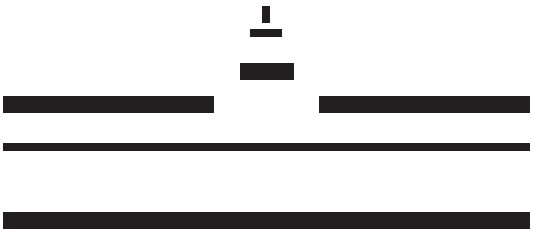


Westfälische Wilhelms-Universität Münster

Diplomarbeit in Mathematik

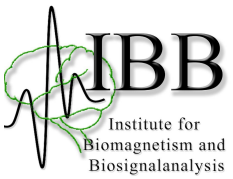
Hierarchical Bayesian Approaches to the Inverse Problem of EEG/MEG Current Density Reconstruction

eingereicht von

Felix Lucka

Münster, 10. März, 2011

Fachbereich 10 Mathematik und Informatik



Gutachter:

Prof. Dr. Martin Burger

Institut für Numerische und Angewandte Mathematik

Priv.-Doz. Dr. Carsten Wolters

Institut für Biomagnetismus und Biosignalanalyse

Abstract

This thesis deals with the inverse problem of EEG/MEG source reconstruction: The estimation of the activity-related ion currents by measuring the induced electromagnetic fields outside the skull is a challenging mathematical inverse problem, as the number of free parameters within the corresponding forward model is much larger than the number of measurements. Additionally, the problem is ill-conditioned due to the smoothing propagation characteristics of the fields through the human tissue. The thesis is devoted to the introduction of a special class of statistical models, called hierarchical Bayesian models to overcome both obstacles. For this sake, it consists of four main parts: The mathematical modeling and challenges of bioelectromagnetism, a theoretical introduction of the model, the algorithmical aspects of the implementation and their practical use and properties within simulation studies. Technically, a focus of interest is on a certain class of inference algorithms that are based on alternated conditional walks through the parameter space. The forward computation will be done with a realistic high resolution finite element (FE) model of a human head.

Acknowledgments

I want to thank everybody who made this thesis and my studies possible, especially:

- ⋆ Carsten Wolters, for introducing me into this thrilling field of research and for providing me with the needed tools and equipment.
- ⋆ Martin Burger, for being open-minded towards this project, for discussion and in particular for sticking to the "quest for the retrieval of the MAP estimate's honor". It proved to be the right intuition in the end.
- ⋆ My parents for their mental and financial support at all times.
- ⋆ Meinen Großeltern für ihre finanzielle Unterstützung.
- ⋆ All my dear friends for trying to maintain my work-life balance.
- ⋆ Everybody who had to correct this document: Sven, Lars, Behrend and last but not least, Esther.

I am grateful that I received a scholarship from the German National Academic Foundation (Studienstiftung des deutschen Volkes) including financial and ideational support.

Studienstiftung
des deutschen Volkes

Contents

List of Figures

List of Tables

Notation and Abbreviations

Most of the notation and abbreviations will be introduced in the corresponding chapters, this listing should serve as a reference for later look-up.

General notation:

A	A *random variable* called "A": $A : \Omega \to X$
a	The concrete *realization* $a \in X$ of the random variable A.
A	A *linear operator* called "A" with no connection to the former two objects
$\tilde{A}$, $\tilde{a}$, $\tilde{\mathrm{A}}$, ...	The corresponding objects in the pseudo source framework: 2.4.3
$A \sim \ldots$	A follows a ... distribution or A is distributed like ...
$\mathrm{A}x \overset{ls}{=} b$	The linear system is solved in a least-squares sense: A.1.1
Id_n	The *identity matrix* in n dimensions
$\mathcal{N}(\mu, \Sigma)$	The *multivariate normal probability distribution* with mean $\mu \in \mathbb{R}^n$ and symmetric, positive semi-definite covariance matrix $\Sigma \in \mathbb{R}^{n \times n}$
$\mathcal{N}_n(\mu, \Sigma)$	The same as the former one, with an *explicit notation of the dimension.*
$\mathcal{N}(x, \mu, \Sigma)$ $\mathcal{N}_n(x, \mu, \Sigma)$	The *value* of the former distributions at x.

Frequently Used Abbreviations:

AO	Alternated optimization: 3.2
AO_MAP	Alternated optimization for MAP approximation: 3.3
AS	Alternated sampling: 3.2
AS_CM	Alternated sampling for CM approximation: 3.3
CDR	Current density reconstruction: 1.3.1
CGLS	Conjugate gradient least squares 3.5
CM	Conditional mean: 2.2
cmAO_MAP	Conditional mean initialized AO_MAP: 3.3
COME	Center of mass error: 1.3.3
DLE	Dipole localization error: 1.3.3
EEG	Electroencephalography
EMD	Earth mover's distance: 1.3.3
FEM	Finite element method
HBM	Hierarchical Bayesian model/modeling: 2.4.2
IAS	Iterative Alternating Sequential: 3.4
MAP	Maximum a-posteriori: 2.2
McmAO_MAP	Multiple conditional mean initialized AO_MAP: 3.3
MCMC	Markov chain Monte Carlo: 3.1
MEG	Magnetoencephalography
MNE	Minimum norm estimate: 1.3.2
MRI	Magnetic resonance imaging
sLORETA	Standardized low resolution brain electromagnetic tomography: 1.3.2
SNR	Signal to noise ratio
SP	Spatial dispersion: 1.3.3
WMNE	Weighted minimum norm estimate: 1.3.2

Important symbols that have a fixed meaning within the whole thesis:

$\Omega \subset \mathbb{R}^3$	The head volume: 1.2
$j^{imp} : \Omega \to \mathbb{R}^3$	Impressed current density: 1.2
$\mathcal{J} \subset \mathcal{D}'(\Omega; \mathbb{R}^3)$	Source model: 1.2
$\mathcal{L}^{em} : \mathcal{J} \to L^2(\partial\Omega) \times C^\infty(\mathbb{R}^3 \backslash \bar{\Omega})^2$	Electromagnetic forward operator: 1.3.1
$\mathcal{J}_n \subset \mathcal{J}$	Finite dimensional subspace defining the discretization: 1.3.1
k	Number of source space nodes: 1.3.1
r_i, $i = 1, \ldots, k$	Location of the source space nodes: 1.3.1
d	Number of basis functions at a single location: 1.3.1
$n = kd$	Total number of source space basis functions: 1.3.1
$s \in \mathbb{R}^n$	Coefficients of the basis functions for the CDR: 1.3.1
$s_{i*} \in \mathbb{R}^d$, $i = 1, \ldots, k$	Coefficients of the basis functions for the CDR for a single location i: 1.3.1
m	Number of measurement sensors: 1.3.1
$\mathrm{L} \in \mathbb{R}^{m \times n}$	Lead-field matrix: 1.3.1
$b \in \mathbb{R}^m$	Measurement vector: 1.3.1
$\lambda \in \mathbb{R}^+$	Regularization parameter: 1.3.2
$\Sigma_\varepsilon \in \mathbb{R}^{m \times m}$	Covariance matrix of the measurement noise: 1.3.2
$\Sigma_s \in \mathbb{R}^{n \times n}$	Covariance matrix of the source activity: 1.3.2
$\sigma^2 \in \mathbb{R}^+$	Variance of the measurement noise: 2.2
$p_{like}(b\|s) \in \mathbb{P}(\mathbb{R}^m)$	Likelihood density: 2.2
$p_{prior}(s) \in \mathbb{P}(\mathbb{R}^n)$	Prior density: 2.2
$p_{post}(s\|b) \in \mathbb{P}(\mathbb{R}^n)$	Posterior density: 2.2
$\hat{s}_{\mathrm{MAP}} : \mathbb{P}(\mathbb{R}^n) \to \mathbb{R}^n$	MAP estimate: 2.2
$\hat{s}_{\mathrm{CM}} : \mathbb{P}(\mathbb{R}^n) \to \mathbb{R}^n$	CM estimate: 2.2
$\boldsymbol{\gamma} \in \mathbb{R}^h$	Hyperparameters: 2.4.2
$p_{hyper}(s) \in \mathbb{P}(\mathbb{R}^h)$	Hyperprior density: 2.4.2
$\mathrm{C}_i \in \mathbb{R}^{n \times n}$, $i = 1, \ldots, h$	Covariance component: 2.4.3
$\mathcal{C} \subset \mathbb{R}^{n \times n}$	Set of covariance components: 2.4.3
h	Number of covariance components: 2.4.3
$f_i : \mathbb{R} \to \mathbb{R}$:	Single hyperprior energy: 2.4.3
$\mathrm{C}^a \in \mathbb{R}^{k \times k}$	Activity covariance components: 2.4.3
$\mathrm{C}^c \in \mathbb{R}^{n \times n}$	Current covariance components: 2.4.3
ϱ_i, $i = 1, \ldots, h$	The rank of C_i: 2.4.3
g	The dimension of the pseudo source space: 2.4.3
A_i, $i = 1, \ldots, h$	The Cholesky factor of C_i: 2.4.3
$\Sigma_b \in \mathbb{R}^{m \times m}$	The total measurement covariance: 2.4.3
Q and R	Burn-in and sample size of the AS_CM scheme: 3.3
T	Number of iterations of AO-based schemes: 3.3
U	Number of seed points for the McmAO_MAP algorithm: 3.3
α and $\beta \in \mathbb{R}^+$	Shape and scale parameter of the generalized gamma distribution: 4.2

Introduction

Electroencephalography (*EEG*) and *magnetoencephalography* (*MEG*) recordings are used in a wide range of applications today, ranging from clinical routine to cognitive science. One aim in EEG and MEG is to reconstruct brain activity by means of non-invasive measurements. This poses challenging mathematical problems: Simulating the field distribution on the head surface for a given current source in the brain is called the EEG/MEG *forward problem* and will be introduced in section 1.2. The reconstruction of the so-called *primary* or *impressed* currents is called the *inverse problem* of EEG/MEG and will be introduced in section 1.3. In its generic formulation, the inverse problem lacks a unique solution: Infinitely many source configurations - often with extremely different properties - can explain the measured fields. All inverse methods rely on the usage of a-priori information on the source activity to choose a particular solution from the set of possible solutions. This a-priori information can reflect computational constraints as well as neurological considerations. Nevertheless, since the problem is heavily under-determined, the results of the different methods for one and the same measurement data differ considerably. Up to date, there is no universal inverse method available: Most methods work well for certain source-configurations while failing to recover others. Therefore, a careful examination of the performance of the methods for different source configurations is still mandatory.
Hierarchical Bayesian modeling is a way to express this a-priori information by modeling the source activity in an explicit but stochastic way. This construction recently emerged as a unifying theoretical framework for EEG/MEG source imaging, comprising most previously established methods as well as offering promising new methods. Chapter 2 outlines the ideas behind this approach, Chapter 3 deals with the algorithms to implement it for practical applications, and in Chapter 4 simulation studies on its performance for certain source scenarios are carried out.

1 Basics of EEG/MEG Source Reconstruction

1.1 Neurophysiological Generators of the EEG/MEG Signals

The human brain is a highly complex organ monitoring and controlling a large number of functions of the human body. Its elementary functional units are approximately 10^{11} electrically excitable cells, called *neurons*. These neurons communicate with each other over $\sim 10^{15}$ *synaptic connections* by means of electrochemical transmission: The firing neuron (called *pre-synaptic* neuron) creates an *action potential* (a rapid change in the electrical cross-membrane potential) which propagates down the neuron's *axon* to a synaptic connection to the *dendrites* of the receiving neuron (called *post-synaptic neuron*). The signal transmission at the synaptic connection is carried out by the release and reception of chemical *neurotransmitters*. These neurotransmitters cause an electric current within the dendrite and the body of the post-synaptic neuron. The accumulation of those currents (called *summation*) can cause the post-synaptic neuron to generate an action potential, and the signal is passed on. The intracellular or extracellular flow of ion currents due to the electric potential and its changes produce electromagnetic fields which propagate through the body's tissue. In principle, they can be measured outside of the skull. However, practically only specific currents produce measurable signals (see Okada, 1993 for a discussion of the limitations of EEG and MEG): The ion currents associated with the action potential are too fast, unstable and their multipole expansion is dominated by the quadrupole term, which makes them hard to detect in a certain distance (see, e.g., Jackson, 1998). The post-synaptic potential is often stable on the timescale of milliseconds and its multipole expansion is dominated by the dipole term. If many neighboring neurons with similar orientations are simultaneously in the post-synaptic excitation state, their impressed currents as well as the ohmic volume current compensating for the charge displacement lead to an electromagnetic field measurable on the outside of the skull. The main contribution to the EEG compared to MEG signal slightly differs: The EEG signal is mainly produced by the extracellular volume currents whereas the MEG signal is mainly produced by the intracellular currents (magnetic field components generated by volume currents tend to cancel out). As a consequence, EEG signals strongly dependents on the surrounding tissue's conductivity, whereas MEG signals are less influenced by that. The need for a large patch of similar oriented neurons to produce a measurable field explains why the mayor contribution to the EEG/MEG signal originates from the $\sim 10^{10}$ *pyramidal cells* (see Nicholson and Llinas, 1971): These neurons form layers where a large number of cells are oriented in a similar way. They are found mainly in the cortical areas

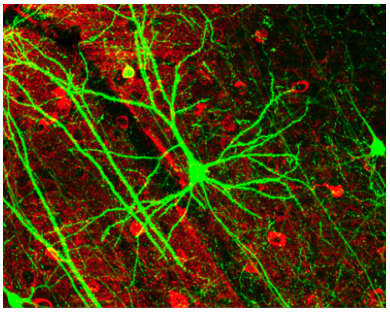


Figure 1.1: Confocal image of pyramidal cell in mouse cortex

Source: Wikimedia Commons, file: GFPneuron.png

of the brain, but also in the Hippocampus and the Amygdala. More details on this topic can be found in Nunez and Srinivasan (2005).

