the prior probability of the parameters, and it is denoted by $p(\boldsymbol{\theta}|\mathcal{M})$. According to its name, it reflects our beliefs and knowledge about the parameter values prior to obtaining the data. It is precisely this prior probability density which will provide the mechanism to limit the space of possible solutions or currents, which we found inevitable in the context of the underdetermined MEG inverse problem.

By using basic probability calculus, we may then derive the conditional probability of the parameters after observing the data, descriptively named the posterior distribution $p(\boldsymbol{\theta}|\mathcal{D}, \mathcal{M})$:

$$p(\boldsymbol{\theta}|\mathcal{D}, \mathcal{M}) = \frac{p(\mathcal{D}|\boldsymbol{\theta}, \mathcal{M})\, p(\boldsymbol{\theta}|\mathcal{M})}{p(\mathcal{D}|\mathcal{M})}. \tag{1.5}$$

This is called Bayes' theorem and it is the cornerstone of Bayesian inference. The denominator $p(\mathcal{D}|\mathcal{M})$ does not depend on the parameter values but is the important normalising constant of the posterior, taking care that the posterior probabili-

ties sum up to unity:

$$p(\mathcal{D}|\mathcal{M}) = \int p(\mathcal{D}|\boldsymbol{\theta}, \mathcal{M})\, p(\boldsymbol{\theta}|\mathcal{M})\, \mathrm{d}\boldsymbol{\theta}. \tag{1.6}$$

Being equal to the conditional probability that the data come from the model $\mathcal{M}$ given our modeling assumptions, the normalising constant is called the marginal likelihood or evidence for model $\mathcal{M}$. The evidence can be used to perform Bayesian model selection or model averaging, under certain circumstances. The posterior distribution contains all information about parameter $\boldsymbol{\theta}$ conveyed from the data $\mathcal{D}$ by our model $\mathcal{M}$.

There is no fundamental distinction between the prior and the posterior. As discussed, we might construct the prior based on some previous experiments or related studies, combine this with the constructed likelihood after obtaining the data, and calculate the posterior distribution, which essentially is a product of these two. Hence, the data "shift" the prior to the direction of the observed data, giving the posterior distribution. Now, if we acquire more data, we can use the old posterior as the "prior" for the new data, combine it with the likelihood by the Bayes' rule and obtain an updated posterior. This coherent way of incorporating new data is one of the most attractive aspects of Bayesian data analysis.

Unfortunately, there is no universal recipe for how to construct a good generative model for the data, which is absolutely critical to obtain reliable results. Sometimes parts of the data generation mechanism are based on known physical principles, as is the case for the MEG inverse analysis, where the forward equation (1.1) dictates how the MEG data are generated by the cortical electric currents. To specify the likelihood, we assume that the real MEG observation is a sum of the forward calculated fields and random Gaussian noise. For the prior, there are various more or less obvious candidates, some of which will be introduced in the next chapter.

Each of the models produce posterior estimates for the currents with different properties, they can all be theoretically justified, and none of them is the "correct". Still, it may and will turn out that some of the models are more useful in practice. In any case, it should be by now evident that the Bayesian framework is a very natural choice for the MEG inverse analysis. The prior constrains the solutions, likelihood incorporates the forward model and gives the probability of the MEG data given the currents, and the posterior yields the "converse" probability of the currents given the data, which is our inverse solution.

1.6.2 Hierarchical models

The true power of the Bayesian approach comes from the possibility to construct and analyse hierarchical models. It might be the case that we feel uncertain about our specification of the prior, and would like to include some unknown parameters

to express this. Adding more parameters to a model is always a risk, as it makes the model more flexible but also prone to overfitting, which means including effects of noise to our estimates of parameters that we think should only be related to the signal. In hierarchical modeling, we introduce unknown parameters to the prior, so that the prior is now expressed conditional on these parameter values.

For concreteness, let us introduce a scalar parameter ϕ to the prior in this way: $p(\boldsymbol{\theta}|\phi, \mathcal{M})$. The parameters of the prior are called hyperparameters, as they are on a subsequent level with respect to the data (the parameters appear also in the likelihood, whereas the hyperparameters do not). The hyperparameter then has its own prior distribution which is suggestively called a hyperprior, and denoted by $p(\phi|\mathcal{M})$. The increase of flexibility without overfitting is possible because in the hierarchical model, the hyperparameter ϕ binds different parameter components θ_i of the vector $\boldsymbol{\theta}$ rather than being a completely separate and independent parameter.

Analysis of the hierarchical models proceeds in theory exactly as for nonhierarchical models. We can find the joint posterior distribution of the parameters and hyperparameters, and marginalise (integrate over) the hyperparameters to obtain the posterior of the parameters only, if the hyperparameters are considered to be auxiliary or nuisance parameters. Even if the introduction of the hyperparameters followed by their immediate marginalisation seems at first like futile juggling, this is not the case – the uncertainty about the hyperparameter values is now included in the marginal distribution of the parameters.

In theory, the number of possible levels of hierarchy is not limited in any way. That is, the hyperprior might also contain some unknown parameters, and we could impose a further prior on these. But the hyperparameters depend on the data only via the parameters: given the values of the parameters, the data and hyperparameters are independent (Goel and Degroot, 1981; Dawid, 1979). Therefore the information about the hyperparameter values comes from the data through the parameters, and it can be shown (for certain models) that this information becomes increasingly vague the higher the parameters lie in the hierarchy (Goel and Degroot, 1981).

Intuitively it is also clear that only as long as there are several exchangeable parameters under the prior at a certain level, the data can tell us something useful about the parameters of this distribution. Exchangeability of a certain set of parameters means that their joint distribution is invariant to any permutation of these parameters (see, *e.g.*, Bernardo and Smith, 2000, pp. 167–172). For instance, if the θ_i's are exchangeable, and there are $N > 1$ such parameters, the posterior distribution of the parameter ϕ could be in principle determined from the data. But if we were to impose a further prior for ϕ with parameter ξ, we would have in a sense only one "degree of freedom" to which we would base our inference on ξ.

In order to complete the model specification, one must assume fixed values to the parameters of the ultimate prior or to choose this prior to be uniform or uninformative in some sense, so that it does not contain any further parameters.

How sensitive the inferences are on this ultimate prior is an important question, and will be of interest in the MEG inverse models studied in this thesis.

1.7 Computational methods

1.7.1 Motivation

Even though theoretically simple and sound, Bayesian data-analysis has practical challenges. The first is calculating the posterior distribution itself: the normalisation factor is an integral over the parameter space. If the parameter vector was the current vector of distributed MEG currents, the dimension would be several thousands, let us say modestly $N = 1000$. If the normalisation constant (1.6) is not analytically solvable, it has to be calculated numerically. The most straightforward way to accomplish this would be to discretise each variable into k possible values, and transform the continuous integral into a sum. The number of terms in this sum would be k^N, for two possible values ($k = 2$) this being already $2^{1000} \approx 10^{301}$. With a supercomputer, performing 10^{13} floating point operations per second, it would take 10^{288} seconds or about 10^{280} years to evaluate the sum if one floating point operation per term would be required. In comparison, estimated age of the universe is 10^{10} years.

The explosion of the sum terms is exponential as a function of the dimension of the parameter space N, and more efficient ways to calculate the integral are needed. Even if we could analytically solve the normalisation constant and obtain the posterior, it would be a function of thousands of variables and difficult to visualise and handle in the general case. Some posterior summary quantities, such as the posterior expectation value of the currents are desired, leading to other high dimensional integrals.

One quantity which is easier to obtain is the parameter values which (locally or globally) maximise the posterior probability density, since this can be calculated without knowing the normalisation constant which does not depend on the parameter values. This is called the Maximum A Posteriori (MAP) estimate, and it can be obtained by optimisation techniques, but the MAP-estimate does not reflect in any way how the posterior probability mass is distributed around this maximum. In the following we introduce the computational methods used in this thesis to evaluate these high-dimensional integrals.

1.7.2 Markov chain Monte Carlo

In Markov chain Monte Carlo (MCMC) the evaluation of high dimensional integrals is performed implicitly. Markov chain is a stochastic process for which

the probability[3] of the next state depends only on the last of the previous states, whereas Monte Carlo refers to a wide class of computational methods in which solution to a problem is obtained by simulating a stochastic system. The MCMC construction is such that the equilibrium or stationary distribution of the Markov chain is the distribution of interest, the posterior distribution in our case. After the chain "forgets its initial state" or converges to the stationary distribution, the states of the Markov chain are distributed according to the target distribution, and we have numerical samples from the posterior. From these samples, any quantity of interest can be obtained: the posterior mean can be obtained just as a sample average and so forth.

Tailoring a suitable Markov chain for a given problem is an art in itself. There are several alternatives, all which give asymptotically correct results, but can perform very differently in practical situations with finite computational resources. Most of the methods have a proposal distribution, from which we know how obtain numerical samples, for example, a Gaussian. This has usually the Markovian property that the new candidate state is randomly proposed based on the previous state only. If the proposal distribution is symmetric meaning it is equally probable to "jump" from the old state to the new and vice versa, the procedure is particularly simple. First, the target distribution (posterior) is evaluated at the new proposed state and the old state and their probability ratio is calculated. If the ratio is greater than one, that is the new state has higher posterior probability than the old, the new state is accepted. If the proposed state has lower probability and the ratio thus is smaller than one, it is accepted with probability equal to this ratio. If the new state is rejected, the new state is set to be the old state, and a new candidate state is proposed.