1.2 The Forward Problem

The physical phenomena of electromagnetic fields produced by living cells, tissue or organisms is called bioelectromagnetism (see Malmivuo and Plonsey, 1995 for a complete introduction). The mathematical modeling of this phenomena will be sketched in the following:
The physical basis to start with are *Maxwell's equations* and the *material equations* (see, e.g., Jackson, 1998). These are four coupled, non-linear and time-dependent PDEs for the field $\mathbf{E}(r)$ and the magnetic field $\mathbf{B}(r)$ in the most general case. As a second step, some simplifying assumptions are used that reduce the complexity of the problem (see Plonsey and Heppner, 1967; Sarvas, 1987; Hämäläinen et al., 1993 for details):

- ⋆ *Primary- and volume currents*: A source current model that separates the whole current density into two parts: A *primary* or *impressed* current, generated actively by the electrochemical processes in the exited cell (cf. 1.1), and dependent on the microscopic details in the vicinity of the cell, and a passive current in the surrounding volume that compensates for the net charge displacement caused by the primary current and is determined by the macroscopic conductivity.
- ⋆ *Non magnetic material*: The magnetic susceptibility of the body's tissue is zero, thus $\mu = \mu_0$.
- ⋆ *Linearity*: The body's tissue is a passive conductor.
- ⋆ *Quasistatic approximation*: The temporal changes of the fields are small compared to their spatial propagation velocity. The tissue is (temporally-) passive, i.e., time-independent and no inductance effects occur.
- ⋆ *Charge-free*: In the body's tissue, no macroscopic charge distributions can aggregate.

An additional assumption made by some approaches and explicitly avoided by others is the electric *isotropy* of all tissues. The method we will use for our forward simulation, namely the *finite element method* (*FEM*), can explicitly account for anisotropy. The quasistatic approximation allows to consider the (scalar) electric potential Φ with $\mathbf{E}(r) = -\nabla\Phi(r)$ instead of $\mathbf{E}(r)$. Furthermore, the concept of impressed currents allows to compute $\mathbf{B}(r)$ directly by Biot-Savart's law, once Φ is known. We can state the *direct* or *forward* problem (in the classical formulation) as:

Definition 1 (Direct problem) *Let $\sigma(r) \in C^1(\Omega; \mathcal{S}^2\mathbb{R}^3)$ be the conductivity and $j^{imp}(r) \in C^1(\Omega; \mathbb{R}^3)$ a primary current density in a bounded, simply connected domain $\Omega \subset \mathbb{R}^3$ with a smooth surface $\partial\Omega$. The forward problem of calculating the electric potential $\Phi \in C^2(\Omega; \mathbb{R}) \cap C^1(\partial\Omega; \mathbb{R})$ is given by solving:*

$$\begin{aligned} \nabla \cdot (\sigma \nabla \Phi) &= \nabla \cdot j^{imp} && \text{in } \Omega \\ n \cdot (\sigma \nabla \Phi) &= 0 && \text{on } \partial\Omega \text{ (no-penetration condition)} \\ \int_{\partial\Omega} \Phi \cdot dS &= 0 && \text{(fix ground potential)} \end{aligned} \tag{1.1}$$

The magnetic field $\boldsymbol{B}$ can then be computed by (Biot-Savart):

$$\boldsymbol{B}(r') = \frac{\mu_0}{4\pi} \int_\Omega \left(j^{imp}(r) - \sigma(r) \cdot \nabla\Phi(r) \right) \times \frac{r' - r}{\|r' - r\|^3} \, d^3r \qquad \text{for } r \in \mathbb{R}^3 \backslash \bar{\Omega} \tag{1.2}$$

However, classical solutions can only be found for quite restrictive assumptions, and their value with regard to more realistic modeling is limited. We will rather have to rely on the weak or even distributional formulation of (1.1) dependent on the regularity of $j^{imp}(r), \sigma(r)$ and $\partial\Omega$ we assume or can provide by the models used. In principle, three points that depend on each other have to be considered:

1. A *source model* for j^{imp}: How can we model the macroscopic current-flows, i.e., to which mathematical space $\mathcal{J} \subset \mathcal{D}'(\Omega; \mathbb{R}^3)$ do we restrict j^{imp}?

2. A *volume conductor model*: How can we model the dielectric properties of the different tissues, i.e., how do we define $\sigma(r)$?

3. A *method* for solving (1.1): Which method is able to deal with our assumptions?

1. Source Model: A commonly accepted mathematical model for the impressed ion currents in the post-synaptic densities is to replace the real current density by a mathematical current dipole with an adequate dipole moment (see Brazier, 1949 for the introduction, and de Munck et al., 1988 for an examination of this approach). Furthermore, many of those current dipoles representing microscopic current flows with the same orientation are replaced by an *equivalent current dipole* $q_{dip}\delta(r - r_{dip})$. Location, amplitude and orientation of this dipole are chosen in such a way that the dipole represents the dipole moment of the resulting macroscopic current flow in the surrounding volume, and the total current $j^{imp}(r)$ is given by a linear combination of such dipoles. This approach offers many analytical advantages, it allows, e.g., for the derivation of asymptotic formulas for Φ for simplified volume conductors. Furthermore, this local model is regarded as an adequate representation of focal brain activity. Nevertheless, since

$$\delta(r) \in H^{-3/2-\varepsilon}(\Omega)\ \forall \varepsilon > 0 \quad \text{and} \quad D^\alpha \delta(r) \in H^{-3/2-|\alpha|-\varepsilon}(\Omega)\ \forall \varepsilon > 0,$$

the sources are modeled very irregular, which is problematic for the weak formulation. Furthermore, the numerical treatment of the singularities with the finite element method imposes theoretical and practical challenges (Wolters et al., 2007; Drechsler et al., 2009). For finite element analysis it might be advantageous to have a less singular current model, e.g., $j^{imp}(r) \in H(\mathrm{div}, \Omega, ; \mathbb{R}^3)$. *Whitney forms* are a family of differential forms on a simplical mesh that provide a hierarchy of basis functions that can be used to represent different electromagnetic quantities (Tanzer et al., 2005), and automatically respect the physically relevant continuity conditions across element boundaries. If the mesh is fine enough, the support of these basis functions is considerably smaller than the spatial extent of source activity that is needed to produce a measurable signal (cf. 1.1). Thus focal activity is well represented by these continuous basis functions as well. For the application of such an approach see, e.g., Pursiainen (2008); Calvetti et al. (2009).

2. Volume Conductor: Creating a realistic, individual head model for each patient or proband is a complex and costly task: The geometries and different tissue compartments have to be segmented from different *magnetic resonance imaging* (*MRI*) or *computerized tomography* (*CT*) scans (CT only for clinical indications). The automated registration of the different scans of one subject and their segmentation are still topics of research. The assignment of conductivities to the segmented compartments is a second challenge since, as studies show, some compartment's conductivities show inter- and intra-subject variations. Recently, the recordings of an MRI-based technique, called *diffusion weighted MRI* (DW-MRI) have been used to determine those conductivities for the white matter compartment (Tuch et al., 2001). For these reasons, simplified head models are used in many applications. The most common ones are realistically shaped head models with three compartments (scalp, skull, brain) and isotropic conductivities and models consisting of concentric spheres whose radii are matched to the human head. It is still a topic

of research in which situations their use introduces negligible errors compared to more realistic models: The importance of using realistically shaped three isotropic compartment head models was shown by Hämäläinen and Sarvas (1989); Roth et al. (1993); Cuffin (1996). Yet, the isotropic three-compartment head model still ignores the three-layeredness of the skull (e.g., Sadleir and Argibay, 2007), whose influence on EEG was shown in Dannhauer et al. (2009, 2010), skull holes and inhomogeneities (e.g., Ollikainen et al., 1999), white matter conductivity anisotropy (e.g., Wolters et al., 2006; Hallez, 2008) and conductivity changes in the vicinity of the source (e.g., Wolters et al., 2005; Rullmann et al., 2009). A general overview on this topic is given in Wolters and de Munck (2007).

3. Solution Method: As mentioned above, an analytical solution is only possible for simplified geometries. Using realistic geometries, only numerical methods are applicable. The most commonly used are *boundary element* (*BE*) methods or *finite element* (*FE*) methods. BE methods only need the surfaces of the compartments as a head model (most often extracted from MRI recordings), but cannot account for anisotropic conductivities. See Hackbusch (1997) for an introduction to this approach and, e.g., Hämäläinen and Sarvas (1989); Kybic et al. (2005) for the application to EEG/MEG. FE methods need a discretization of the whole brain volume into elementary geometries but can handle anisotropy and complex anatomical details. On the other hand, as discussed above, an irregular modeling of the source activity may impose theoretical and practical challenges for the FE method. For general introductions to finite element analysis, see, e.g., Braess (2007) and Brenner and Scott (2008).

1.3 The Inverse Problem

1.3.1 Formulation and General Properties

The *physical* inverse problem in EEG/MEG-based source reconstruction can be stated as:

> "*Estimate brain activity non-invasively by measuring the induced electromagnetic fields outside of the skull.*"

Using the notations from Section 1.2 we will derive the corresponding mathematical formulation of this problem:
Formally, since Φ is related to j^{imp} via the linear operator $\nabla \cdot \sigma \nabla$ with Neumann boundary conditions (cf. (1.1)), it can be expressed in terms of the corresponding *Neumann–Green's function* $\mathcal{G}_N(r', r)$ (Calvetti et al., 2009), i.e., Φ is given by a *Fredholm integral equation of first kind*:

$$\Phi(r') = \int_\Omega \mathcal{G}_N(r', r)\, \nabla \cdot j^{imp}(r)\, dr \tag{1.3}$$

In general, since $\nabla \cdot \sigma \nabla$ is not translation invariant due to the inhomogeneous σ, $\mathcal{G}_N(r', r)$ is not a convolution kernel. Equation (1.3) defines a linear operator $\mathcal{L}^e : \mathcal{J} \to L^2(\partial\Omega)$ such that:

$$\Phi = \mathcal{L}^e j^{imp}, \quad \text{with} \quad \mathcal{L}^e[j^{imp}](r') := \int_\Omega \mathcal{G}_N(r', r) \nabla \cdot j^{imp}(r)\, dr$$

Using (1.2) we can define a linear operator $\mathcal{L}^m : \mathcal{J} \to C^\infty(\mathbb{R}^3 \backslash \bar{\Omega})^3$ such that:

$$\mathbf{B} = \mathcal{L}^m j^{imp}, \quad \text{with} \quad \mathcal{L}^m[j^{imp}](r') := \frac{\mu_0}{4\pi} \int_\Omega \left(j^{imp}(r) - \sigma(r) \cdot \nabla \mathcal{L}^e[j^{imp}](r') \right) \times \frac{r' - r}{\|r' - r\|^3}\, d^3 r$$

Normally, not $\mathbf{B}$ but the *magnetic flux* is measured (by magnetometers) and spatial changes of that (scalar) quantity (by axial- or planar gradiometers). The magnetic flux is the (scalar) surface integral of $\mathbf{B}$ over the area spanned by the sensor coil, which becomes the normal component

of **B** in the limiting case of vanishing coil area. Gradiometers measure the spatial derivatives of the normal component of **B** in that case. In the following, we will assume point-like sensors for the magnetic field as well ((1.1) assumes point-like sensors for the electrical potential, in contrast to *complete electrode models*, see e.g., Somersalo et al., 1992; Pursiainen, 2008). To formalize the measurement of the magnetic flux in the continuous setting, we assume that a normal direction field $n(r')$ and a gradiometer direction field $v(r')$ are given in $\mathbb{R}^3\backslash\bar{\Omega}$, indicating a normal direction and the gradiometer direction for every potential sensor location. We then define

$$\begin{aligned} \mathcal{L}^m_n : \mathcal{J} \to C^\infty(\mathbb{R}^3\backslash\bar{\Omega}) \quad \text{via} \quad \mathcal{L}^m_n[j^{imp}](r') &:= \langle n(r'), \mathcal{L}^m[j^{imp}](r')\rangle \\ \mathcal{L}^m_{n,v} : \mathcal{J} \to C^\infty(\mathbb{R}^3\backslash\bar{\Omega}) \quad \text{via} \quad \mathcal{L}^m_{n,v}[j^{imp}](r') &:= \partial_{v(r')}\langle n(r'), \mathcal{L}^m[j^{imp}](r')\rangle \end{aligned}$$

Hence we have three operator equations for the forward mapping, which we will combine by introducing a single electromagnetic forward operator $\mathcal{L}^{em}$ which maps $j^{imp} \in \mathcal{J}$ to the measurements $u \in L^2(\partial\Omega) \times C^\infty(\mathbb{R}^3\backslash\bar{\Omega})^2$:

$$u := (\Phi, \mathrm{B}_n, \mathrm{B}_{n,v})^t = (\mathcal{L}^e j^{imp}, \mathcal{L}^m_n j^{imp}, \mathcal{L}^m_{n,v} j^{imp})^t =: \mathcal{L}^{em} j^{imp} \tag{1.4}$$

Definition 2 (Continuous inverse problem) *Let $\sigma(r)$ be a volume conductor model for Ω and $\mathcal{J} \subset \mathcal{D}'(\Omega, \mathbb{R}^3)$ a source model. For given measurements u, the (continuous) inverse problem of EEG/MEG is to find the impressed current $j^{imp} \in \mathcal{J}$ satisfying (1.4).*

Most practical methods to solve the inverse problem rely on choosing a finite dimensional subspace $\mathcal{J}_n \subset \mathcal{J}$ on which the problem is formulated in a discrete setting. There are two main categories of subspace models:

- ⋆ *Focal current models*: The current consists of a small (either predefined or flexible) number of elementary sources having arbitrary location and orientation within the source compartment (usually dipoles are chosen as a source model, cf. Section 1.2).
- ⋆ *Distributed current models*: The current consists of a large number of focal elementary sources having a fixed location and orientation within the source compartment. This is intended as a localized discretization of the underlying continuous current distribution and is called *current density reconstruction* (*CDR*).