The basic algorithm just described is called the Metropolis algorithm according to its inventors, and the generalisation to nonsymmetric proposal distribution is called Metropolis-Hastings. In cases where the conditional posteriors of separate parameter components given the others can be directly sampled, these conditional distributions can be used as "proposal distributions" and the method is called Gibbs sampling; with Gibbs sampling the proposed states are always accepted. With Reversible Jump MCMC (RJMCMC), models with unknown parameter dimensionality can be handled. These can be mixed in all manners, for instance in our hypothetical hierarchical model, the hyperparameter ϕ could perhaps be sampled with Metropolis and the parameters $\boldsymbol{\theta}$ given the hyperparameter with Gibbs and so on.

The efficiency of MCMC over the brute force "uniform" numerical integration comes from the fact that the chain drifts towards those regions of the parameter space where the posterior probability is high (with a symmetric proposal distribution, jumps to more probable parameter states are always accepted). No

[3] In this informal treatment we use the word probability for discrete and continuous densities.

time is wasted for the regions where the probability is zero anyways, contributing nothing to the integrals. A key feature of Metropolis-type algorithms is that the acceptance/rejection calculation is based on posterior probability ratios and the normalising constant does not need to be known as it cancels out. Even though we can calculate in principle arbitrary expectations and marginal distributions over the posterior with the numerical samples, we cannot directly calculate the normalisation constant based on these, because it is a different type of integral[4]. There are suitable sampling techniques for this purpose also, but they tend to be computationally very intensive.

Of course, many difficulties pertain to the MCMC approach. Establishing the convergence of the chain, which means determining when the states begin to be true representatives of the posterior, is a nontrivial problem. The convergence is not deterministic, and it can be a very slow process. If the consecutive states of the Markov chain are correlated, the number of independent posterior samples obtained is much smaller than the number of MCMC iterations. This can happen, for instance, if the proposal distribution is badly chosen – then there can be either too many rejections, and the chain will stay in the same state for several steps, or too few rejections, and the chain will explore the parameter space in too small steps, showing up as slow drifts in the trends of the chain. If the posterior distribution is multimodal, that is, there are many separate areas of high probability, the chain can have difficulties in jumping between the modes. This is especially true if the areas of high probability are separated with areas of extremely low probability, because jumps to parameter states of smaller probability are also accepted with a smaller probability. In this case the samples may not represent the posterior in full.

1.7.3 Variational Bayes

Because MCMC tends to be computationally heavy, different analytical approximation methods have also been developed for performing Bayesian inference. Whereas MCMC represents the intractable posterior by numerical samples, in Variational Bayesian methods a simpler, analytically tractable approximate for the true posterior is searched. Similar approximations have been used in problems of statistical physics, where the term mean-field method is used. The class of variational approximations is rather diverse, but most of the methods still follow roughly the same logic.

First, we must define the form of the approximative distribution in some sense. We could say, for instance, that we want a Gaussian approximation for the posterior, as we know how to calculate various statistics for Gaussian distributions. These kinds of approximations are called fixed-form. If the full model belongs

[4] A different sort of Monte Carlo estimate for the evidence can be calculated from the MCMC samples, but it is generally unstable (see, *e.g.*, Kass and Raftery, 1995).

to a suitable subclass of models, notably to the conjugate-exponential family, we may obtain a tractable approximation by assuming that the posterior factorises over some subsets of variables. The form of the factorial distributions then falls off the conjugate-exponential model structure, and such approximations are called free-form. For instance, if our hierarchical model was a conjugate-exponential, we might assume that the joint posterior factorises into separate terms containing parameters $\boldsymbol{\theta}$ and hyperparameter ϕ.

Second important component of the variational analysis is to choose a mathematical measure of similarity/distance/closeness of two probability distributions, for then we can search for the approximate posterior distribution which is in this sense as close as possible to the true[5]. The asymmetric Kullback-Leibler (KL) divergence is one of the most popular choices, because it often leads to convenient analytical calculations[6]. The minimisation of the KL-divergence yields also a lower bound on the (logarithm) of the normalisation constant of the true posterior.

The VB-approach in a nutshell is that we define the form of the approximate distribution and choose a cost function, such as the KL-divergence, and seek the distribution which extremises this cost function. The setup resembles calculus of variations, whence historically the name Variational Bayes. The VB-method is intimately connected with the Expectation-Maximisation (EM) algorithm, which is an iterative method to find, for instance, a MAP-estimate for marginal posterior of the parameters, integrated over the hyperparameters in hierarchical models.

The VB-scheme seems attractive as we get not only an approximate posterior but a lower bound for the normalising constant of the true posterior, the evidence. This lower-bound functional is called free energy by the analogy to statistical physics, and plays a central role in many VB-analyses. The KL-minimisation is more of an optimisation problem (the integrations involved are analytically tractable, because of the simplified posterior), leading to considerable computational savings. The algorithm itself is deterministic, and the convergence can be mathematically proven for some approximation classes.

The key question is, of course, how good the approximation of the posterior and the lower bound on the evidence actually are, and there is no straightforward way to answer this. The VB-results can be compared for instance to those obtained by sampling techniques, as is done in this thesis. Again, if the true posterior is multimodal and the variational approximation (by definition) unimodal, the inferences based on the variational distribution underestimate the real uncertainty.

[5] Selection of a suitable distance measure is naturally connected to the form of the approximate posterior.

[6] The choice of KL-divergence is sometimes identified with the term Variational Bayes.

1.8 Recent Bayesian approaches to the inverse problem

Despite being rather extensively studied from the very beginning of MEG measurements, the inverse problem field has remained very active. This is partly due to the increase in computational resources opening possibilities to study more elaborate models and algorithms – especially the Bayesian approach has gained increasing popularity. Here we will give a run-through of those recent studies and ideas which are closest to and have most influenced the present work.

One of the early papers, containing the key elements of modern Bayesian analysis, was by Schmidt et al. (1999), entitled "Bayesian inference applied to the electromagnetic inverse problem". While not in any way the first Bayesian inverse approach, the authors took the first steps to scan the space of "all" probable solutions to the inverse problem with MCMC, in contrast to just trying to find the MAP or some other "best fitting" point estimate.

As most of the inverse approaches, the prior distribution for the currents was set to be a Gaussian with zero mean. As a novelty, the authors defined a set of activation parameters: the number, locations and spatial extents of the activations. The covariance of the Gaussian prior was then affected by these: if belonging to an "activated area", the prior covariance for those locations was increased, thus allowing larger currents in that region. The Gaussian prior enabled reducing the dimension of the parameter space by analytical marginalisation over the currents, and the remaining unknown parameters of the posterior were the activation parameters. The posterior was sampled with MCMC, upon which the following inferences were based.

Interestingly, the posterior distribution for the number of the activated regions, as well as their spatial extents, was obtained. Even though simulated and empirical data were analysed with the model, very little details were given about the exact implementation of the MCMC-scheme, convergence results, possible multimodality and so forth. Thus the feasibility of these rather aspiring ideas were subsequently analysed by several other researchers.

1.8.1 Multidipole models

The problem of determining the number of activations with MCMC was further analysed by Bertrand et al. (2001a,b), with the more straightforward way of using ECDs as an alternative to the activation parameters. In these articles, MCMC-techniques of Parallel Tempering (PT) and RJMCMC were applied to construct a chain which can jump between inverse solutions with different number of dipoles. The method was tested with simulated and empirical data, and it showed that the number of dipoles in the chains indeed vary, giving a distribution of different possible solutions. Even though the authors give the probability distribution of "number of dipoles" in Bertrand et al. (2001a), it is briefly discussed in Bertrand

et al. (2001b) that the different runs of the MCMC chain do not always contain all of the solutions. This indicates that the MCMC-method does not necessarily mix between all of the modes – strictly speaking the chains may have not converged globally but only locally. If that is the case, one can not reliably calculate the true posterior probability mass proportions of the solutions with different numbers of dipoles, based on the separate MCMC runs.

In another study, Kincses et al. (2003) looked at the possibility of estimating the spatial extent of sources. An MCMC-based estimator was constructed, with a similar "activation parameterisation" to Schmidt et al. (1999), but the activation was anatomically constrained to be a cortical patch and flat priors were assumed for the parameters. Simulation results and analysis of empirical somatosensory data were provided. The results support the feasibility of estimating the spatial source extent. However, only one source is assumed, which makes the inverse problem much easier. The location parameter space is not the whole cortex, but limited close to the true source location – if the algorithm would have been initialised far from the true source, it might have been trapped in a local minimum, a manifestation of the multimodality again. There is even a small technical issue in the implementation, as the location parameters are sampled with MCMC and the amplitude is obtained by a least-squares fit: the Markov chain will not be reversible producing a slight bias to the results.