Using focal current models to solve the inverse problem leads to methods aiming to find the best number, location, and magnitude of the elementary sources used. This can be done in a least-squares sense to fulfill the data (e.g., Mosher et al., 1992), or in a probabilistic sense (Jun et al., 2008). The resulting source model usually comprises far less parameters than measurements available, thus the inverse problem is usually well-posed in the sense of Hadamard (Hadamard, 1923), i.e., it has a unique solution and the solution depends continuously on the data. When the number of sources is unknown or the current distribution might have a larger spatial extent, focal current models are not suitable. We will restrict ourselves to the discussion of CDRs in this thesis. Assume that we have k locations r_i, $i = 1, \ldots, k$ within the brain and place d focal elementary sources with different orientations $j_{i,l}$, $i = 1, \ldots, k$, $l = 1, \ldots, d$ (i.e., dipole or dipole-like sources, cf. Section 1.2) at each of these locations. The case $d = 1$ is most often chosen when reliable information about the local normal direction of the gray matter layer is given (cf. Section 1.1), and is therefore called *normal constraint*, whereas $d = 3$ is chosen, if no such information is available. The normal constraint leads to a better relation between the number of measurements and the number of parameters that have to be estimated. However, since the measurements are extremely sensitive to the orientation of a single source, invalid information on the normal directions can have serious negative impact on the solution. Now, $n = d \cdot k$ and an element j of $\mathcal{J}$ can be approximated in $\mathcal{J}_n$ by a linear combination of the basis functions

$j_{i,l}$. The corresponding coefficients $s \in \mathbb{R}^n$ will become the main parameters of interest in the following (also called *sources*):

$$j \approx \sum_{i=1}^{k} \left(\sum_{l=1}^{d} s_{i,l}\, j_{i,l} \right)$$

To simplify some notation, the d-dimensional $i-th$ *source vector* s_{i*} contains all d coefficients of a single location: $s_{i*} = (s_{i,1}, \ldots, s_{i,d})^t$. Hence the ℓ_2 norm of s_{i*} is the source amplitude at location r_i. Furthermore, the double-index (i,l) will be replaced by a single index $((i-1)d+l)$, i.e., the d basis functions of a single location are numbered consecutively and $s := (s_{1*}^t, \ldots, s_{k*}^t)^t$. That will ease the notation in the later parts, as the groups of these d basis functions will be the atomic components on which the hierarchical extension of the model will be build up.
Now let $r_i' \in \partial\Omega$, $i = 1, \ldots, m_e$ be the locations of the electrodes, $r_i' \in \mathbb{R}^3 \backslash \bar{\Omega}$, $i = m_e+1, \ldots, (m_e + m_{mag})$ be the locations of the magnetometers, and $r_i' \in \mathbb{R}^3 \backslash \bar{\Omega}$, $i = (m_e + m_{mag}) + 1, \ldots, (m_e + m_{mag} + m_{grad}) =: m$ be the locations of the gradiometers.

Definition 3 (Lead-field matrix) *The matrix elements of $\mathcal{L}^{em}$ with respect to r_i', $i = 1, \ldots, m$ and j_l, $l = 1, \ldots, n$ define the lead-field or gain matrix* $\mathrm{L} \in \mathbb{R}^{m \times n}$

As a result, the columns of the lead-field matrix represent the electric potential and/or magnetic field strength or gradient measured at the sensors caused by the corresponding single elementary source of the discretization (see Hämäläinen and Ilmoniemi, 1984; Sarvas, 1987; Hämäläinen et al., 1993 for the introduction of this concept). Now, the field-measurements $b := u(r_i')_{i=1,\ldots,m}$ caused by s can be calculated via:

$$b = \mathrm{L}\, s \tag{1.5}$$

Definition 4 (Discrete inverse problem) *Let* L *be the lead-field matrix of a discretization of a continuous problem as in definition 2. For given measurements b, the discrete inverse problem of EEG/MEG is to find $s \in \mathbb{R}^n$ satisfying* (1.5).

We will refer to this definition, when speaking of "the" inverse problem in the following. Note that we will mainly restrict ourselves to the solution of the *instantaneous* in contrast to the *dynamical* inverse problem (Kaipio and Somersalo, 2005). This means that we are only looking at one single time slice of the whole data stream.

Remark: Classically, the term *lead-field* is introduced in a different way: Since Φ depends on j^{imp} in a linear way, the evaluation of Φ at a given sensor location r_i' is a linear functional on $\mathcal{J}$. Due to Riesz representation theorem, there is a vector field $\mathcal{L}_i^e : \Omega \to \mathbb{R}^3$ such that:

$$\Phi(r_i) = \int_\Omega \mathcal{L}_i^e(r') \cdot j^{imp}(r')\, dr'$$

$\mathcal{L}_i^e$ represents the sensitivity distribution of the i-th electrode, and is called its (electrical) lead-field. A similar reasoning defines a magnetic lead-field for each magneto- or gradiometer. The discretization of these lead-fields in terms of the n basis functions in source space form the rows of the lead-field matrix. However, this is a sensor-based perspective of the problem, which is not very useful for the framework we will develop. We will rather choose a source location based perspective, i.e., the columns of the lead-field matrix will play an important role. The column of a single basis function will be called its *gain vector*, whereas for a given source location the term *3-gain* will denote the $m \times d$ matrix formed by the gain vectors of the d basis functions in that location.

Important properties of the inverse problem: In this formulation the problem is apparently *ill-posed* (Hadamard, 1923):

- ⋆ The number of dipoles n should be much larger than the number of sensors m. Hence, (1.5) is under-determined.
- ⋆ The formal solution of the (linear) forward equation (1.1) is a Fredholm integral equation of the first kind (cf. (1.3)). The forward operator of the continuous problem is therefore a compact linear operator. A closer analysis reveals that its singular values are decaying even exponentially fast leading to a very strong smoothing of the electromagnetic fields during their propagation through the tissue (Gencer and Williamson, 2002). The corresponding inverse operator is unbounded, thus the inversion is ill-posed. The discrete inverse problem inherits the properties of the continuous problem in form of the properties of the lead-field matrix, which is ill-conditioned. One speaks of an *exponentially-ill-posed* problem and although this fact is seldom discussed in publications, it is at least as important as the previous point (see, e.g., Engl et al., 1996 for a discussion of the properties of inverse problems involving Fredholm operators).
- ⋆ EEG/MEG recordings suffer from very low *Signal-to-Noise-Ratios* (*SNR*), which is especially crucial in combination with the last point.

For CDR in EEG/MEG source reconstruction, two conceptual approaches to overcome these difficulties have been established:

- ⋆ *Global, source space based methods*: *A-priori information* on the global properties of the solution is incorporated in an explicit or implicit way.
- ⋆ *Local, spatial scanning methods/beamforming*: The estimate of the activity at a single location or a small region of interest is optimized concerning a specific quantity while suppressing crosstalk from all other areas. This is repeated for all source locations (the source space is "scanned").

We will not further pursue beamforming techniques here but refer to Sekihara et al. (2001, 2005); Hillebrand et al. (2005); Greenblatt et al. (2005); Sekihara and Nagarajan (2008); Steinsträter et al. (2010) for their use in EEG/MEG source reconstruction. In the following, we will outline the main developments of source space based methods for CDR.

1.3.2 Development of Source Space Based Methods for CDR

Preliminary remark

Two main concepts dominated the treatment of the inverse problem:

- ⋆ *Regularization*: The originally ill-posed problem is *approximated* by a well-posed problem in such a way that the solution of the well-posed problem favors features consistent with available a-priori knowledge on the solution. The degree of approximation is controlled by means of a *regularization parameter* λ, such that $\lambda \to 0$ corresponds to the solution given by a generalized inverse of L (see, e.g., Engl et al., 1996).
- ⋆ *Bayesian statistics*: The high uncertainty and under-determinateness of the problem is explicitly accounted for by formulating the inverse problem as a *statistical estimation problem*. The aim is to make statistical inferences about the real source configuration based on the information given by the measurements and the a-priori knowledge about the underlying brain activity (see, e.g., Kaipio and Somersalo, 2005).

While seeming quite different from the conceptual point of view, they often lead to equal algorithms and solutions, and in these situations, it is a personal favor which description to use. The latest publications tend to favor the Bayesian perception, which often leads to the belief that they represent new ideas whereas regularization based methods such as MNE or LORETA are regarded as "old". However, historically, this is not true: Hämäläinen et al. (1987) introduced the Bayesian perception of the inverse problem at the same time when regularization techniques were first introduced (Sarvas, 1987). In the following, ideas or interpretations from both frameworks were used alternately to improve the methods for CDR. A Bayesian formulation of the inverse problem as a starting point is quite common in the regularization community nowadays, as well. At first glance, this might give the impression that all recent methods rely on ideas from the field of Bayesian statistics, which is not true. In several situations, it was even the other way round: A Bayesian formulation was found for an already existing regularization technique (see Chapter 4 and Section A.1.8).
Since we will focus on the Bayesian approach in Chapter 2, we will use the framework of regularization to introduce some standard methods in the remaining part of this chapter. Revisiting these methods within the Bayesian framework will ease the interpretation of the key concepts thereof.

Chronological Order

Hämäläinen and Ilmoniemi (1984) is the first publication proposing a practical method for CDR, called *minimum norm estimate* (*MNE*):

$$\hat{s}_{\text{MNE},84} = \mathrm{L}^{+}b \overset{A.1.1}{=} \mathrm{L}^{t}(\mathrm{L}\mathrm{L}^{t})^{-1}b$$

This approach uses the Moore–Penrose pseudoinverse A^{+} of a matrix A (Ben-Israel and Greville, 2003), which maps b to the least-squares solution of (1.5). Shortly after, the importance of a regularization of the problem was discussed: Relying on Parker (1977), Sarvas proposed the use of a *truncated singular value decomposition* and the use of the χ^2 criterion to determine the truncation value (Sarvas, 1987). The minimum norm solution normally considered in EEG/MEG source reconstruction (Hämäläinen and Ilmoniemi, 1994) is given by the classical ℓ_2-norm Tikhonov regularization (Tikhonov and Arsenin, 1977):

$$\hat{s}_{\text{MNE}} = \underset{s\in\mathbb{R}^n}{\operatorname{argmin}} \left\{\|b - \mathrm{L}\,s\|_2^2 + \lambda\|s\|_2^2\right\} \overset{A.1.1}{=} (\mathrm{L}^t\mathrm{L} + \lambda\,\mathrm{Id}_n)^{-1}\mathrm{L}^t b \overset{(A.7)}{=} \mathrm{L}^t(\mathrm{L}\mathrm{L}^t + \lambda\,\mathrm{Id}_m)^{-1}b$$

Compared to the original MNE the problem has been approximated by a quadratic minimization problem. Since this has a linear first-order condition the estimate is still given by a linear mapping of the measurements b.
Ioannides et al. (1990) were the first who introduced a weighting of source locations into the inversion process, intending to encode a-priori information on the probability of source activity at certain locations. Dale and Sereno (1993) extended this approach and showed with a statistical argument that in case of normally distributed noise with zero mean and known covariance matrices Σ_ε for the sensors and Σ_s for the sources, the optimal linear estimator for s is given by:

$$\hat{s}_{\text{WMNE}} = \Sigma_s\mathrm{L}^t(\mathrm{L}\Sigma_s\mathrm{L}^t + \Sigma_\varepsilon)^{-1}b \overset{A.1.1}{=} \underset{s\in\mathbb{R}^n}{\operatorname{argmin}} \left\{\|\Sigma_\varepsilon^{-1/2}(b - \mathrm{L}\,s)\|_2^2 + \|\Sigma_s^{-1/2}s\|_2^2\right\} \tag{1.6}$$

We will call this solution *weighted minimum norm estimate* (*WMNE*), although some authors use that term only for special cases of Σ_ε and Σ_s. Another commonly used parameterization of WMNE is given by:

$$\hat{s}_{\text{WMNE}} = \underset{s\in\mathbb{R}^n}{\operatorname{argmin}} \left\{\|b - \mathrm{L}\,s\|_2^2 + \lambda\|\mathrm{W}s\|_2^2\right\} \overset{A.1.1}{=} (\mathrm{W}^t\mathrm{W})^{-1}\mathrm{L}^t(\mathrm{L}(\mathrm{W}^t\mathrm{W})^{-1}\mathrm{L}^t + \lambda\mathrm{Id}_m)^{-1}b$$

Both formulations can be transfered into each other by choosing $\Sigma_\varepsilon = \sigma^2 \mathrm{Id}_m$ and $\Sigma_s = \frac{1}{\lambda}(\mathrm{W}^t\mathrm{W})^{-1}$. The frequently used *LORETA* (Pascual-Marqui et al., 1994) method is a special case of WMNEs aiming at spatially smooth estimates. Due to the well known blurring property of ℓ_2-norm based solutions, ℓ_1-norm based approaches were proposed to recover focal current configurations (termed *minimum current estimate* (*MCE*) Matsuura and Okabe, 1995; Uutela et al., 1999). The *FOCUSS* algorithm (Gorodnitsky et al., 1995) is an *iteratively reweighted least squares* algorithm implicitly minimizing a concave cost function[1] (Gorodnitsky and Rao, 1997). In the original formulation it aims at minimizing $\sum_{s_i \neq 0} \log(|s_i|),\ s.t.\ \mathrm{L}s = b$. This leads to a highly unstable algorithm converging to a suboptimal local minimum almost always. Practical implementations use modified and relaxed versions minimizing $\|b - \mathrm{L}\,s\|_2^2 + \frac{\lambda}{p}\|s\|_p^p,\ 0 < p \leq 2$. See Wipf (2006); Wipf and Nagarajan (2009, 2010) for a discussion. Another class of methods are re-weighted schemes like *sLORETA* (Pascual-Marqui, 2002) or *dSPM* (Dale et al., 2000) which produce standardized estimates of source activation. As we use sLORETA in our studies in Chapter 4, we will sketch it here: For a given λ, let $\hat{s}_{\mathrm{MNE}}(\lambda)$ be the MNE, then the sLORETA estimate $\varphi_{\mathrm{sLORETA}}$ at location r_i is computed by:

$$\varphi_{\mathrm{sLORETA},i} = \hat{s}^t_{\mathrm{MNE},i*}(\lambda)\left(\mathrm{R}_{(i,i)}\right)^{-1}\hat{s}_{\mathrm{MNE},i*}(\lambda) \quad \text{where} \quad \mathrm{R} := \mathrm{L}^t(\mathrm{L}\mathrm{L}^t + \lambda\,\mathrm{Id}_m)^{-1}\mathrm{L}$$

is the *resolution matrix* of the MNE, and $R_{(i,i)}$ its $d \times d$ diagonal block belonging to the source location i. Many spatio-temporal schemes have been proposed as well, but will not be discussed here (e.g., Schmitt and Louis, 2002; Schmitt et al., 2002).
The class of weighted minimum norm solution schemes led to a generalization of the methods: Σ_s encodes our modeling of the source activity, but can also be used to compensate for unwanted features. For instance, three possible choices considered by various studies are:

- ⋆ $\Sigma_s = \mathrm{diag}(\mathrm{L}^t\mathrm{L})^{-1} := (\mathrm{W}^t\,\mathrm{W})^{-1}$, with $\mathrm{W} = \mathrm{diag}(\|\mathrm{L}_{(\cdot,1)}\|, \ldots, \|\mathrm{L}_{(\cdot,n)}\|)$. This was proposed to compensate for a phenomena called *depth bias* which will be introduced in Section 4.1. See, e.g., Ioannides et al. (1990) for this approach.

- ⋆ $\Sigma_s = (\mathrm{W}^t\,\mathrm{B}^t\,\mathrm{B}\,\mathrm{W})^{-1}$, with B being the discrete Laplacian. This models spatial correlations between the source locations, and features a weighting as in the previous proposal (Pascual-Marqui et al., 1994).

- ⋆ $\Sigma_s = \mathrm{Id}_n + \alpha\mathrm{A}_{\mathrm{fMRI}}$, where $\mathrm{A}_{\mathrm{fMRI}}$ is a diagonal matrix encoding the activity measured by a *functional magnetic resonance imaging* (*fMRI*) scan. This is usually done in a binarized form, i.e., $(\mathrm{A}_{\mathrm{fMRI}})_{(i,i)} = 1$ if the fMRI activation exceeds a certain threshold and 0 else (Liu et al., 1998; Phillips et al., 2002a; Henson et al., 2010). The parameter α weights the impact of the fMRI information (see Liu et al., 2002, 2006a,b for a discussion).

Since all three covariances encode valuable information and their use improves certain features of the solution over the ordinary MNE, the idea to combine them came up (Phillips et al., 2002a):

$$\Sigma_s = \sum_{i=1}^{h} \gamma_i \mathrm{C}_i, \quad \mathrm{C}_i \in \mathcal{C}, \quad |\mathcal{C}| = h,$$

where $\mathcal{C}$ is a predefined set of covariance matrices C_i like the ones discussed before, termed *covariance components* from now on, and γ_i are *hyperparameters* controlling the weighting between them. This construction immediately poses the question of how to choose these hyperparameters. Whereas early approaches tried to determine the optimal weighting by running simulation studies (e.g., Liu et al., 1998), Phillips et al. (2002a) suggested to estimate them from the data, too. The theoretical consolidation of this approach leads to *hierarchical* Bayesian models, which is the main topic of this thesis and will be introduced in detail in the following chapter. Sato

[1] It is actually a modified Newton's method

et al. (2004) were the first to come up with such a model for EEG/MEG source reconstruction, further important contributions are Friston et al. (2002b,a); Sato et al. (2004); Phillips et al. (2005), Mattout et al. (2006); Nummenmaa et al. (2007b,a); Wipf et al. (2007); Friston et al. (2008); Wipf and Nagarajan (2009); Calvetti et al. (2009).

1.3.3 Validation, Performance Measures and Inverse Crimes

The appropriate validation of inverse methods in EEG/MEG source-reconstruction is a difficult task, since the spatio-temporal properties of real brain activity related source currents are not sufficiently known yet. There are three most commonly used validation means:

1. *Real-data:* The validation with data originating from real experiments has the methodical advantage that one does not need to impose artificial assumptions concerning the underlying source model. The disadvantage is of course that no quantitative evaluation of the results can be made. Qualitative evaluation needs the expertise of a neurologist, and for many experiments, even that might not be sufficient.

2. *Synthetic-data:* When using synthetic data produced by an invented source-configuration, it is crucial to avoid an *inverse crime*, i.e., model and reality are identified (Kaipio and Somersalo, 2005). For our problem, one should not produce synthetic data with the same lead-field matrix used for the inversion, which would correspond to the assumption that the real current sources are also restricted to the chosen source space nodes (in fact, one should not use the lead-field matrix-concept for the data-generation at all but use mesh-free forward computations). Strictly, it is even an inverse crime to use the concept of current-dipoles for the data-generation, as it is a modeling assumption, too.

3. *Semi-synthetic-data:* A third way between the two previous ones, is to use real data and real noise as a starting point and to then construct a source-configuration that would lead to similar data. This synthetic data is then mixed with the real noise and used for the inversion, and the results are compared to the constructed source-configuration. See e.g., Friston et al. (2008) for this procedure.

In this thesis, validation by synthetic-data will be carried out to assess the reconstruction properties of different inverse methods. In the following, the measures we will use to evaluate their performance will be introduced. This thesis also aims at a careful examination of their practical value for arbitrary source configurations. A general problem is that the measures used in most publications rely on an inverse crime per se, i.e., real and estimated source are assumed to come from the same space ($\mathbb{R}^n$ in our case). We would need a mapping of the real sources to the space of estimated sources to apply these measures, but this choice has a non trivial impact on the results in any case, and is not very convincing from the theoretical perspective as well. Furthermore, some measures are only applicable for linear inverse methods (they calculate properties of the *resolution matrix*). We will thus only rely on mesh-free measures applicable for any inverse method and underlying real source configuration.
If the real source configuration is modeled as a single dipole, the *dipole localization error* (*DLE*) (e.g., Molins et al., 2008) is the most commonly used measure:

Definition 5 (Dipole localization error (DLE)) *Let $j_r = M \cdot \delta_{r_d}$ be the real source configuration consisting of a dipole at position r_d. Then*

$$d_{DLE}(s, j_r) := \|r_d - r_j\|, \qquad \textit{with} \qquad j = \operatorname*{argmax}_i \{\|s_{i*}\|_2\}\,.$$

For this thesis we will also need a relaxed, continuous version of the DLE. Our proposal for this will be called p^{th} *center of mass error* (*p-COME*):

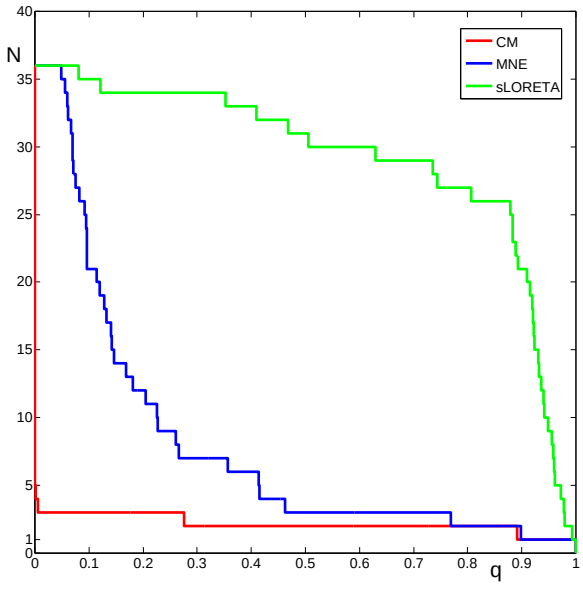


Figure 1.2: The curves of $N = \mathring{f}(s,q)$ for $s = \hat{s}_{\text{MNE}}$ (blue), $s = \varphi_{\text{sLORETA}}$ (green), and $s = \hat{s}_{\text{CM}}$ (red, see 4.2) for a simplified model.

Definition 6 (p^{th} center of mass error (p^{th}-COME))

$$d^p_{COME}(s, j_r) := \|r_d - r_{p\text{-}COM}\|, \quad \textit{where} \quad r_{p\text{-}COM} = \frac{1}{\tau}\sum_{i=1}^{k} \|s_{i*}\|^p \, r_i, \quad \textit{with} \quad \tau = \sum_{i=1}^{k} \|s_{i*}\|^p,$$

and r_d has the same meaning as in the DLE definition.

For $p = 1$ this is the normal mass center of the current distribution, for $p \to \infty$ it converges to the DLE.
Furthermore, we will need a continuous measure how focal or blurred the estimate is, i.e., a measure for the *spatial dispersion*. A standard approach would be to define a threshold q, and count the percentage of sources whose amplitude is above q times the maximal source amplitude $\max\|s_{i*}\|_2$. We will call this measure $\mathring{f}(s,q)$. However, $\mathring{f}(s,q)$ is not continuous, and involves some arbitrariness, since q has to be chosen ad hoc. In Figure 1.2 three plots of $\mathring{f}(s,q)$ as a function of q are depicted for a simplified model geometry. The curves for focal and widespread CDRs show quite obvious differences. We therefore propose to use a normalized version of the area below the curve as a measure for the spatial dispersion:

Definition 7 (Spatial dispersion (SD))

$$\Gamma_{SP} := \frac{1}{(k-1)}\left(\int_0^1 \mathring{f}(s,q)\, dq - 1\right) = \frac{1}{(k-1)}\left(\sum_{i=1}^{k} \frac{\|s_{i*}\|_2}{a_{\star,\infty}} - 1\right), \quad \textit{with} \quad a_{\star,\infty} = \max_j \|s_{j*}\|_2$$

Note that this measure does not compare the spatial spread of real and estimated source, but only yields information about the estimate.
Finally, a measure that combines the aspects of right localization and right spatial dispersion would be desirable. In addition, the proposed localization measures are not easy to extend to more complex patterns of source activity in an intuitive way: The number of possible misfits between real and estimated activity becomes very large, and thus a number of heuristic measures applicable for a fixed number of source dipoles have been proposed. However, in this thesis, we will examine the use of measures based on *optimal transport* to tackle both problems: *Wasserstein metrics* are distance measures between probability distributions (Ambrosio et al., 2008):

Definition 8 (Wasserstein metric) *Let μ and ν be two probability measures on a Radon space (Ω, d) that have a finite p^{th} moment for some $p \geq 1$. Then the p^{th} Wasserstein distance $W_p(\mu, \nu)$ is defined as:*

$$W_p(\mu, \nu) = \left(\inf_{\gamma \in \Gamma(\mu,\nu)} \int_{\Omega \times \Omega} d(x, y)^p \, d\gamma(x, y) \right)^{1/p}, \tag{1.7}$$

where $\Gamma(\mu, \nu)$ denotes the class of all transport maps, i.e., measures on $\Omega \times \Omega$ with marginals μ and ν.

The intuitive explanation behind this quantity dates back to Monge who published it in 1781 as an optimal transport problem: The idea is to think of the first probability measure as an amount of sand piled on a space Ω, and of the second as a hole with the same size. For a given distance function d, the minimum-cost transport of the sand into the holes has to be found (where the cost of a single assignment is understood as classical physical work in terms of distance times amount of sand). This minimal cost is the Wasserstein distance between the two measures. Due to that analogy, it is also often called *earth mover's distance* (*EMD*). In the following, we will speak of the p^{th}-EMD when referring to the p^{th} Wasserstein distance (and omit the prefix for the case $p = 1$). The concrete implementation will be discussed in Section 3.7.
Wasserstein metrics have three big advantages for the topic we are dealing with:

1. They are sensitive to both mis-localization and mismatches in spatial extent.

2. The concept of a measure is a very flexible tool for our purpose: It allows to compare (normalized) source estimates of arbitrary form, e.g., dipole fits or CDRs with arbitrary scaling (some CDR produce statistical values rather than real current amplitudes) with real sources of arbitrary form (including continuous formulations).

3. In simple scenarios, it reduces to intuitive measures (e.g., for two dipoles, the spatial distance between them).

However, it still looks like a rather abstract concept for the practical task we are aiming at, but the lack of a more simple measure that is commonly accepted may be rooted in the fact that the task is not that simple after all: A good measure has to mimic the way source estimates from inverse methods are interpreted by the user, and compare this with the real source activity. Tools from pattern recognition may be promising for this purpose as well.

2 Statistical Inverse Problems

2.1 Basic Concepts of Bayesian Modeling

A native way to deal with high uncertainty and under-determinateness of a problem is to account for them explicitly by formulating the problem in a *statistical framework*. In our setting, the inverse problem is recasted in form of a statistical *quest for information*:

1. All variables are modeled as *random variables*.
2. This randomness is *not* a property of the objects to be recovered, but rather reflects our *lack of information* about their concrete values.
3. Every information available concerning their concrete values, (e.g., their mean value or the scale on which we expect them to be) is explicitly encoded in their *probability distributions*.
4. The solution of the inverse problem is a *posterior probability distribution* over the unknown variables.