Jun et al. (2005) reformulated the model of Schmidt et al. (1999) in terms of ECDs and RJMCMC, and generalised it to contain a temporal smoothness constraint for the activation timecourses. Considerable effort was also put to including a full spatio-temporal model for the noise covariance as well as its initial estimation. The noise covariance was later marginalised from the posterior, with the aim of producing a smoother distribution. Once again, the method is shown to provide reasonable estimates for simple simulated and empirical data, but the questions of the existence of multiple modes and if the chain actually jumps between these, and whether it is possible to really perform "full Bayesian analysis" on the posterior, are not explicitly addressed. In Jun et al. (2006), the model was augmented to incorporate the possibility of a dipole being active only a part of the time window under analysis.

Our group has adapted the model of Jun et al. (2005) to the cortically constrained case (Auranen et al., 2007a), with some improvements to computational efficiency, and studied the structure of the solutions with empirical and simulated data. The results show that the multimodality is an issue, especially with empirical data and larger number of sources. Subsequently, in (Auranen et al., 2007b) we studied the performance of the algorithm with a visual MEG dataset and compared the source localisations to corresponding fMRI data. We also investigated the possibility of using the fMRI data to create a better, informed proposal distribution to aid the mixing of the chain.

Finally, there are Bayesian multidipole approaches which are not based on

MCMC but to different Monte Carlo sampling techniques, such as particle filters (Somersalo et al., 2003). Whether these provide a more efficient alternative than MCMC in practical MEG data analysis will be seen in the future.

1.8.2 Distributed models

Most of the approaches to the distributed inverse problem are based on assuming a Gaussian prior with zero mean for the currents. This is mainly motivated by mathematical convenience: assuming also Gaussian noise, the posterior of the currents will be a Gaussian. Actually, this is true only if we know the covariance of the Gaussian prior. The most simple kind of a prior covariance would be proportional to the identity matrix, the constant of proportionality being the prior variance of the currents. The prior variance determines how large the currents can be, and with a usually assumed smallish value dictates that the currents are of rather similar magnitude throughout the brain.

The prior variance corresponds to what is called a regularisation parameter in the classical inverse problem literature, as it is used to constrain the overall "magnitude" of the inverse solution. How to determine a suitable value for the prior variance is an immediate question. In the classical literature, methods such as the "L-curve" (Hansen, 1992) have been used, where the inverse solutions are calculated with several values of the regularisation parameter, and the value which provides a compromise between the data fit and the solution magnitude is found. The generalisation to more complicated prior covariance structures is not so straightforward for these types of methods.

In the Bayesian terminology the prior variance is a basically a hyperparameter, that is an unknown parameter of the prior, and methods used in hierarchical modeling can be used for its estimation. The basic problem is that the joint posterior of the currents (parameters) and the prior variance (hyperparameter) is not of tractable form. Phillips et al. (2002) solve this with an EM-type algorithm, in which the values of the hyperparameter and the current amplitudes can be simultaneously estimated. The method finds the marginal or "restricted" maximum likelihood estimate of the hyperparameter, and given this, the MAP estimate for the currents.

In Phillips et al. (2005), the procedure is generalised to handle several covariance components. This is useful if we have fMRI data, for instance, and would like to let the prior variance be possibly larger for the fMRI-active source locations. Then the Gaussian prior would comprise of two covariance components, each with its own hyperparameter. The framework is further extended by Mattout et al. (2006) to include comparison of different prior combinations, by using a second-level inference procedure based on the evidence. With this, in theory, best combinations of the fMRI, smoothness, or whatever prior covariance components could be selected. How to select the candidates for prior covariance components

is, once again, not completely trivial.

In the approach of Trujillo-Barreto et al. (2004), the cortex is initially parcellated to (about 70) anatomical patches, and each of these, or any combination, can be assumed to contain a source, and hence to contribute a larger variance component. The evidences of the models obtained in this manner are estimated, and a Bayesian model average of the corresponding solutions is calculated. The practical problem is the combinatorial explosion of the number of possible prior covariance models. For one active area, there are 70 possible priors, for two 2415, for three 54740, and so forth. To overcome this problem, the authors choose the models in a nested way, so that those with vanishing evidence are not estimated, but the process is still somewhat heuristic. The performance of the method with a larger number or closely located sources has not been demonstrated, to the best of authors knowledge.

A framework, which in a sense contains the above covariance component models as special cases, is presented by Sato et al. (2004), as each of the cortical locations can have an individual prior variance parameter[7]. A second-level hyperprior is imposed on these prior variances. The currents and their variances are estimated from the MEG data with a VB-method. The Gaussian prior with same small variance throughout the brain produces solutions, which tend to be diffuse – small current "ripples" are distributed over the whole reconstruction grid, but the hierarchical model of Sato et al. (2004) allows some of the currents and their prior variances to obtain large values, while setting the others close to zero leading to more focal estimates.

This is the model which most of the work presented here concerns. We compare the estimates obtained by MCMC and VB, discuss the multimodality of the posterior, see how the hyperprior selection affects the inverse solutions, and how the model performs in empirical data analysis (Nummenmaa et al., 2007a,b,c). In (Auranen et al., 2005) we also study another continuous family of prior models, which include the Gaussian case, but also others which have been used to produce more focal inverse solutions. We then attempt to obtain a hierarchical estimate with MCMC methods, which includes the uncertainty about the prior selection.

[7] Actually, the model is formulated with precision or inverse variance parameters.

Chapter 2

Models and methods

2.1 The MEG observation model

In this work, we adopt the simple statistical model for the generation of the MEG data, in which the forward computed fields are mixed with additive Gaussian noise:

$$\boldsymbol{B}(t) = \boldsymbol{G}\boldsymbol{J}(t) + \boldsymbol{N}(t). \tag{2.1}$$

The measurement noise $\boldsymbol{N}(t)$ is assumed to be independent of time and to have a Gaussian distribution with zero mean and inverse noise covariance $\boldsymbol{\Sigma}_{\boldsymbol{G}}$. We do not include any possible uncertainties of the forward model (1.1) to the observation model and hence we get the likelihood[1] by solving $\boldsymbol{N}(t)$ from Eq. (2.1) and substituting to its postulated Gaussian distribution:

$$p(\boldsymbol{B}(t)|\boldsymbol{J}(t), \mathcal{M}) = \frac{|\boldsymbol{\Sigma}_{\boldsymbol{G}}|^{1/2}}{(2\pi)^{M/2}} \exp\left(-\frac{1}{2}(\boldsymbol{B}(t) - \boldsymbol{G}\boldsymbol{J}(t))'\boldsymbol{\Sigma}_{\boldsymbol{G}}(\boldsymbol{B}(t) - \boldsymbol{G}\boldsymbol{J}(t))\right), \tag{2.2}$$

where M is the number of MEG sensors. We have included the assumption of the Gaussianity of the noise, and the (fixed) noise covariance $\boldsymbol{\Sigma}_{\boldsymbol{G}}$ and gain matrix $\boldsymbol{G}$ to the symbol $\mathcal{M}$; in what follows, we will incorporate all fixed modeling assumptions to this symbol, assuming that its meaning can be deduced from context.

In most conventional MEG experiments, the same stimulus is presented several times, and a mean of these "trials" is calculated to enhance the signal-to-noise ratio (SNR). The noise in the averaged $\boldsymbol{B}(t)$ is an averaged random variable, and the central limit theorem justifies the Gaussianity assumption even if the noise in the single trials was not Gaussian, if the number of trials is reasonable.

[1] The likelihood function is the probability of the data given the parameters considered as a function of the parameters and hence it is not a probability density as such.

2.2 The minimum-norm and minimum-current priors

The work presented in this thesis involves hierarchical Bayesian generalisations of models related to two "traditional" inverse estimation techniques, the Minimum-Norm Estimate (MNE) and Minimum-Current Estimate (MCE), and these models are briefly introduced first.

The MNE stems from the assumption that the currents have a Gaussian prior distribution with zero mean, and a uniform precision (inverse variance) ρ^2 throughout the brain:

$$p_{MNE}(\boldsymbol{J}(t)|\rho^2, \mathcal{M}) \propto \exp\left(-\frac{\rho^2}{2}\boldsymbol{J}(t)'\boldsymbol{J}(t)\right) = \exp\left(-\frac{\rho^2}{2}\|\boldsymbol{J}(t)\|^2\right), \quad (2.3)$$

where the Euclidean norm of a vector $\boldsymbol{X}$ is defined as

$$\|\boldsymbol{X}\| = \left(\sum_i \boldsymbol{X}_i^2\right)^{1/2}. \quad (2.4)$$

Consequently, the prior considers less probable such current configurations for which the Euclidean norm of the currents is larger.

This is not the only reasonable functional choice to limit the current magnitudes; equally well, we could define another norm, denote it by $\|\boldsymbol{X}\|_1$,

$$\|\boldsymbol{X}\|_1 = \left(\sum_i |\boldsymbol{X}_i|\right). \quad (2.5)$$

The corresponding prior, which now penalises for the sum of the absolute values of the currents and not their squares, gives rise to the MCE:

$$p_{MCE}(\boldsymbol{J}(t)|\tau, \mathcal{M}) \propto \exp\left(-\frac{\tau}{2}\|\boldsymbol{J}(t)\|_1\right). \quad (2.6)$$

The prior of the MCE model is formally a Laplace distribution.

In both MNE and MCE models, the hyperparameters ρ^2 and τ are used to "regularise" the solutions, that is they set the overall scale for the currents. The smaller the prior width, the smaller the currents are constrained to be. Some properties of the inverse solutions associated with these priors can be deduced directly. Because a Gaussian distribution has very short tails, none of the currents can be particularly large in comparison with the others, as the prior precision is the same for the amplitudes at different locations. This causes the MNEs to be spatially diffuse, as the current is distributed quite evenly throughout the brain. Same is true for the MCE model, but to a lesser extent as the Laplace distribution has heavier tails, and the MCE prior can advocate also more focal solutions.

2.2.1 The minimum-norm estimate

Because in the MNE-model both the prior and the noise distributions are Gaussian, the posterior of the currents given the hyperparameter ρ^2 is Gaussian, and the posterior expectation (which coincides with the MAP in this case) of the currents can be analytically obtained. This procedure gives the MNE for the currents:

$$\hat{\boldsymbol{J}}_{MNE}(t) = \frac{1}{\rho^2}\boldsymbol{G}'\boldsymbol{\Sigma}_{\boldsymbol{B}}\boldsymbol{B}(t) \equiv \boldsymbol{W}\boldsymbol{B}(t), \quad \text{where} \quad \boldsymbol{\Sigma}_{\boldsymbol{B}}^{-1} = \frac{1}{\rho^2}\boldsymbol{G}\boldsymbol{G}' + \boldsymbol{\Sigma}_{\boldsymbol{G}}^{-1}. \quad (2.7)$$

The MNE is computationally very efficient, as an inverse operator (matrix) $\boldsymbol{W}$ can be formed, and the estimate is simply obtained by multiplying the observations with this matrix.

2.2.2 The minimum-current estimate

Because the prior of the MCE model is not Gaussian but Laplacian, calculating the posterior expectation or MAP estimate is not tractable analytically. In the implementation of Uutela et al. (1999), for instance, the problem is formulated as seeking a solution to the optimisation problem

$$\min \| \boldsymbol{J}(t) \|_1 \quad (2.8)$$

constrained by the observation equation

$$\boldsymbol{B}(t) \approx \boldsymbol{G}\boldsymbol{J}(t). \quad (2.9)$$

The problem can be solved efficiently by linear programming.

2.3 The ℓ^p-norm prior

Both of the norms defining MNE and MCE belong to the class of ℓ^p-norms, which is defined as:

$$\| \boldsymbol{X} \|_p = \left(\sum_i |\boldsymbol{X}_i|^p \right)^{1/p}, \quad p \geq 1. \quad (2.10)$$

The MNE corresponds to the case ℓ^2 and MCE to ℓ^1, and intermediate values of p yield norms and models with properties in between these. The ℓ^p-norm gives rise to prior

$$p_{\ell^p}(\boldsymbol{J}(t)|\kappa, p, \mathcal{M}) \propto \exp\left(-\frac{\kappa^p}{2} \| \boldsymbol{J}(t) \|_p^p \right). \quad (2.11)$$

We note that the ℓ^p-model contains two hyperparameters, namely the prior width (regularisation) parameter κ and the norm order p. While these are certainly subject to uncertainty, the norm order has been typically fixed in previous studies,

for computational convenience. But we have in principle no reasons *a priori* to exclude continuous norm orders and "mixture" models between the MCE and MNE. We therefore consider the model as a hierarchical Bayesian generalisation of the MCE and MNE, and attempt to estimate the joint posterior of the parameters and hyperparameters.

2.3.1 Estimating the ℓ^p-norm model

With the general ℓ^p-model, the posterior is analytically intractable, and we resort to MCMC methods. More specifically, the method of Slice Sampling (SS) is used (Neal, 2003), which is based on the idea of sampling uniformly under the graph of a one-dimensional distribution. The virtue of SS is that it can automatically adapt to different scales of the sampled distribution, does not require an explicit proposal distribution or the target distribution to be of any specific functional form. Samples of the joint posterior are obtained by sampling in turn the conditional distribution of each individual unknown variable (currents and hyperparameters) given the others.

2.4 The hierarchical Gaussian prior

Because the prior precision of the MNE-model is assumed to be uniform over the current reconstruction grid, more dipole-model like solutions where the current is large at few locations while being close to zero elsewhere, do not emerge from this framework. Hence, we study another hierarchical generalisation of the MNE-model (here called hMNE-model) in which individual prior precision parameters are allowed at each cortical location.

If $\boldsymbol{\alpha} = [\alpha_1, \ldots, \alpha_N]'$ is the vector of the prior precisions and $\boldsymbol{A} = \mathrm{diag}(\boldsymbol{\alpha})$ the corresponding diagonal matrix, the prior for the currents becomes

$$p_{hMNE}(\boldsymbol{J}(t)|\boldsymbol{\alpha}, \mathcal{M}) = \frac{|\boldsymbol{A}|^{1/2}}{(2\pi)^{N/2}} \exp\left(-\frac{1}{2}\boldsymbol{J}(t)'\boldsymbol{A}\boldsymbol{J}(t)\right). \qquad (2.12)$$

Then, a Gamma-hyperprior is assumed for the prior precisions

$$p_{hMNE}(\alpha_i|\alpha_0, \gamma_0, \mathcal{M}) = \frac{1}{\alpha_i}\left(\frac{\alpha_i\gamma_0}{\alpha_0}\right)^{\gamma_0}\Gamma(\gamma_0)^{-1}\exp\left(-\frac{\alpha_i\gamma_0}{\alpha_0}\right), \qquad (2.13)$$

where $\Gamma(\cdot)$ is the Euler Gamma function.

The hyperprior contains further parameters, namely the mean α_0 and the degrees of freedom γ_0 of the Gamma-distribution. Thus, α_0 sets the scale of the prior precisions, and γ_0 then determines how much they can vary around this mean value. Consequently, by letting $\gamma_0 \to \infty$, all α_i's are constrained to be equal to α_0, and we get the basic MNE model. However, assuming a smaller value for

γ_0 makes it possible for some of the source locations to acquire a small prior precision and large current, while the others remain small yielding more focal or multidipole-like inverse solutions.

The hMNE-model stems also from a well-known scale mixture representation of the Student t-distribution (Gelman et al., 2003; Geweke, 1993). Namely, with fixed α_0 and γ_0, marginalising the α_i's from the *prior* yields an independent Student t-distribution prior for the distributed current at each source location with zero mean (exists when $\gamma_0 > 1/2$), degrees of freedom $2\gamma_0$, and variance $\frac{\gamma_0}{\alpha_0(\gamma_0-1)}$ (exists when $\gamma_0 > 1$), with the t-distribution parameterised as in (Gelman et al., 2003). This offers an alternative way of considering how the prior becomes more Gaussian with large γ_0, and more heavy-tailed endorsing focal solutions when γ_0 gets smaller.

2.4.1 Estimating the hMNE-model

Again, the joint posterior of the currents $\boldsymbol{J}(t)$ and their prior precisions $\boldsymbol{\alpha}$ is not of tractable form. With this model, we perform the estimation with both VB and MCMC to obtain a more complete view on the structure of the posterior distribution and the inverse estimates.

Because the model has the conjugate-exponential structure, a free-form variational posterior can be obtained by assuming that it factorises over the currents $\boldsymbol{J}(t)$ and their prior precisions $\boldsymbol{\alpha}$. The minimisation of the KL-divergence from the true posterior to the approximate yields update equations for the parameters of the variational distribution (this turns out to be a product of Gaussian and Gamma-distributions), for which a fixed-point can be iteratively found.

With MCMC, the conjugate-exponential model enables using Gibbs sampling for posterior simulation. Namely, the joint conditional posterior of the currents given the prior precisions is Gaussian, and these can all be updated in one step. The joint posterior of the prior precisions given the currents is a product of univariate Gamma-distributions, and the prior precisions can be independently sampled. We also study the possibility of estimating the parameter α_0 from the data, and in this case SS is used to sample from its conditional posterior given the rest of the variables.

2.5 Afterthoughts

Since the introduction of the MNE by Hämäläinen and Ilmoniemi (1984), it has been widely used and studied from various viewpoints, such as incorporation of anatomical and physiological constraints and combining MEG with EEG (Dale and Sereno, 1993; Liu et al., 1998, 2002), noise-sensitivity normalisation (Dale et al., 2000), depth-weighting (Köhler et al., 1996; Lin et al., 2006b), and utilisation of the cortical orientation constraint (Lin et al., 2006a). By generalising

the prior (inverse) covariance to be proportional to a nondiagonal matrix, current estimates with spatial smoothness constraints and the corresponding "standardised" versions can be obtained (see, *e.g.*, Pascual-Marqui, 2002). No matter how much the MNE-model is turned, the estimates will be spatially rather diffuse, if the prior is assumed to be of the basic ℓ^2-form and our MCMC-studies of the ℓ^p-norm brings nothing new to this aspect.