The inference about statistical systems based on both a-priori information and measurements is subject to the *Bayesian approach* to statistics. Bayesian statistics are based on a subjective concept of probability, defining probability as the individual degree of belief in a statement, given the available evidence (Jaynes and Bretthorst, 2003)[1]. This concept is very suitable for the treatment of inverse problems (Kaipio and Somersalo, 2005). Bayesian statistics can be viewed as regularizing ill-posed statistical problems by incorporating a-priori information, but on the level of probability distributions rather than on the level of point estimates.
In the following, we will formulate the inverse problem of EEG/MEG source-reconstruction in a statistical framework.

Notation and Remark: *Subsequently, all random variables are denoted by italicized, upper case letters (e.g., X), their corresponding concrete realizations by italicized, lower case letters (e.g., $X = x$). Their probability distributions are denoted by $P(X)$ and their probability density functions by $p(x)$. In contrast to random variables, linear operators are denoted by unitalicized, upper case letters (e.g.,* L*). It should further be noted, that the setting we will deal with for the formulation of the inverse problem does not involve any critical issues from the point of probability theory: All random variables defined in this thesis are finite dimensional random variables on $\mathbb{R}^n$ and have a probability distribution that is absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}^n$ or are a Dirac measure. All these measures are Radon measures and problems potentially arising with conditional probability densities are not relevant in this context, since a regular version of the conditional probability densities exists (*Ambrosio et al., 2008; Klenke, 2008*). We will speak of distributions and densities instead of probability distributions and probability densities in the following. The multivariate normal distribution will simply be termed "Gaussian distribution", and random variables distributed according to this distribution will be termed "Gaussian random variables". Note that the following sections intend to give a gentle introduction of the main concepts rather than a solid mathematical formulation in the sense of probability theory. For this sake, the terms probability, probability distributions and*

[1] In contrast, the more common approach to statistics is called *frequentist statistics*, and is based on an objective concept of probability, defining probability as the limit of relative frequency in a large number of similar trials.

probability density are used somewhat loosely, as often encountered in applied statistics when it is clear that only absolutely continuous random variables are considered.

2.2 Bayesian Formulation of the Inverse Problem of EEG/MEG

Starting of with (1.5) we account for the measurement-noise generation explicitly by adding a random variable $\mathcal{E}$:

$$B = \mathrm{L}\, s + \mathcal{E} \tag{2.1}$$

Note that thereby, B has become a random variable, too. It is a good approximation to assume that $\mathcal{E}$ follows a Gaussian distribution with zero mean and a covariance matrix Σ_ε (cf. 1.3.2), which can be estimated in a data-preprocessing step. We further assume that a decorrelation of the measurement-channels is performed, which diagonalizes the covariance matrix such that $\Sigma_\varepsilon = \sigma^2 \mathrm{Id}_m$. For a given s, (2.1) then states that the conditional density $p(b|s)$ of B is given by:

$$p_{like}(b|s) = \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{m}{2}} \exp\left(-\frac{1}{2\,\sigma^2}\|b - \mathrm{L}\, s\|_2^2\right) \tag{2.2}$$

The density $p_{like}(b|s)$ is called *likelihood function* or short *likelihood* of the measurements b, as it encodes the probability that s generates b. This first step is a simple reformulation of the inverse problem that does not change any of its properties and does not provide any improvement. A classical statistical inference approach to estimate s, given b is the *maximum likelihood estimation*, which tries to find the value of s that maximizes (2.2). Apparently, this would lead to (1.5) again.

The next step is the central one in the Bayesian approach: s is considered to be a random variable itself (in our notation: $s \to S$). Its density $p_{prior}(s)$ reflects our a-priori assumptions and knowledge on its typical values. Hence, $p_{prior}(s)$ is called *a-priori density* (and the corresponding probability distribution is called *a-priori distribution*) or short *prior*. The choice of the likelihood and the prior fully determine and specify the model that is used for inversion. However, as the likelihood is normally determined by the process of data acquisition, the proper choice of $p_{prior}(s)$ is the most important part. It determines the different approaches and methods of model-inversion that are applicable. Therefore, most of the following sections are dedicated to the discussion how to choose and construct priors. Note nevertheless that it is the task of the prior to render the estimation problem well-posed. Thus it has to confine the source space sufficiently, which means that it has to distribute the probability tightly over the source space (one speaks of *informative* priors, as it should have a low entropy)

Given the likelihood and the prior, via the definition of conditional densities (Klenke, 2008), a quality called *model-evidence* can be computed:

$$p(b) = \int p(b,s)ds = \int p_{like}(b|s)p_{prior}(s)ds$$

The model evidence encodes the probability that the formulated model could generate the observed measurement b. It can be used to perform *model averaging* or *model selection*, i.e., different possible models are considered, and the one producing the largest model evidence is chosen. This offers a new level of inference, used in a wide range of applications, especially for the processing of real data rather than simulated data, where the underlying generating model is known. For EEG/MEG see Sato et al. (2004); Henson et al. (2009a) for the choice of the source space and the forward model on the basis of model-evidence, and, e.g., Henson et al. (2009b, 2010) for the validation of the benefits of multimodal integration by a comparison of the model-evidence.

Once a likelihood and a prior have been specified, Bayes rule can be applied to invert the model:

$$p_{post}(s|b) = \frac{p_{like}(b|s)p_{prior}(s)}{p(b)} \tag{2.3}$$

This is the conditional density of S given B. It is called *posterior density/distribution* or short *posterior* as it represents all our information on the unknown parameter S given the realization of $B = b$ by merging our knowledge *before* the measurement (the prior) with the information gained *after* performing the measurement (the likelihood). In Bayesian inference this density is the complete solution to the inverse problem.
The result being a probability density rather than a point estimate allows for various ways of inference. Furthermore, it offers the opportunity to quantify their reliability, e.g., the resolution power of the data is reflected by the width of the posterior. In cases where the posterior shows to be multimodal or suffers from extreme skewness, point estimates can become rather useless or uninformative. Nevertheless, the common way to exploit the information contained in the posterior is to infer a point estimate for the value of S out of it. There are two popular choices:

1. *Maximum a-posteriori*-estimate (MAP): $\hat{s}_{\text{MAP}} := \operatorname{argmax}_{s\in\mathbb{R}^n} \; p_{post}(s|b)$. Practically, this is a high-dimensional *optimization* problem.

2. *Conditional mean*-estimate (CM): $\hat{s}_{\text{CM}} = \mathbb{E}\left[s|b\right] = \int_{\mathbb{R}^n} s\; p_{post}(s|b)ds$. Practically, this is a high-dimensional *integration* problem.

At first glance, these two estimates do not seem to differ that much and for many distributions, the value they aim to estimate even coincides. Still, in Section A.1.3 in the appendix, we briefly introduce the theoretical framework of statistical estimation theory which reveals that the estimates rely on quite contradictory approaches. A first hint to that is the different nature of the practical tasks to obtain them. There are more sophisticated estimation methods that need the introduction of multiple classes of parameters, and then mix the above estimation methods. These methods will be discussed after the general introduction of this kind of modeling. For certain constructions of the prior, a direct link to the methods outlined in Section 1.3.2 can be established. As this will clarify the role of the distributions introduced here, we will sketch it in the next section.

2.3 Reformulation of Tikhonov-type Regularization Methods

We start with the most commonly used a-priori assumption, i.e., we choose a Gaussian distribution with zero mean and known covariance matrix Σ_s as a prior on S. Using Bayes rule (2.3) and the definition of the likelihood (2.2), the posterior becomes:

$$\begin{aligned} p_{post}(s|b) &= p(b)^{-1}(2\pi)^{-m/2}|\Sigma_\varepsilon|^{-1/2}\exp\left(-\frac{1}{2\,\sigma^2}\|b-\mathrm{L}\,s\|_2^2\right)\cdot(2\pi)^{-n/2}|\Sigma_s|^{-1/2}\;\exp\left(-\frac{1}{2}s^t\Sigma_s^{-1}s\right) \\ &\propto \exp\left(-\frac{1}{2}\left(\|\Sigma_\varepsilon^{-1/2}(b-\mathrm{L}\,s)\|_2^2+\|\Sigma_s^{-1/2}s\|_2^2\right)\right) \end{aligned} \tag{2.4}$$

Computing the MAP estimate of this posterior directly gives the WMNE from Section 1.3.2 as formulated by Dale and Sereno (1993), or in the alternative formulation:

$$\begin{aligned} \hat{s}_{\text{MAP}} &= \operatorname*{argmax}_{s\in\mathbb{R}^n}\left\{\exp\left(-\frac{1}{2}\left(\|\Sigma_\varepsilon^{-1/2}(b-\mathrm{L}\,s)\|_2^2+\|\Sigma_s^{-1/2}s\|_2^2\right)\right)\right\} \\ &= \operatorname*{argmin}_{s\in\mathbb{R}^n}\left\{\|\Sigma_\varepsilon^{-1/2}(b-\mathrm{L}\,s)\|_2^2+\|\Sigma_s^{-1/2}s\|_2^2\right\} \\ &\overset{1.3.2}{=} \operatorname*{argmin}_{s\in\mathbb{R}^n}\left\{\|b-\mathrm{L}\,s\|_2^2+\lambda\|\mathrm{W}s\|_2^2\right\}, \quad \text{where} \quad \mathrm{W}=\tfrac{\sigma}{\sqrt{\lambda}}\Sigma_s^{-1/2} \\ &= \hat{s}_{\text{WMNE}} \end{aligned}$$

In essence, the exchange of perspectives from the indirect modeling via classical Tikhonov regularization to the explicit one-level Gaussian modeling in the Bayesian framework works over

a direct reciprocal relationship of $\mathrm{W}^t\,\mathrm{W}$ and Σ_s or respectively of W and $\Sigma_s^{1/2}$. This is a well known equivalence for Gaussian priors. Some further explanations on it are given in Section A.1.5 in the appendix. More generally, Tikhonov-type regularization with an arbitrary penalty functional $\mathcal{P}(s)$ corresponds to MAP-estimation with a prior $p_{prior}(s) \propto \exp\{-\frac{1}{2}\,\lambda\,\mathcal{P}(s)\}$, i.e., the penalty functional defines the *energy*[2] of the prior, which is its negative natural logarithm:

$$\begin{aligned}\hat{s}_{\mathrm{MAP}} &= \underset{s\in\mathbb{R}^n}{\operatorname{argmax}} \left\{ \exp\left(-\frac{1}{2\,\sigma^2}\|b-\mathrm{L}\,s\|_2^2\right) \exp\left(-\frac{1}{2}\,\lambda\,\mathcal{P}(s)\right)\right\} \\ &= \underset{s\in\mathbb{R}^n}{\operatorname{argmin}} \left\{ \frac{1}{\sigma^2}\|b-\mathrm{L}\,s\|_2^2 + \lambda\mathcal{P}(s)\right\}\end{aligned}$$

Since many penalty functionals can be seen as abstractions of energies arising in real physical systems, this equivalence is far more than a pure technical transformation.
While the above equivalences are quite well known, it is less known that "adaptive" regularization methods as well as practical optimization algorithms for certain penalty functionals (including the often used EM-algorithm) often correspond to algorithms to compute MAP or empirical Bayesian (γ-MAP) estimates in hierarchical Bayesian models, which we will introduce in the next section. Details on these correspondences will be revisited after introducing a special hierarchical model in Section 4.2.

2.4 Hierarchical Models

2.4.1 Motivation

Brain activity as the general subject to be modeled is a very complex process that comprises many different spatial patterns. As mentioned in Section 2.2, a prior modeling all of these phenomena would have to distribute probability of the same order of magnitude on a wide range of points in the source space. This would lead to a *flat* or *uninformative* prior that is not able to render the estimation problem well-posed anymore. This problem can be handled by introducing an adaptive, data-driven element into the estimation process. The mathematical consolidation of this approach leads to a *hierarchical* multi-level model structure, which we will introduce and discuss in the remaining part of this chapter.

2.4.2 General Construction and Point Estimates

The idea of *hierarchical Bayesian models* (*HBM*) is to let the same data determine the appropriate model used for the inversion of this data. At first glance, this looks like an obvious overestimation of the information given by the data: We use an estimator constructed by our data to invert our data, therefore our estimator has to be biased in some way. This methodical conflict can be solved by extending our model by a new dimension of inference: Not only S, but also the prior on S itself is not fixed anymore but random, determined by the values of additional parameters $\boldsymbol{\gamma}$, called *hyperparameters*. These parameters follow an a-priori assumed distribution (the so called *hyperprior*) and are subject to estimation schemes, too. This construction follows a top-down scheme, as the parameters of each level completely control the distribution of the parameters of the level below, thus this modeling approach is called *hierarchical* modeling:

$$p(s,\boldsymbol{\gamma}) = p_{prior}(s|\boldsymbol{\gamma})\,p_{hyper}(\boldsymbol{\gamma}) \quad \Rightarrow \quad p_{prior}(s) = \int p_{prior}(s|\boldsymbol{\gamma})\,p_{hyper}(\boldsymbol{\gamma})d\boldsymbol{\gamma} \tag{2.5}$$

The likelihood does not depend on $\boldsymbol{\gamma}$:

$$p_{like}(b|s,\boldsymbol{\gamma}) = p_{like}(b|s)$$

[2] In statistical physics, the probability of a state i is given by the *Boltzmann distribution* (*Gibbs measure*), i.e., $p_i \sim \exp(-\beta\,E_i)$, where E_i is the energy of the state, and β is the inverse of the *fundamental temperature* $k_b T$.