For MCE, there are some interesting differences in the characteristics of the solutions obtained by MCMC, in comparison with those obtained by a classical implementation (such as by Uutela et al., 1999). In their method the solutions are also regularised by making a Singular Value Decomposition (SVD) to the observation equation, and leaving only the "most significant" MEG sensor combinations. The SVD cutoff value determines the number of constraining equations and the number of nonzero currents allowed in the solution, and the "MAP-solutions" are sparse (number of nonzero currents is less than or equal to the number of equations). In our MCMC studies we find "small current ripples" in the solutions, almost everywhere, and the solutions are not sparse.

One immediate explanation for this difference would be the additional regularisation provided by the SVD cutoff, but even without it the number of nonzero currents would still be less than 306 in the classical MCE. It could be also related to the differences of MAP-estimate and posterior expectation of the currents, which are not necessarily equivalent for the MCE model. Practically, the marginal densities of the currents could have maximum at zero, but posterior probability mass distributed asymmetrically. In our approach, the solutions are "averaged" over the prior variance, whereas in MCE the SVD-regularisation is fixed. Interestingly, Rao et al. (2003) analyse the ℓ^p-prior model[2] with a generic linear "inverse problem" and their general results imply also that the MAP-estimate of the ℓ^1-norm model is sparse, when estimated with fixed regularisation parameter.

One could also argue that the presence of the small values is only due to the Monte Carlo error in the estimated posterior expectation values. We have reasons to believe that the difference is real as the Laplace prior does not really force the currents to be zero in any way. Our results are in a sense concordant with those of Lassas and Siltanen (2004), where the use of total variation prior for edge-preserving Bayesian inversion was studied. Their results show, among other things, that the MAP and posterior expectation estimates are quite different, especially when the number of discretisation points is increased. Why this matter of the small vs. exactly zero currents is of any importance is that as we will later see, the small current ripples can give rise to large MEG fields if they suitably sum up – therefore the thresholding of the estimates is not a completely trivial operation.

In context of MEG inverse problem, the ℓ^p-norm model has been analysed by

[2] In this study the parameter p is actually allowed to be smaller than unity.

Beucker and Schlitt (1996); Bücker and Beucker (2004), but more from an optimisation point of view than Bayesian estimation. Solutions for different values of p are obtained, but no quantitative means for comparison of these solutions are discussed. We note also that our actual analysis involves a slightly different but equivalent parameterisation of the ℓ^p-norm prior (see, pp. 156-160, Box and Tiao, 1973).

Relatives of the hierarchical Gaussian prior have been used to perform Automatic Relevance Determination (ARD) based input selection for artificial neural networks (Neal, 1996) and Sparse Bayesian Learning (SBL) of regression and classification models (Tipping, 2001). As a basis for MEG inverse estimation it was first officially published by Sato et al. (2004) using the VB-method. Unaware of this work, we were simultaneously analysing the same model with MCMC and presented preliminary results in Nummenmaa et al. (2004). It turned out that, with proper understanding of some technical aspects of the model, the VB-algorithm produces similar results to the MCMC but with considerably less computational effort. Therefore, we give Sato et al. (2004) full credit of inventing the method.

Finally, we mention that both the ℓ^p- and the hierarchical Gaussian model were studied in a setup where the noise covariance actually contained an unknown scale parameter, which was estimated from the data as well. This detail was deliberately left out from the introduction of the models and methods to keep the focus on the relationships of the different priors.

Chapter 3

Summary of studies

3.1 General methodology

We acknowledge the following general methods, software tools and measuring instruments used in this work. The MEG measurements were obtained with Neuromag Vectorview device (Elekta Neuromag Oy), located at the Brain Research Unit, Low Temperature Laboratory, Helsinki University of Technology. MRI and fMRI measurements were made with GE Signa EXCITE 3T scanner, at the Advanced Magnetic Imaging Centre, Helsinki University of Technology (MRIs of previous studies acquired elsewhere were also utilised in part). The FreeSurfer software was used to segment the grey matter – white matter surface from the anatomical MRIs (Dale et al., 1999; Fischl et al., 1999, 2001), which then served as the space of possible source locations. The orientations of the sources were constrained to be perpendicular to this surface. The forward computations were made with the MNE software package by Matti Hämäläinen, Martinos Center HMS/MGH/MIT, using the one layer boundary-element method (see, *e.g.*, Hämäläinen et al., 1993). FMRIB's Software Library was used in analysing the fMRI data (Smith et al., 2004). MEG signal processing, inverse estimation, and visualisation of the results were implemented with MATLAB (The Mathworks, Inc.).

3.2 Bayesian analysis of the ℓ^p-norm model (P I)

3.2.1 Rationale

The ℓ^p-norm family includes the prior models of MCE (ℓ^1) and MNE (ℓ^2) as special instances. In previous studies, MCE has been shown to produce more focal estimates, whereas MNEs are spatially very diffuse. Any norm order p between 1 and 2 could equally well be used, and these evidently produce estimates "between" the MCE and MNE models. It seems natural to pose questions such as

what do the MEG data tell about the posterior distribution of the norm order p, and would it be feasible to infer whether MCE or MNE or any norm in between is more appropriate for a given dataset. Since MCE is known to produce more focal solutions than MNE, it would be also interesting to see if the spatial extent of the source is somehow reflected in the posterior distribution of p.

3.2.2 Methods

We use both simulated and empirical data in this study. In simulations, sources of different spatial extent are created to the cortical surface, and the forward calculated fields are corrupted with Gaussian noise of various levels. The empirical data consist of averaged MEG fields evoked by self-paced index finger lifts. By using slice sampling, the joint posterior of the hyperparameters (norm order, prior width) and the parameters (noise covariance scale, current amplitudes) is numerically estimated. The inverse solutions are obtained for several source spaces with different grid densities. Due to the computational intensiveness of the MCMC-scheme, the analysis is restricted to a single timepoint.

3.2.3 Results

The results show that the information about the norm order p in the MEG data is rather limited. The marginal distribution of p "leans" strongly to $p = 1$ for almost all studied cases, and is rather flat for larger values. With increasing reconstruction grid sizes and decreasing SNR, there seems to be more mass in the $p = 2$ side of the distribution. The reason for this could be just the increase of uncertainty with more parameters and noise. On the other hand, there can also be some mathematical subtleties involved – in the study of Lassas and Siltanen (2004), it is shown that the "edge-preserving" total variation prior becomes effectively Gaussian when the number of discretisation points becomes very large. The spatial extent of the source affects the marginal posterior of p very weakly, if at all. The analysis of the empirical somatomotor data gives largest current peaks close to the expected areas.

3.2.4 Comments

The value of this brute-force analysis is rather the theoretical insight obtained than the ability of the ℓ^p-method to produce "better" solutions than MCE or MNE in some sense. Actually, the results indicate that the value of p would "like" to be even smaller than 1, which was not allowed in this study as the ℓ^p-norm is actually a norm only for $p \geq 1$. In other respects, the ℓ^p-prior parameterisation could be used also for $0 < p \leq 1$. In this range, even simple MAP-estimation with fixed hyperparameters and optimisation techniques becomes more difficult due to the

nonconvexity of the (negative natural logarithm) of prior. Extrapolating the results a bit for the MCMC and unknown p, it is very likely that the posterior of p would lean on its smallest admitted positive value, and the obtained solutions would become more focal, but the prior's "less restrictive" nature would probably cause multimodality and convergence problems for the MCMC-method. The MCMC-based estimates with different values of $1 \leq p \leq 2$ and are not really that different and the focality of the solutions is also partly a matter of thresholding.

3.3 Theoretical aspects of the hierarchical Gaussian model (P II)

3.3.1 Rationale

The hierarchical Gaussian prior offers an alternative way to construct a more flexible model for the distributed MEG source reconstruction problem. The question of how to choose the hyperparameters α_0, γ_0 immediately raises, and what are the effects of this selection. In the VB-approach of Sato et al. (2004), this matter was solved by letting $\gamma_0 \to 0$ and $\alpha_0 =$ undefined, which leads to the "noninformative" Jeffrey's prior (see, Eq. 2.13), which is an improper (unnormalisable) distribution. Improper priors are often used, but in this case the improper prior causes the posterior also to become improper. With proper posteriors, MCMC is in theory capable of numerically representing the true posterior, whereas the VB-approach assumes inherently an analytically simpler form for the posterior. We therefore discuss the theoretical consequences of using the noninformative hyperprior, and make a comparison of the VB and MCMC estimates to obtain a more complete view on how the algorithms behave with different selections of the Gamma-hyperprior.

3.3.2 Methods

Since this is more of a theoretical article we use only simulated data, so that a "ground truth" of some sort is available for comparison. The data are generated by assuming two sources with similar timecourses and additive Gaussian noise. We demonstrate how the posterior becomes singular both by analytical calculations and by studying the numerical behaviour of the VB and MCMC algorithms. Then we analyse the hyperprior sensitivity and consider the possibility of estimating the hyperprior parameter α_0 from the data, and augment the previously presented VB-algorithm in this respect. Possible multimodality of the posterior is studied by initialising the algorithms randomly and observing whether they converge to different regions of the parameter space.

3.3.3 Results

The results show that especially the parameter γ_0 has a clear effect on the solutions and the behaviour of the both algorithms. When the γ_0 gets smaller and closer to the singular case, convergence of the algorithms slows down. Estimates obtained with different γ_0's also show significant variability. Estimating the hyperparameter α_0 from the data is technically possible for fixed γ_0, but the computational cost increases significantly for the present methods. The multimodality of the true posterior is evident as several different "solution candidates" emerge from the randomly initialised MCMC- and VB-algorithms. The multimodality is not to be confused with the (possibly multiple) spatial current peaks indicating likely source locations. The posterior distribution is difficult to handle because the Markov chain does not mix between the different modes, at least for feasible computational times. For such values of α_0 and γ_0 that the algorithms show robust (local) convergence, both methods produce rather similar estimates, the VB-method being computationally much more efficient.

3.3.4 Comments

Technically, the Markov chain used in this analysis does not converge globally at all, but only locally, since it gets trapped in different regions of the parameter space depending on initialisation and does not jump between these regions (for feasible computation times). Calculation of the posterior probability mass proportions of the different modes based on the separate MCMC-runs is not directly possible. For the VB-method used in this study, the variational posterior is always unimodal. From theoretical viewpoint, using the unimodal variational posterior as a proxy for the multimodal true posterior can lead to overinterpretations of the results, if the limitations of the technique are not properly understood.

3.4 The hierarchical model in practical MEG analysis (P III)

3.4.1 Rationale

The studies of the hierarchical Gaussian model with the simplistic simulated datasets give limited information on how the model behaves with more complex real MEG data. Hence, we analyse an empirical dataset with several experimental stimulus combinations with the hierarchical method and study how the theoretical aspects such as hyperprior selection and multimodality show up in practical MEG data analysis. Tutorial-like elements are included to this presentation in order to make it more attractive and accessible from the empirical researcher's viewpoint.

3.4.2 Methods

The experimental data consist of MEG fields evoked by simple auditory tones and visual checkerboard images, presented alone in conditions (A) and (V), and in combination (AV). The stimuli are such that drastic audiovisual integration effects should be absent, and the MEG fields add roughly linearly: (A) + (V) $\approx$ (AV). We demonstrate one specific virtue of the hierarchical approach in its ability to produce more focal and dipole-like solutions, which is related to the thresholding of the estimates. With the basic MNE-model, all current amplitudes are of rather similar magnitude, and contribute roughly equally to the "data fit". Thus, showing heavily thresholded estimates can lead to a situation in which the subthreshold amplitudes actually explain a larger proportion of the variability in the data. By "favouring" the more focal solutions we do not mean to say that the activations would in reality be pointlike or dipolar, but being able to model the data with a drastically smaller "effective number of parameters" makes the interpretation of the results in a sense more robust. We also demonstrate that the nature of the inverse problem causes difficulties to model selection based on the evidence or its variational free energy approximate. The role of the γ_0 as a regularisation parameter is studied by comparing data fits (and free energies) associated with its different values. We study the nonlinearity of the VB-algorithm by fixing the hyperprior and calculating the inverse estimates for the different data cases (A), (V), and (AV). Finally, we discuss the circumstances under which the free energy value could be used to calculate approximate posterior mass proportions of the different modes, and study the multimodal structure of the solutions obtained by random initialisation of the VB-algorithm.

3.4.3 Results

With thresholding, the hierarchical inverse estimate is able to produce the same degree of data fit with only a couple of sources for which hundreds of MNE-sources are needed. In the model selection, the free energy based procedure leads to choosing a very sparse reconstruction grid and a small value of γ_0, but for which the solution looks visually quite implausible. The regularisation analysis shows that larger values of γ_0 correspond to more regularised solutions which is quite natural as the hyperprior becomes more "informative" and restricting. Studying the nonlinearity shows that with particular hyperprior selections, the nonlinearity effects can be rather drastic – for the (A) and (V) cases reasonable estimates with large source amplitudes are obtained, but with the (AV)-case the auditory sources remain very small. It means that even when the data add linearly, the solutions necessarily do not. Analysis of the multimodality suggests that the hyperprior selection can have a significant effect also on the free energy estimated posterior mass proportions, and as the hyperprior selection is done on an *ad hoc* basis, the

trustworthiness of these numerical estimates is suspect.

3.4.4 Comments

The thresholding problem most likely touches not just MNE but all methods in which the variance of the currents is assumed to be rather uniform over the cortex, and their noise sensitivity normalised and standardised versions. Actually, in the original article we refer to this as the "nonstatistical" thresholding problem in contrast to statistical thresholding of fMRI activation maps. From theoretical viewpoint, this study in a sense demonstrates that the full Bayesian analysis of this general hierarchical model is very difficult, if not impossible. Still, after the disenchantment, the VB-method provides reasonable and robust inverse estimates in a rather automated way and computationally efficient manner, appearing as a viable option for practical MEG inverse analysis.

3.5 Sparse hierarchical solutions and comparison with fMRI (P IV)

3.5.1 Rationale

The thresholding problem could be circumvented by obtaining sparse solutions to the MEG inverse problem, that is, forcing the small amplitude values to be zero. The mathematically most elegant way of obtaining this property would be to assume a (Dirac delta) point mass prior for the current being exactly zero, but such a cosntruction would possibly lead to similar problems as in multidipole models with an unknown number of dipoles. Hence, we study how to obtain effectively sparse solutions by using the hierarchical Gaussian prior. Complex visual responses have been successfully analysed by manual multidipole approaches, and hence it is interesting to see how well a largely automated inverse algorithm can perform in a similar situation. The visual system has been rather extensively studied, electrophysiologically and otherwise, and the evoked responses are challenging to analyse because the sources of different visual areas are spatially close and have temporally overlapping activation patterns. Data from fMRI can be used as a qualitative reference for the MEG source locations in the absence of the "ground truth".

3.5.2 Methods

Three subjects participated in identical experiments carried out in MEG and fMRI. The stimulus was a drifting grating located in the lower left quadrant of the visual field. The drifting grating stimulus activated the retinotopic areas and the motion sensitive area. In addition, the retinotopic visual areas were localised by using a

multifocal fMRI paradigm developed by one of the co-authors in order to facilitate the functional identification of the drifting grating activations. The sparsity is obtained by setting the α_0 parameter controlling the overall magnitude of the prior precisions to a very large value, which forces most of the current amplitudes to be effectively zero. Using several values of γ_0, the inverse estimates are calculated, and the value which produces intermediate degree of regularisation is chosen. We also study the possibility of utilising the fMRI information in the inverse estimates by initialising the algorithms according to this.

3.5.3 Results

The fMRI activations to the drifting grating stimulus are found to be physiologically plausible. However, both fMRI data and MEG responses show significant intersubject variability, especially in terms of SNR. Consequently, there is also significant variability among the inverse estimates, especially in the number of locations estimated to contain a large current amplitude. The overall source locations and timecourses are found to be consistent with previous studies. The locations are somewhat superficial in comparison to the fMRI, which is due to the fact that the hierarchical method also favours solutions with small amplitudes. The initialisation with the fMRI data has only a slight effect on the inverse estimates.

3.5.4 Comments

The sparsity assumption and implementation with the hierarchical prior appears to be quite strong, as the fMRI-initialisation produces very little effects. The reason is also in that the prior precisions, which dictate where the large amplitudes are located, are really estimated from the MEG data only; a more symmetric model should be considered for true integration of MEG and fMRI.

3.6 Corrigenda for (P II)-(P III)

Unfortunately, there was a small error in the numerical implementation of the free energy used in P II and P III, see the Appendix of (Nummenmaa et al., 2007a). In the code, $\Gamma(\gamma_0)$ was erroneously written as $\Gamma(\gamma_0 + T/2)$. This has no effects whatsoever on the estimates themselves – it only causes errors to comparisons of the free energies obtained with different values of γ_0. This computational error actually promoted the conclusions, too hastily drawn, that the free energy increases and approaches a finite value in the limit $\gamma_0 \to 0$. This matter is not so easily settled and it is actually quite dangerous to write equations related to the improper distributions, such as $p(\alpha_i) = 1/\alpha_i$, because there is an infinite normalising constant missing. By looking at the Gamma-distribution in the limit $\gamma_0 \to 0$, this

normalisation actually is precisely the $\Gamma(\gamma_0)$-term miscalculated in the free energy. One should be more careful with taking limits and notation with improper distributions. Even if these infinite normalisation constants do not affect the VB-update equations, which are basically obtained by taking derivatives of the free energy, they might affect model selection or other conclusions made based on the free energy approximation of the evidence.

We also note that due to insuperable typesetting difficulties beyond the authors' control, the manuscript version of P II is included instead of the official print which has limited readability.