The posterior takes the form:

$$p_{post}(s, \boldsymbol{\gamma}|b) \propto p_{like}(b|s)\, p_{prior}(s|\boldsymbol{\gamma})\, p_{hyper}(\boldsymbol{\gamma}) \tag{2.6}$$

Using a large number of hyperparameters, this construction allows for the construction of both very complex and flexible models and the incorporation of qualitative rather than quantitative information on the level of the hyperparameters. The estimation process leads to a convenient mixing of these different types of information, and data-driven hyperparameter estimation can automatically determine relevant model-components. On the theoretical side, hierarchical models are central to modern Bayesian statistics, for allowing a more "objective" approach to inference by estimating the parameters of the prior distributions from data rather than requiring them to be specified using subjective information. See MacKay (2003) and Gelman et al. (2003) for a general reference. In the next section, a specific realization of such a model and the mentioned properties will be discussed in detail. In the following, we will introduce the main inference methods this construction scheme offers at this abstract stage: Note that the posterior (2.6) now depends on two kinds of parameters, the ones of main interest s, and the hyperparameters $\boldsymbol{\gamma}$. Five main ways to deal with this situation are established:

Full-CM: Integrate $p_{post}(s, \boldsymbol{\gamma}|b)$ w.r.t. s and $\boldsymbol{\gamma}$.
Full-MAP: Maximize $p_{post}(s, \boldsymbol{\gamma}|b)$ w.r.t. s and $\boldsymbol{\gamma}$.
S-MAP: Integrate $p_{post}(s, \boldsymbol{\gamma}|b)$ w.r.t. $\boldsymbol{\gamma}$, and maximize over s. (*Type I approach*)
$\boldsymbol{\gamma}$-MAP: Integrate $p_{post}(s, \boldsymbol{\gamma}|b)$ w.r.t. s, and maximize over $\boldsymbol{\gamma}$, first.
Then use $p_{post}(s, \hat{\boldsymbol{\gamma}}(b)|b)$ to infer s (*Type II approach, Hyperparameter MAP, Empirical Bayes*).
VB: Assume approximative factorization of $p_{post}(s, \gamma|b) \approx \hat{p}_{post}(s|b)\, \hat{p}_{post}(\gamma|b)$.
Approximate both with distributions that are analytically tractable
(VB = *Variational Bayes*).

In the traditional Bayesian framework, all kinds of parameters should be treated equally, that is why the first two schemes are also referred to as *fully-Bayesian* methods. Still, practically, the hyperparameters have been introduced with the explicit intention that they have a different meaning than the normal parameters, hence a different treatment can be justified from the methodical point of view. The corresponding schemes, S-MAP and $\boldsymbol{\gamma}$-MAP, are usually classified as *semi-Bayesian* methods. Variational Bayesian techniques (often referred to as *approximate-Bayesian* methods) actually rely on more advanced considerations than a simple approximation, but this cannot be pursued in detail here (Friston et al., 2007; Nummenmaa et al., 2007a; Wipf and Nagarajan, 2009). We will focus on fully-Bayesian methods in this thesis.

2.4.3 Gaussian Scale Mixture Models

Construction: The special construction of the prior on s that is proposed, is called a *Gaussian scale mixture* or *conditionally Gaussian hypermodel*, which means that the density of s is controlled by hyperparameters $\boldsymbol{\gamma}$, and for every fixed value of $\boldsymbol{\gamma}$, it is a Gaussian density as in Section 2.3. Consequently, varying the hyperparameters can, e.g., vary the typical scale of the distribution (the standard deviation in the one-dimensional case). See, e.g., Dempster et al. (1977); Palmer et al. (2006) for a general introduction and treatment of such models. Note that this approach is different from a *mixture of Gaussians* which refers to the weighted sum of Gaussian densities (often intended as an approximation to the real density): The resulting density must not be a Gaussian density itself (it hardly ever is).
We compose the total covariance matrix as a weighted sum of covariance components C_i belonging to a predefined set $\mathcal{C} \subset \mathbb{R}^{n \times n}$ of symmetric, positive, semi-definite matrices. The weighting

between them is controlled by a (positive) hyperparameter $\boldsymbol{\gamma} \in \mathbb{R}^h$:

$$S|\boldsymbol{\gamma} \sim \mathcal{N}(0, \Sigma_s(\boldsymbol{\gamma})) \qquad \text{with} \qquad \Sigma_s(\boldsymbol{\gamma}) = \sum_{i=1}^{h} \gamma_i \mathrm{C}_i \qquad \text{where} \qquad \mathrm{C}_i \in \mathcal{C}$$
$$\Rightarrow\ p_{prior}(s|\boldsymbol{\gamma}) = (2\pi)^{-n/2} |\Sigma_s|^{-1/2} \exp\left(-\frac{1}{2}\|s\|^2_{\Sigma_s^{-1}}\right),$$

where $|\Gamma|$ denotes the determinant of a matrix. For the distribution of the hyperparameter $\boldsymbol{\gamma}$, a hyperprior that factorizes and takes a special form is assumed:

$$p_{hyper}(\boldsymbol{\gamma}) \propto \prod_{i=1}^{h} \exp\left(-\frac{1}{2} f_i(\gamma_i)\right) \quad \text{and} \quad \gamma_i > 0 \tag{2.7}$$

The f_i are assumed to be fixed and known at this point and will be specified later. The posterior then takes the form:

$$p_{post}(s, \boldsymbol{\gamma}|b) \overset{(2.6)}{\propto} \exp\left(-\frac{1}{2}\left(\frac{1}{\sigma^2}\|b - \mathrm{L}\, s\|_2^2 + \|s\|^2_{\Sigma_s^{-1}} + \ln|\Sigma_s| + \sum_{i=1}^{h} f_i(\gamma_i)\right)\right) \tag{2.8}$$

Note that the computation of the implicit prior on s involves the integration over all hyperparameters which (in the case of normalized C_i) represent the value of the variances and therefore the length scale on which s takes its values. This explains the term "scale mixture". We could easily extend the presented framework by adding more levels, but we will focus on the impact of this one new level of the hyperprior[3]. Note that this is the main extension compared to the WMNE scheme (cf. 1.3.2), which is (virtually) included in this framework, if we choose just one variance component, namely $\mathrm{C} := \left(\mathrm{W}^t\,\mathrm{W}\right)^{-1}$ and equip $\boldsymbol{\gamma}$ with a singular hyperprior $p_{hyper}(\boldsymbol{\gamma}) = \delta(\boldsymbol{\gamma} - \frac{\sigma^2}{\lambda})$.
The analytical advantage of such a model over other possible approaches is that the expression within the brackets in (2.8) is quadratic with respect to s and the γ_i's are mutually independent. This allows for a reformulation of the model in terms of *pseudo sources* which simplifies and accelerates many practical computations with this model (see the last but one paragraph of this section on page 21 and Chapter 3).

Choice of covariance components and hyperpriors: The choice of a specific set of covariance components and the corresponding hyperpriors fully specifies the prior used, thus it is the important choice in this framework. As the covariance matrix expresses the likelihood of single and simultaneous activity of sources, the choice can be motivated by neurological as well as mathematical considerations. When encoding a-priori information into a covariance component, one has to bear in mind that we usually formulate our a-priori knowledge in terms of the scalar valued, positive source activity, and that source activity is the main information users are interested in. However, our framework is formulated in terms of the vector valued currents. The transfer of a-priori information about source activity correlation into the vector valued framework needs some care, as we usually do not intend to enter any correlation about the orientation of the sources as well: Just assigning a positive correlation to all d basis functions of both source locations chains the source activity, but also chains the orientations of the basis functions. For similar reasons, our model is not able to model inhibition, i.e., activity in one location attenuates activity in another location: Assigning a negative correlation to two (oriented) basic sources just means that they show the same amplitude of activity, but in opposed directions. An extension of our model to deal with this issue is proposed in the outlook in Section 6. To ease the

[3] An extension by one additional level was examined in Nummenmaa et al. (2007a), and did not show convincing results. The authors concluded that the measurements do not contain enough information about this stage.

notation in the following, we will distinguish between *activity covariance components* (*ACC*) $\mathrm{C}^a \in \mathbb{R}^{k\times k}$ encoding the covariance structure of the source activity at different locations and *current covariance components* (*CCC*) $\mathrm{C}^c \in \mathbb{R}^{dk\times dk}$ encoding the covariance structure between all basis functions. For the reasons presented above, we will restrict ourselves to CCCs of the form $\mathrm{C}^c = \mathrm{C}^a \otimes \mathrm{Id}_d$. The following components have been used or proposed so far:

- ⋆ In the most simple case, $\mathrm{C}^a = \mathrm{Id}_k$ is used.
- ⋆ *Spatial smoothness* can be enforced by adding ACCs that impose a covariance structure based on an appropriate distance, e.g., in the form $(\mathrm{C}^a)_{i,j} = \exp(-\operatorname{dist}(r_i, r_j))$. Components of that kind are used, e.g., in Phillips et al. (2005) and Mattout et al. (2006).
- ⋆ ACCs called *location priors* are diagonal matrices, only consisting of ones and zeros, hence they promote source-activity only in some locations. The extreme case is given by ACCs of the form $(\mathrm{C}^a_q)_{(i,j)} := (\boldsymbol{e}_q \boldsymbol{e}_q^t)_{(i,j)} = \delta_{r_i,r_q} \cdot \delta_{r_j,r_q}$, which promote source-activity only in one location. If they are used in combination with hyperpriors enforcing the sparsity of $\boldsymbol{\gamma}$, their choice leads to sparse source-configurations (see 4.2). Apart from location priors that can be incorporated without any given information on the location of the source-activity, location priors encoding some a-priori information are often used. This information can be derived from other imaging modalities giving anatomical or functional information like MRI, *functional magnetic resonance tomography* (*fMRI*) or *positron emission tomography* (*PET*), invasive devices like *electrocorticography* (*ECoG*) or it can be chosen manually by the clinician. Even so, the blindfolded use of location priors is dangerous, if the information is invalid (Wipf and Nagarajan, 2009). One should either combine them with covariance components compensating for that, or include this information on another level, as explained in the next paragraph on page 20.
- ⋆ $\mathrm{C}^c := \operatorname{diag}(\mathrm{L}^t\mathrm{L})^{-1}$ is often used to compensate for *depth-bias* (cf. 4.1). The choice of this CCC (which corresponds to certain WMNE schemes (e.g., Fuchs et al., 1999) is crucial from the Bayesian point of view: It encodes the a-priori information that deep-lying sources show larger activity than superficial ones (on average). This cannot be justified from any neurophysiological findings.
- ⋆ $\mathrm{C}^a_i := \psi_i \psi_i^t$ with $\psi_i \in \mathbb{R}^k$ is a general construction often used in signal processing, leading to harmonic analysis. ψ_i, $i = 1, \ldots, l$ are a set of basic (spatial) waves, e.g.:
  - ∗ *Resolution kernels*: $\psi_{k,v}$ is centered on the source locations k with scale v, therefore $l = n \cdot v$, e.g., $\psi_{k,v}$ is a Gaussian centered on $\mu = k$ with variance $\sigma = \sigma_0 \cdot 10^v$. This leads to a multi-resolution decomposition learned from the data, often using sparsity-enforcing hyperpriors and methods (Wipf and Nagarajan, 2009).
  - ∗ *Multiple sparse priors* (Friston et al., 2008) is a technique to derive these basic waves out of a spatial coherence matrix.
  - ∗ Learned empirically from the data, see, e.g., Phillips et al. (2002a); Mattout et al. (2005).

In practice, instead of using the full CCC or ACC, often only the diagonal part of it is used (referred to as *variance component*). This is not only due to computational considerations, but also showed better results in some studies (e.g., Henson et al., 2010). In addition, the problems concerning the unwanted chaining of orientations can also be avoided by just using variance components: These components just enlarge the single variances of both locations relative to other locations. The final amplitude and especially the orientation will be determined by the measurements. This does not ensure simultaneous activity but remind that once the hyperparameters are fixed, the model collapses to a WMNE. WMNEs usually incorporate every basis function with a significant variance that can account for an aspect of the measurements.

Consequently, both locations will usually show significant activity if they contribute to the data in some way and their variances are larger than those of other locations.
Note that concerning the covariance components, the framework is constructed in a modular way, as one can easily include or exclude certain covariance components, and test how combinations of them work. Facing this variety of possible covariance components, the central question is, how to choose a set $\mathcal{C}$? The answer to this question relies heavily on the hyperpriors chosen, and the inversion method used. With each covariance component, a hyperparameter γ_i is associated that controls the relevance of the component. Usually, a kind of competition between the hyperparameters, leading to the pruning of most of them, and as a result to a sparse $\boldsymbol{\gamma}$, is desired. This may also be needed, as discussed in the last paragraph of this section on page 23.
The next important choice are the hyperpriors. Hyperpriors have a different function for the whole inversion process than priors. As pointed out in Section 2.2, the prior has to regularize the estimation problem, thus it has to reduce the dynamical range of the parameters sufficiently. The hyperprior just controls the parameters defining the prior, hence it has to confirm that if it gives moderate probability to a set of these parameters, the resulting prior can fulfill its function. One important property is that it has to prohibit *overfitting*, i.e., the solution for s is found in a $\boldsymbol{\gamma}$-region that corresponds to models that are complex with regard to s. This will be explained in a more general setting in the last paragraph of this section on page 23.
For some densities, these demands allow the hyperprior to be nearly arbitrary. This concept is often used for Gaussian priors, e.g., hyperpriors are used that are *non-informative* (i.e., *scale-invariant* for a *scale parameter* like $\boldsymbol{\gamma}$, see MacKay, 2003), flat and not even normalizable. This just means that the hyperprior allows that the values of the hyperparameters are estimated solely data-driven (the hyperprior encodes our a-priori information on the typical values of $\boldsymbol{\gamma}$, if we don't have any information at hand, we completely rely on our data). However, the use of these *improper* (not normalizable) hyperpriors needs some care: It is not always clear, whether the resulting posterior is proper: For the non-informative hyperprior, i.e., $p_{hyper}(\gamma_i) \sim \gamma_i^{-1}$ the resulting posterior is in fact improper (Gelman, 2006; Nummenmaa et al., 2007a), whereas a flat hyperprior, i.e., $p_{hyper}(\gamma_i) \sim 1$ results in a proper posterior. To prevent this, the concept of *weakly-informative* hyperpriors is used, i.e., a distribution is chosen that is proper, but intentionally provides weaker information than any a-priori knowledge that is available (Gelman, 2006).
Further common choices include the gamma and inverse-gamma distribution (e.g., Sato et al., 2004; Nummenmaa et al., 2007a; Calvetti et al., 2009, see 4.2) and the log-normal distribution (Friston et al., 2008). These distributions lead to a *sparse* $\boldsymbol{\gamma}$. The pruning process leading to this result is usually regarded as an automated, data-driven learning of the relevant model features, termed *automatic relevance determination* (*ARD*, see MacKay, 1991; Neal, 1994). More information on the whole topic can be found in Gelman (2006).