Chapter 4

Discussion

The MEG inverse problem has been subject for intensive study during several decades and real progress is hard won. In this thesis we have studied some hierarchical Bayesian generalisations of the classical MEG inverse models and methods. The tenacious reader who has waded through the material presented up to this point may also have realised that the more theoretical results are largely independent of the physical realisation of MEG. Analogous analyses could be performed for whatever inverse problem, where the forward problem is linear (for other instances of linear inverse problems, see, *e.g.*, Kaipio and Somersalo, 2005). The practical results of course depend on the special characteristics of the system under study. Having said that, actually a very similar variational setup has been introduced for Bayesian regression and classification (Bishop and Tipping, 2000), as is studied here in the context of MEG inverse problem. While the algorithms and mathematical formulae are closely related, there are differences. The hyperprior selection and the potential improperness of the posterior that we found to have such a drastic effect on the inverse solutions is left rather untouched in (Bishop and Tipping, 2000; Tipping, 2001). The reason could be simply that basic regression is a somewhat less ill-posed problem than the MEG inverse, despite the mathematical similarities. The simulated datasets used in the regression examples are also of quite different complexity from empirical MEG data.

This brings us to the subject of inverse crimes in simulation studies. Inverse crimes are committed by assuming aspects of the model and data generation process which are in reality subject to uncertainty to be exactly known when inverting the data. For MEG these unknown aspects could include conductivity values of the biological tissues, locations of the MEG sensors with respect to the head, accuracy of the cortical surface reconstruction, forward model computation and so on. Most fundamental of these crimes is simulating the data with exactly the same discretisation of the cortex which is later used in the inverse estimation. This type of procedure yields always too optimistic results, to what extent depends again

on the inverse model. In more general terms, the data generation and inverse estimation are both based on a "true" model, a case which is seldom realised in practical analyses. The effect of having the "true parameter" which generates the data among the candidate solutions gives too good results especially when maximisation of model evidence is involved in the model or solution selection. We have tried to avoid at least the most aggravating circumstances in our simulations, but admittedly simulated data tend to be just too simple, especially for cases involving complex cognitive functions. More such empirical datasets are needed where fMRI or other imaging modalities are used to assess the performance of different inverse approaches – in a recent study Bai et al. (2007) compared many different distributed inverse algorithms using intracranial electrocorticograms and fMRI as a reference.

In the present work, especially the approach based on the hierarchical Gaussian prior is found to yield plausible solutions for both empirical and simulated data, when the properties of the estimates are properly understood. As with most inverse methods, there are limitations and room for improvement. Because the hierarchical prior still favours solutions with smaller current amplitudes, the estimated source locations tend to be somewhat too superficial. The depth weighting used with classical MNE could be implemented also to the hierarchical method, but the effects would most likely be similar to those obtained by the simple fMRI initialisation. The basic reason for the "depth bias" in the MNE-type estimates is simply physical: the magnetic field of a current dipole dies off like inverse square as a function of distance, hence deeper sources produce smaller fields. If both smaller superficial and larger deep source configurations produce similar MEG observations, a "minimum-something" estimate will always favour the more superficial. Of course, the depth weighting can be used to push the estimates deeper and is mathematically a perfectly viable operation. From statistical inference viewpoint, such a procedure incorporating conflicting priors seems a bit dissatisfying.

By adopting a specific normalisation scheme for the currents, spatial activation maps with "zero localisation error" have been reported (Pascual-Marqui, 2002). Afterwards, the zero localisation error property was demonstrated analytically to hold for case of one dipolar source (Sekihara et al., 2004) and high signal to noise ratio. While this is theoretically and practically interesting, Wagner et al. (2004) also showed by simulations that for several sources and noisy data, the localisation error is nonzero, which hardly comes as a big surprise. The zero localisation error just means that the maximum peak of the possibly spatially rather diffuse activity map is concordant with the true source location. Almost needless to say, the concepts of true source and zero localisation error are prone to our previous discussion on inverse crimes.

Things that are not concerned in the present work at all include spatial and dynamical modeling of the sources. That is, we could incorporate some prior

assumptions about smoothness of the spatial and temporal characteristics of the sources into the model. These types of models have been introduced, but typically the computational cost of estimating such a model is rather high. Of course, the integration of fMRI and MEG/EEG is and will continue to be an area of special interest, and the way in which this could be realised in the hierarchical framework is currently under investigation. Very recently, a more symmetric method for EEG/fMRI fusion based on a VB-framework rather similar to the one studied here was introduced by Daunizeau et al. (2007). Hopefully in the future these types of models and methods will help to add some pieces to the puzzle of human brain functions.

References

Auranen, T., Nummenmaa, A., Hämäläinen, M. S., Jääskeläinen, I. P., Lampinen, J., Vehtari, A., and Sams, M. (2005). Bayesian analysis of the neuromagnetic inverse problem with ℓ^p-norm priors. *NeuroImage*, 26(3):870–884.

Auranen, T., Nummenmaa, A., Hämäläinen, M. S., Jääskeläinen, I. P., Lampinen, J., Vehtari, A., and Sams, M. (2007a). Bayesian inverse analysis of neuromagnetic data using cortically constrained multiple dipoles. *Human Brain Mapping*, 28(10):979–994.

Auranen, T., Nummenmaa, A., Vanni, S., Vehtari, A., Hämäläinen, M. S., Lampinen, J., and Jääskeläinen, I. P. (2007b). Automatic fMRI-guided MEG multidipole localization for visual responses. *Submitted*.

Bai, X., Towle, V. L., He, E. J., and He, B. (2007). Evaluation of cortical current density imaging methods using intracranial electrocorticograms and functional MRI. *NeuroImage*, 35(2):598–608.

Baillet, S., Mosher, J. C., and Leahy, R. M. (2001). Electromagnetic brain mapping. *IEEE Signal Processing Magazine*, pp. 14–30.

Bernardo, J. M. and Smith, A. F. M. (2000). *Bayesian Theory*. John Wiley & Sons, Ltd.

Bertrand, C., Hamada, Y., and Kado, H. (2001a). MRI prior computation and parallel tempering algorithm: A probabilistic resolution of the MEG/EEG inverse problem. *Brain Topography*, 14(1).

Bertrand, C., Ohmi, M., Suzuki, R., and Kado, H. (2001b). A probabilistic solution to the MEG inverse problem via MCMC methods: The reversible jump and parallel tempering algorithms. *IEEE Transactions on Biomedical Engineering*, 48(5).

Beucker, R. and Schlitt, H. A. (1996). On minimal lp-norm solutions of the biomagnetic inverse problem. Technical Report KFA-ZAM-IB-9614, Research Center Jülich, Germany.

Bishop, C. M. and Tipping, M. E. (2000). Variational relevance vector machines. In Boutilier, C. and Goldszmidt, M., editors, *Uncertainty in Artificial Intelligence Proceedings*. Morgan Kaufmann.

Box, G. E. P. and Tiao, G. C. (1973). *Bayesian Inference in Statistical Analysis*. John Wiley and Sons, Inc.

Bücker, H. M. and Beucker, R. (2004). Using automatic differentiation for the solution of the minimum p-norm estimation problem in magnetoencephalography. *Simulation Modelling Practice and Theory*, 12(2):105–116.

Dale, A. M., Fischl, B., and Sereno, M. I. (1999). Cortical surface-based analysis I: Segmentation and surface reconstruction. *NeuroImage*, 9:179–194.

Dale, A. M., Liu, A. K., Fischl, B. R., Buckner, R. L., Belliveau, J. W., Lewine, J. D., and Halgren, E. (2000). Dynamic statistical parametric mapping: Combining fMRI and MEG for high-resolution imaging of cortical activity. *Neuron*, 26:55–67.

Dale, A. M. and Sereno, M. I. (1993). Improved localization of cortical activity by combining EEG and MEG with MRI cortical surface reconstruction: A linear approach. *Journal of Cognitive Neuroscience*, 5(2):162–176.

Daunizeau, J., Grova, C., Marrelec, G., Mattout, J., Jbadbi, S., Pélégrini-Issac, M., Lina, J.-M., and Benali, H. (2007). Symmetrical event-related EEG/fMRI information fusion in a Variational Bayesian framework. *NeuroImage*, 36:69–87.

Dawid, A. P. (1979). Conditional independence in statistical theory. *Journal of the Royal Statistical Society. Series B (Methodological)*, 41(1):1–31.

Dayan, P. and Abbot, L. F. (2001). *Theoretical neuroscience: computational and mathematical modeling of neural systems*. MIT Press.

Fischl, B., Liu, A., and Dale, A. M. (2001). Automated manifold surgery: Constructing geometrically accurate and topologically correct models of the human cerebral cortex. *IEEE Transactions on Medical Imaging*, 20(1):70–80.

Fischl, B., Sereno, M. I., and Dale, A. M. (1999). Cortical surface-based analysis II: Inflation, flattening, and a surface-based coordinate system. *NeuroImage*, 9:195–207.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2003). *Bayesian Data Analysis*. Chapman & Hall/CRC, second edition.

Geweke, J. (1993). Bayesian treatment of the independent Student-t linear model. *Journal of Applied Econometrics*, 8(Supplement):S19–S40.

Goel, P. K. and Degroot, M. H. (1981). Information about hyperparameters in hierarchical models. *Journal of the American Statistical Association*, 76(373):140–147.

Hansen, P. C. (1992). Analysis of discrete ill-posed problems by means of the L-curve. *SIAM review*, 34(4):561–580.

Harville, D. A. (1999). *Matrix Algebra From a Statistician's Perspective*. Springer-Verlag.

Hämäläinen, M. S., Hari, R., Ilmoniemi, R. J., Knuutila, J., and Lounasmaa, O. V. (1993). Magnetoencephalography — theory, instrumentation, and applications to noninvasive studies of the working human brain. *Reviews of Modern Physics*, 65(2):413–497.

Hämäläinen, M. S. and Ilmoniemi, R. J. (1984). Interpreting measured magnetic fields of the brain: Estimates of current distributions. Technical Report TKK-F-A559, Helsinki University of Technology, Department of Technical Physics.

Huettel, S. A., Song, A. W., and McCarthy, G. (2004). *Functional Magnetic Resonance Imaging*. Sinauer.

Jezzard, P., Matthews, P. M., and Smith, S. M., editors (2001). *Functional MRI*. Oxford University Press.

Jun, S. C., George, J. S., Paré-Blagoev, J., Plis, S. M., Ranken, D. M., Schmidt, D. M., and Wood, C. C. (2005). Spatiotemporal Bayesian inference dipole analysis for MEG neuroimaging data. *NeuroImage*, 28(1):84–98.

Jun, S. C., George, J. S., Plis, S. M., Ranken, D. M., Schmidt, D. M., and Wood, C. C. (2006). Improving source detection and separation in a spatiotemporal Bayesian inference dipole analysis. *Physics in Medicine and Biology*, 51:2395–2414.

Kaipio, J. P. and Somersalo, E. (2005). *Statistical and Computational Inverse Problems*, volume 160 of *Applied Mathematical Sciences*. Springer.

Kandel, E. R., Schwartz, J. H., and Jessell, T. M., editors (2000). *Principles of Neural Science*. McGraw-Hill, fourth edition.

Kass, R. E. and Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, 90(430):773–795.

Köhler, T., Wagner, M., Fuchs, M., Wischmann, H.-A., Drenckhahn, R., and Theissen, A. (1996). Depth nornalization in MEG/EEG current density imaging. In *Conference Proceedings of the 18th Annual International Conference of the Engineering in Medicine and Biology Society of the IEEE*.

Kincses, W. E., Braun, C., Kaiser, S., Grodd, W., Ackermann, H., and Mathiak, K. (2003). Reconstruction of extended cortical sources for EEG and MEG based on a Monte-Carlo-Markov-chain estimator. *Human Brain Mapping*, 18:100–110.

Lappalainen, H. and Miskin, J. W. (2000). *Advances in Independent Component Analysis*, chapter Ensemble Learning, pp. 75–92. Springer-Verlag.

Lassas, M. and Siltanen, S. (2004). Can one use total variation prior for edge-preserving Bayesian inversion? *Inverse Problems*, (20):1537–1563.

Lin, F.-H., Belliveau, J. W., Dale, A. M., and Hämäläinen, M. S. (2006a). Distributed current estimates using cortical orientation constraints. *Human Brain Mapping*, 27:1–13.

Lin, F.-H., Witzel, T., Ahlfors, S. P., Stufflebeam, S. M., Belliveau, J. W., and Hämäläinen, M. S. (2006b). Assessing and improving the spatial accuracy in MEG source localization by depth-weighted minimum-norm estimates. *NeuroImage*, 31:160–171.

Liu, A. K., Belliveau, J. W., and Dale, A. M. (1998). Spatiotemporal imaging of human brain activity using functional MRI constrained magnetoencephalography data: Monte Carlo simulations. *Proceedings of the National Academy of Sciences of the United States of America*, 95(15):8945–8950.

Liu, A. K., Dale, A. M., and Belliveau, J. W. (2002). Monte Carlo simulation studies of EEG and MEG localization accuracy. *Human Brain Mapping*, 16:47–62.

Mattout, J., Phillips, C., Penny, W. D., Rugg, M. D., and Friston, K. J. (2006). MEG source localization under multiple constraints: An extended Bayesian framework. *NeuroImage*, 30:753–767.

Neal, R. M. (1996). *Bayesian Learning for Neural Networks*. Springer-Verlag.

Neal, R. M. (2003). Slice sampling. *The Annals of Statistics*, 31(3):705–767.

Nummenmaa, A., Auranen, T., Hämäläinen, M. S., Jääskeläinen, I. P., Lampinen, J., Sams, M., and Vehtari, A. (2004). A hierarchical Bayesian approach in distributed MEG source modelling. In 10th Annual Meeting of the Organization for Human Brain Mapping (CD-ROM, WE 265), June 13–17, Budapest, Hungary.

Nummenmaa, A., Auranen, T., Hämäläinen, M. S., Jääskeläinen, I. P., Lampinen, J., Sams, M., and Vehtari, A. (2007a). Hierarchical Bayesian estimates of distributed MEG sources: Theoretical aspects and comparison of variational and MCMC methods. *NeuroImage*, 35(2):669–685.

Nummenmaa, A., Auranen, T., Hämäläinen, M. S., Jääskeläinen, I. P., Sams, M., Vehtari, A., and Lampinen, J. (2007b). Automatic relevance determination based hierarchical Bayesian MEG inversion in practice. *NeuroImage*, 37(3):876–889.

Nummenmaa, A., Auranen, T., Vanni, S., Hämäläinen, M. S., Jääskeläinen, I. P., Lampinen, J., Vehtari, A., and Sams, M. (2007c). Sparse MEG inverse solutions via hierarchical Bayesian modeling: evaluation with a parallel fMRI study. *Submitted*.

Opper, M. and Saad, D., editors (2001). *Advanced Mean Field Methods: Theory and Practice*. MIT Press.

Pascual-Marqui, R. D. (2002). Standardized low resolution brain electromagnetic tomography (sLORETA): Technical details. *Methods & Findings in Experimental & Clinical Pharmacology*, 24:5–12.

Phillips, C., Mattout, J., Rugg, M. D., Maquet, P., and Friston, K. J. (2005). An empirical Bayesian solution to the source reconstruction problem in EEG. *NeuroImage*, 24:997–1011.

Phillips, C., Rugg, M. D., and Friston, K. J. (2002). Systematic regularization of linear inverse solutions of the EEG source localization problem. *NeuroImage*, 17:287–301.

Purves, D., Augustine, G. J., Fitzpatrick, D., Katz, L. C., LaMantia, A.-S., McNamara, J. O., and Williams, M. S., editors (2001). *Neuroscience*. Sinauer, second edition.

Rao, B. D., Engan, K., Cotter, S. F., Palmer, J., and Kreutz-Delgado, K. (2003). Subset selection in noise based on diversity measure minimization. *IEEE Transactions on Signal Processing*, 51(3):760–770.

Sato, M.-A., Yoshioka, T., Kajihara, S., Toyama, K., Goda, N., Doya, K., and Kawato, M. (2004). Hierarchical Bayesian estimation for MEG inverse problem. *NeuroImage*, 23:806–826.

Schmidt, D. M., George, J. S., and Wood, C. C. (1999). Bayesian inference applied to the electromagnetic inverse problem. *Human Brain Mapping*, 7:195–212.

Sekihara, K., Sahani, M., and Nagarajan, S. S. (2004). Localization bias and spatial resolution of adaptive and non-adaptive spatial filters for MEG source reconstruction. *NeuroImage*, 25:1056–1067.

Smith, S. M., Jenkinson, M., Woolrich, M. W., Beckmann, C. F., Behrens, T. E. J., Johansen-Berg, H., Bannister, P. R., Luca, M. D., Drobnjak, I., Flitney, D. E., Niazy, R. K., Saunders, J., Vickers, J., Zhang, Y., Stefano, N. D., Brady, J. M., and Matthews, P. M. (2004). Advances in functional and structural MR image analysis and implementation as FSL. *NeuroImage*, 23:S208–S219.

Somersalo, E., Voutilainen, A., and Kaipio, J. P. (2003). Non-stationary magnetoencephalography by Bayesian filtering of dipole models. *Inverse Problems*, 19:1047–1063.

Tipping, M. E. (2001). Sparse Bayesian learning and the relevance vector machine. *Journal of Machine Learning*, 1:211–244.

Trujillo-Barreto, N. J., Aubert-Vázquez, E., and Valdés-Sosa, P. A. (2004). Bayesian model averaging in EEG/MEG imaging. *NeuroImage*, 21:1300–1319.

Uutela, K., Hämäläinen, M. S., and Somersalo, E. (1999). Visualization of magnetoencephalographic data using minimum current estimates. *NeuroImage*, 10:173–180.

Wagner, M., Fuchs, M., and Kastner, J. (2004). Evaluation of sLORETA in the presence of noise and multiple sources. *Brain Topography*, 16(4):277–280.

ISBN 978-951-22-9142-7 (printed)
ISBN 978-951-22-9143-4 (PDF)
ISSN 1455-0474