A-priori information embedding: As indicated in the last paragraph, a-priori information can be incorporated at different levels of the hierarchical structure of the model. In general, *qualitative* information is more suitable for higher levels, whereas *quantitative* information can be incorporated in the lower stages. The higher levels of the model determine the general behavior, whereas the lower levels need to regularize the statistical problem, thus they need to constrain the effective range of the main parameters sufficiently once the higher levels are set up. A crucial point in this context is the validity of the a-priori information. If invalid a-priori information is included in the lowest levels of the model, it can hardly be corrected by the data, and usually leads to invalid results (see, e.g., Liu et al., 2006b for the impact of invalid fMRI location priors). One should therefore consider how "hard" or "soft" the a-priori information given is, and include it at the right stage. For our concrete model, this can be done in the following ways:

- ⋆ Structural information with high reliability, e.g., given by CT or MRI-scans are usually used to confine the source space beforehand, and are further used in the forward compu-

tation. They are included into the model implicitly over the lead-field matrix (Dale and Sereno, 1993).

- ⋆ Structural information with lower reliability can be included over covariance components as sketched above, but should only be used in combination with other covariance components that can correct their influence if necessary (Wipf and Nagarajan, 2009).
- ⋆ Structural information from *standard probability maps* given by computational brain atlases, has to be registered to the individual source space first, its validity is crucial, especially in cases, where the anatomy of the patients head shows anomalies (Trujillo-Barreto et al., 2004).
- ⋆ Information learned empirically from the same or comparable data can be included over covariance components (Phillips et al., 2002a; Mattout et al., 2005).
- ⋆ Functional information with low spatio or temporal resolution, e.g., given by fMRI or PET-scans can either be included as covariance components, but as above, only in combination with other components, but it was suggested that this information should rather be included in the hyperprior (Sato et al., 2004). This holds for functional information given by probability maps in general (fMRI information is usually given in this format), especially concerning information given by computational brain atlases (Wipf and Nagarajan, 2009).

Generalized transformation to standard form: In the following, we will formulate the scale mixture model in terms of *generalized* or *pseudo* sources. This will ease the formulation and implementation of many algorithms and clarifies the rule of the covariance components C_i. In principle, an artificial generative source model is created that has a larger dimension, but a simpler structure and is equivalent to our original model in terms of the processing of external and a-priori information (i.e., completely equivalent in the sense of Bayesian statistics, but somewhat less natural in the description). This concept generalizes the transformation to standard form for weighted Tikhonov regularization (e.g., Engl et al., 1996) which is used to improve the condition of the underlying problem by a change of variables, and corresponds to a whitening transform of the variables of interest in the Bayesian framework (cf. A.1.5 for the details). For Gaussian scale mixture models, it can be extended and has a figurative meaning: We will shift the perspective from the amplitudes of the single basis functions $S_i,\ i = 1,\dots,n$ to clusters (termed pseudo sources) of (potentially coupled) basis functions determined by C_i and sharing the same variance $\gamma_i,\ i = 1,\dots,h$. Since the γ_i are independent, where the S_i are not, this perspective effectively decouples the problem but normally, the transformation from source to pseudo sources is not unique. Still, the inverse transformation from pseudo sources to sources is unique, thus if the algorithms are implemented based on the description of the pseudo source level, the desired result for s can be achieved afterwards. The basic technique behind this approach is an *affine mixing* of independent Gaussian random variables (see Section A.1.4 for details on that): Let $g := \sum_{i=1}^{h} \varrho_i$, where ϱ_i is the rank of C_i. We will then generate the n-dimensional random variable S with covariance $\sum_{i=1}^{h} \gamma_i \mathrm{C}_i$ from an affine mixing of g independent one-dimensional Gaussian distributed random variables:
Let $\tilde{S}_i$ be a ϱ_i-dim. Gaussian random variable with zero mean and a diagonal covariance matrix $\tilde{\mathrm{C}}_i := \gamma_i \mathrm{Id}_{\varrho_i}$ (i.e., the components of $\tilde{S}_i$ are independent):

$$\tilde{S}_i \sim \mathcal{N}_{\varrho_i}(0, \gamma_i \mathrm{Id}_{\varrho_i}), \quad \Rightarrow \quad p(\tilde{s}_i|\gamma_i) \propto \exp\left(-\frac{\tilde{s}_i^t \tilde{s}_i}{2\gamma_i}\right), \qquad \forall i = 1,\dots,h \tag{2.9}$$

Now let the $\tilde{S}_i$ be independent for all $i = 1,\dots,h$ and let $\tilde{S} := [\tilde{S}_1^t, \cdots, \tilde{S}_h^t]^t$. It follows that $\tilde{S} \in \mathbb{R}^g$ and (2.9) states that:

$$\tilde{p}_{prior}(\tilde{s}|\boldsymbol{\gamma}) = \prod_{i=1}^{h} \mathcal{N}_{\varrho_i}(\tilde{s}_i, 0, \gamma_i \mathrm{Id}_{\varrho_i}) = \mathcal{N}_g(\tilde{s}, 0, \tilde{\Sigma}_{\tilde{s}}), \quad \text{with} \quad \tilde{\Sigma}_{\tilde{s}} = \mathrm{diag}(\gamma_1 \mathrm{Id}_{\varrho_1}, \dots, \gamma_h \mathrm{Id}_{\varrho_h})$$

Next we will attain S with the desired covariance from an affine mixing of the $\tilde{S}_i$. Let A_i be the Cholesky factor of C_i, i.e., $\mathrm{C}_i = \mathrm{A}_i\mathrm{A}_i^t$. Further let $\mathrm{A} = [\mathrm{A}_1, \cdots, \mathrm{A}_h]$, and define S by:

$$S := \sum_{i=1}^{h} \mathrm{A}_i \tilde{S}_i = \mathrm{A}\tilde{S}$$

It follows from the affine transformation lemma (A.1.4) that, given $\boldsymbol{\gamma}$, S follows a Gaussian distribution with zero mean and a covariance given by

$$\Sigma_s = \mathrm{A}\tilde{\Sigma}_{\tilde{s}}\mathrm{A}^t = \sum_{i=1}^{h} \gamma_i \mathrm{A}_i\mathrm{A}_i^t = \sum_{i=1}^{h} \gamma_i \mathrm{C}_i \tag{2.10}$$

Note that this means, that through this construction, C_i and γ_i have the same impact on S as before. To formulate our model completely in terms of the pseudo sources, we define $\tilde{\mathrm{L}} := [\tilde{\mathrm{L}}_1, \cdots, \tilde{\mathrm{L}}_h] := \mathrm{L} \cdot [\mathrm{A}_1, \cdots, \mathrm{A}_h] = \mathrm{L} \cdot \mathrm{A}$ as the lead-field of the pseudo sources. This way $\tilde{\mathrm{L}}\tilde{s} = \mathrm{L}\mathrm{A}\tilde{s} = \mathrm{L}s$, and the complete model is determined by:

$$\begin{aligned}
&\tilde{p}_{like}(b|\tilde{s}) \propto \exp\left(-\frac{1}{2\,\sigma^2}\|b - \tilde{\mathrm{L}}\,\tilde{s}\|_2^2\right) \\
&\tilde{p}_{prior}(\tilde{s}|\boldsymbol{\gamma}) \propto |\tilde{\Sigma}_{\tilde{s}}|^{-1/2} \exp\left(-\frac{1}{2}\|\tilde{s}\|^2_{\tilde{\Sigma}_{\tilde{s}}^{-1}}\right) = \exp\left(-\frac{1}{2}\sum_{i=1}^{h}\frac{\tilde{s}_i^t\tilde{s}_i}{\gamma_i} - \frac{1}{2}\sum_{i=1}^{h}\varrho_i \ln(\gamma_i)\right) \\
&p_{hyper}(\boldsymbol{\gamma}) \propto \prod_{i=1}^{h} \exp\left(-\frac{1}{2}f_i(\gamma_i)\right) \quad \text{and} \quad \gamma_i > 0 \\
&\tilde{p}_{post}(\tilde{s}, \boldsymbol{\gamma}|b) \propto \exp\left(-\frac{1}{2}\left(\frac{1}{\sigma^2}\|b - \tilde{\mathrm{L}}\,\tilde{s}\|_2^2 + \sum_{i=1}^{h}\frac{\tilde{s}_i^t\tilde{s}_i}{\gamma_i} + \varrho_i \ln(\gamma_i) + f_i(\gamma_i)\right)\right)
\end{aligned} \tag{2.11}$$

Note that in contrast to (2.8) the posterior in (2.11) completely factorizes over γ_i. To see, that this is an equivalent HBM to (2.8), it is more instructive to regard the following points, than to rely on an algebraic transformation:

1. Both models encode our a-priori information in the same way:
   - ⋆ Our information on $\boldsymbol{\gamma}$ is given by $p_{hyper}(\boldsymbol{\gamma})$ which is the same for both models.
   - ⋆ Our information on the elementary sources is encoded in

     $$p_{prior}(s) = \int p_{prior}(s|\boldsymbol{\gamma})\; p_{hyper}(\boldsymbol{\gamma}) d\boldsymbol{\gamma},$$

     which is also the same in both models, due to the previous point and (2.10).
2. The interaction of the a-priori information with the data B is the same in both models:
   - ⋆ Since $\tilde{\mathrm{L}}\tilde{S} = \mathrm{L}S$, it follows that $p_{like}(b|s) = \tilde{p}_{like}(b|\tilde{s})$ since $s = \mathrm{A}\tilde{s}$, therefore the interaction of S and B is the same in both models.
   - ⋆ $\boldsymbol{\gamma}$ only interacts with B over s respectively $\tilde{s}$. Since the interaction of s respectively $\tilde{s}$ with b is the same we can analyze the interaction of b and $\boldsymbol{\gamma}$ over s in both models. As noted above, the relationship of $\boldsymbol{\gamma}$ and $S = \mathrm{A}\tilde{S}$ is the same in both models, thus the interaction with B is also the same. To see this explicitly, we can consider $p(\boldsymbol{\gamma}|b)$ and $\tilde{p}(\boldsymbol{\gamma}|b)$ which describe this interaction in both models. Remind, that $p(\boldsymbol{\gamma}|b) \propto p(b|\boldsymbol{\gamma})p(\boldsymbol{\gamma})$, that $B = \mathrm{L}S + \mathcal{E}$ and that, given $\boldsymbol{\gamma}$, $S \sim \mathcal{N}(0, \Sigma_s)$. Therefore, $p(b|\boldsymbol{\gamma}) \sim \mathcal{N}(0, \Sigma_b)$, with $\Sigma_b := \Sigma_\varepsilon + \mathrm{L}\Sigma_s\mathrm{L}^t$ (see A.1.4) and

     $$p(\boldsymbol{\gamma}|b) \propto \exp\left(-\frac{1}{2}\left(b^t\Sigma_b^{-1}b + \ln|\Sigma_b| + \sum_{i=1}^{h} f_i(\gamma_i)\right)\right) \tag{2.12}$$

As a result, $p(\boldsymbol{\gamma}|b)$ relies on the source level of the model only through the *projected source covariance* $\mathrm{L}\Sigma_s\mathrm{L}^t$. Since $\mathrm{L}\Sigma_s\mathrm{L}^t = \tilde{\mathrm{L}}\tilde{\Sigma}_{\tilde{s}}\tilde{\mathrm{L}}^t$, it follows that $p(\boldsymbol{\gamma}|b) = \tilde{p}(\boldsymbol{\gamma}|b)$.

In summary, the pseudo source framework is an over-parameterization of the original generative model with the intention to reduce dependencies between the main parameters. This eases the implementation of many practical estimation algorithms and further speeds up their convergence (Gelman et al., 2007).

Overfitting, Degrees of Freedom and Entropy: A common objection against the use of the HBM presented here instead of simpler schemes like MNE is that the model comprises much more parameters. This introduction of a large number of additional parameters that need to be estimated seems counterproductive, since already in simpler models, the number of main parameters n is much lager than the number of measurements m. To sum up the arguments, an objection could be stated like this:

"The information given by the measurement is not sufficient to determine the free parameters within the simpler model, so why should we make it even more complex?"

A misconception of three related but somewhat different concepts is underlying this statement: *Information*, *degrees of freedom* and *model complexity*. A complete and sound treatment of these issues is beyond the scope of this thesis, and needs a detailed commitment to statistical estimation theory. In the following we will therefore sketch some considerations that will not fully resolve the misconception, but will point out, in which aspects one needs to take more care. We start with *model complexity*: The relation between the parameters of interest and the observables is usually considered to be a simple one, but corrupted by noise. A complex model of the parameters of interest within this situation might lead to *overfitting*: It will mainly describe noise instead of the underlying relationship - the complexity makes it *unspecific*. To illustrate the further arguments, we introduce a simple example of such a situation: Let y_i be measurements which originate from a noisy measurement of a simple linear relationship: $y_i = bx_i + a + \varepsilon$, where a and b are constants and ε an independent additive noise term. A simple, but robust model in this situation would be to assume a linear relationship $y_i = \theta x_i + \vartheta$ which can be fitted to the data via a least-squares-fit, hence involve some residual error. A more complex model would, e.g., be to assume a polynomial relationship of the order of the number of measurement points y_i. The parameters one has to estimate are the coefficients of the polynomial. This approach can explain the data without residual error, but it rather fitted the noise than the underlying relationship: It is neither robust against noise, nor has a good *predictive power* concerning the value of y in other points than x_i, e.g., when used to inter- or extrapolate the data. For these reasons, a main paradigm in statistical modeling is to always start with a simple model in the first place. People became aware of this danger for HBMs as well. Hierarchical models were constructed with the intention to contain various different concrete models for S in one superior model. This means that the model complexity with regard to both S and $\boldsymbol{\gamma}$ is very high as mentioned in the general introduction to hierarchical modeling above. Yet, this must not mean that for every fixed $\boldsymbol{\gamma}$ the model complexity with regard to S is still high: If this $\boldsymbol{\gamma}$ represents a very specific model on S it can actually be quite low (for the concrete model considered in Section 4.2, if $\boldsymbol{\gamma}$ is sparse, the model complexity for S is much lower than for the classical MNE). As a result, if values for $\boldsymbol{\gamma}$ leading to a complex model for S have low probability within the model, overfitting with regard to S will practically not occur. To illustrate that in our simple example from above, think of a HBM for this situation as a polynomial with coefficients as the main parameters of interest, and a similar number of hyperparameters, each controlling the probable range of one main parameter in a proportional way. If the corresponding hyperprior favors sparse hyperparameter configurations, the interplay with the likelihood will ensure that only the coefficients for the zeroth and the first monomial will be nonzero, thus effectively the simple model $y_i = \theta x_i + \vartheta$ is used for the inversion as well. However, the advantage over the

simple model is that if the data would come from a higher order polynomial with few nonzero coefficients, the HBM is likely to reconstruct this, too.
In summary, the important question is, given a certain prior depending on hyperparameters, how do we have to choose the hyperpriors to enforce simple models on S? If we had some measure of the model complexity as a function of $\boldsymbol{\gamma}$, say $c(\boldsymbol{\gamma})$, the choice $p(\boldsymbol{\gamma}) \propto \exp(-\beta c(\boldsymbol{\gamma}))$ would be tempting[4]. This approach would lead to:

$$p(\boldsymbol{\gamma}|b) \propto p(b|\boldsymbol{\gamma})p(\boldsymbol{\gamma}), \quad \text{with} \quad p(b|\boldsymbol{\gamma}) = \int p_{like}(b|s)p(s|\boldsymbol{\gamma})\, dS$$

and on the log scale:

$$\log p(\boldsymbol{\gamma}|b) \propto \log p(b|\boldsymbol{\gamma}) - \beta c(\boldsymbol{\gamma}) \tag{2.13}$$

The density $p(\boldsymbol{\gamma}|b)$, i.e., the marginal posterior of $\boldsymbol{\gamma}$, shows, how the model on S is adopted to b. Equations (2.13) tell us that this adoption is controlled by two opposing terms in a regularization-like scheme: The first is the marginal likelihood of b given $\boldsymbol{\gamma}$ and therefore represents the fit of the model to the data. When maximizing $p(\boldsymbol{\gamma}|b)$ this term will lead to a high model complexity in order to best fit the data. The second term tries to prohibit this increase of the model complexity, hence the whole equation has the form "goodness of fit - penalization of model complexity" (van der Linde, 2001). Coming from this abstract scheme, we will now examine how the situation looks like for the Gaussian scale mixture model (2.8): Here, each C_i corresponds to a specific model on S, thus a sparse $\boldsymbol{\gamma}$ should lead to a low model complexity. Computing (2.13) for the concrete posterior (2.8) leads to (cf. (2.12)):

$$\mathcal{L}_{\mathrm{II}}(\boldsymbol{\gamma}) := -2\log p(\boldsymbol{\gamma}|b) \propto \underbrace{b^t\Sigma_b^{-1}b}_{\substack{\text{data fit}}} + \underbrace{\log|\Sigma_b|}_{\substack{\text{volume-based}\\ \text{regularization}}} + \underbrace{\sum_{i=1}^{h} f_i(\gamma_i)}_{\substack{\text{hyperprior-based}\\ \text{regularization}}},$$

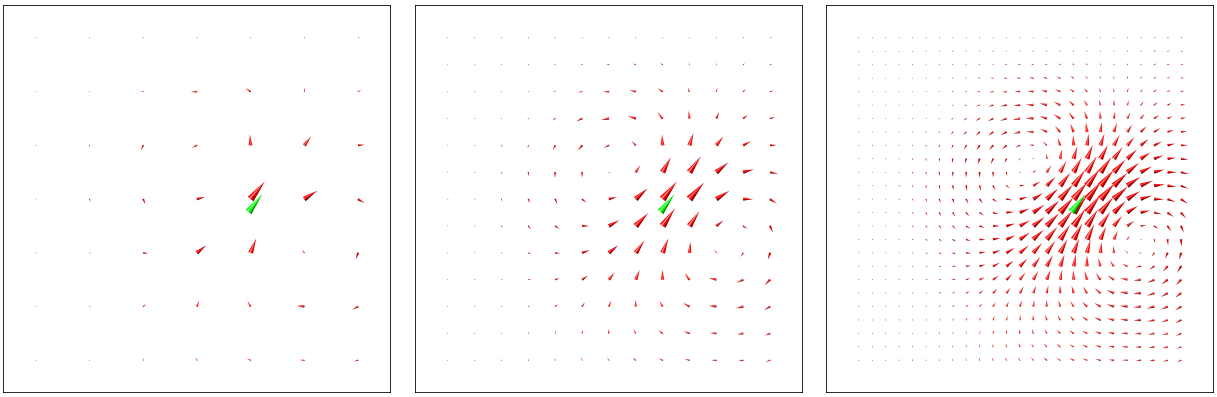
The cost function $\mathcal{L}_{\mathrm{II}}$ is called *Type II cost function*, its minima $\hat{\boldsymbol{\gamma}}$ correspond to probable models on S given b. Interestingly, an explicitly sparsity-enforcing hyperprior is not even needed to maintain a low model complexity in the Gaussian scale mixture model: The additional term $\log|\Sigma_b|$ measures the volume formed by the *total sensor covariance* Σ_b[5]. The volume of high dimensional objects is minimized most effectively by collapsing single dimensions as close to zero as possible due to the *curse of dimensionality* (opposed to an isometrically reduction of all dimensions). That means that Gaussian scale mixture models favor sparse $\boldsymbol{\gamma}$'s *intrinsically*. The hyperprior can correct, damp or amplify this property, but if this is not explicitly intended a Gaussian scale mixture model will hardly lead to an overfitting, regardless the large number of additional parameters it can comprise (Wipf and Nagarajan, 2009).
This leads to the second point, the *degrees of freedom*: In classical statistics, calculating the degrees of freedom is just a way to keep score on the remaining variability in the parameters, given the observed data or after estimating certain quantities (which is mathematically the same). For instance, assume we have observed/estimated the mean of a set of N independent points x_i. We might ask ourselves to which extend this determines the data, and how much variability is still in it. At first, we have N degrees of freedom, as much as independent measurements. The observation/estimation of the mean uses one degree of freedom, as the remaining variability of the data (i.e., the residuals after subtracting the mean) is constrained to lie in the linear

[4]This means that the energy of the hyperprior is given by the measure of the model complexity rather than some analogue of a physical energy (cf. Section 2.3) which clarifies that the hyperparameters are purely internal parameters of the statistical model.
[5]The determinant of a positive symmetric matrix is a volumetric measure

subspace determined by $\frac{1}{N}\sum(x_i - \bar{x}) = 0$ which has the dimension $N - 1$. That is the reason, why a subsequent *unbiased* estimation of the variance is given by $\frac{1}{N-1}\sum(x_i - \bar{x})^2$ instead of $\frac{1}{N}\sum(x_i - \bar{x})^2$. Such classical estimators always correspond to linear operators which project the residuals into a subspace in an orthogonal ℓ_2 sense (*least-squares estimators*). As a result, the remaining variability of the data always lies in a proper linear subspace, and the degrees of freedom the data is precisely the dimension of this subspace. This is the intuitive meaning of degrees of freedom underlying the statement from above. In this sense, it is clear that the data is not sufficient to determine the parameters of the model, since $n > m$. However, the estimators encountered in inverse problems are usually not based on a least-squares projection of the data, and so measuring the remaining degrees of freedom in terms of dimensionality is generally not useful for these procedures. A concept called *effective degrees of freedom* is used instead. In our setting (equation (1.5)) we start with m measurements. For classical Tikhonov regularization, the effective degrees of freedom of the parameters given this observed data b is (Wahba, 1990):

$$\tau = m - \sum_{i=1}^{m} \frac{\sigma_i^2}{\sigma_i^2 + \lambda}, \qquad \text{where } \sigma_i \text{ are the singular values of L} \tag{2.14}$$

The meaning of this expression can be explained by looking at the limits $\lambda \to 0/\infty$: In general, b gives m pieces of independent information. For $\lambda \to 0$ (i.e., the pseudo inverse, cf. 1.3.2) it follows $\tau = 0$, so the estimation "consumes" the whole information given by the data. If the data is contaminated with noise, this leads to unwanted results, as the information given by the noise is also used completely. The limit $\lambda \to \infty$ leads to $\tau = m$, which means, that the method actually uses none of the information provided by b, which is clear, since $s = 0$ for $\lambda \to \infty$ independent from b. For every λ in between, τ reflects, how the variability of b is projected to the variability of s. This variability expresses the number of independent "units" of information that the method can use to determine the values of the main parameters in some way. If it is less than the number of these main parameters (as in our case, since $\tau < m < n$) it is clear that not every main parameter is free to vary on the full range of its possible values independent from all other parameters, but dependencies between the values of the main parameters are not avoidable. For MNE, the spatial structure of these dependencies is obvious: Neighboring locations are forced to show similar activity, leading to a blurred solution, which is not able to retain the spatial resolution offered by the source grid. In Figure 2.1 the MNE for a single dipole is depicted using a different number of source space nodes: A grid of 4×4 MEG gradiometers is located in a plane parallel to a plane containing a grid of source nodes (see Calvetti et al., 2009 for details). For the different source spaces, a hierarchy of three source grids that emerge from



(a) k = 7×7, $\lambda = 0.12$, $\tau = 5.41$ (b) k = 13×13, $\lambda = 0.45$, $\tau = 5.48$ (c) k = 25×25, $\lambda = 1.76$, $\tau = 5.52$

Figure 2.1: MNE using different source grid resolutions. Larger Versions of these figures can be found in the appendix on page XV

each other by successive refinement is used. The regularization parameter λ has been chosen according to the discrepancy principle (assuming a noise level of 5%, see Section 4.4.1 and 4.4.3). The effective degrees of freedom have been calculated according to formula (2.14). In each figure, the cones for the MNE solution have been scaled individually. Note that the effective degrees of freedom of the model, τ, hardly change, which means, that the information consumed by the MNE is almost equal, regardless the number of k. This is also visible in the figures: Adding new source space nodes does not add new features to the solution. Actually, the current values in these points could have been interpolated from their neighbors as well, without changing the result in a significant way (we have done that for illustration, see Figure A.5 on page XIV in the appendix). This means that by increasing k beyond a certain value, no new "free" parameters are added to the model, but parameters which are chained to the other parameters so strong that their inclusion does not offer any new insights[6]. For sparse reconstruction methods the spatial structure of the dependency between the parameters is more subtle, it is an inhibitory structure: If a bit of the variability of the data is used to determine the value of one parameter, several other parameters are forced to be zero, thus none of the variability is spent for them.

In summary, even for simple linear inverse methods, the meaning of degrees of freedom is not trivial anymore. For nonlinear, like most of the methods used for HBMs, this will be even less intuitive (I am not even aware of a general definition). At least, the pure number of "free" parameters in the model is not the important point but the *uncertainty* concerning their value (i.e., how "free" they are really allowed to be).

Note that the reason for this dilemma relies in the fact that we first discretized the continuous inverse problem (cf. Section 1.3) and then formulated our methods in this discrete setting. There is no guarantee that this procedure leads to a consistency between different discretization levels. Especially in Bayesian inversion it is known that a consistent, discretization invariant formulation of a-priori information needs some care (Lassas and Siltanen, 2004; Kaipio and Somersalo, 2005; Lassas et al., 2009). Another approach is to stick to the continuous formulation of the inverse problem as long as possible and to formulate the inverse methods in the continuous setting as well. Discretization is only performed for the practical computation at the end. This approach is quite common in the regularization community (Engl et al., 1996). Concerning Bayesian inversion and especially hierarchical modeling, the theoretical foundation for this approach is far less developed until now (for a discussion, see e.g., Lassas et al., 2009; Helin and Lassas, 2009; Helin, 2010b,a). Yet, if it is possible to formulate a continuous stochastic model, the Bayesian approach offers interesting tools to infer, examine and enhance discrete models derived from it: The field of *statistical model reduction* is concerned with accounting for discretization errors explicitly by the use of *enhanced noise models* and inverse crimes in conjunction with discretization (Kaipio and Somersalo, 2005, 2007).

Finally, we address the last remaining point of this paragraph, i.e., *information*:

Statements like "The information given by EEG/MEG measurements is not sufficient to fully determine the source activity" are problematic: It aims at the under-determinateness of (1.4) or respectively (1.5), i.e., the continuous or discrete forward operators $\mathcal{L}^{em}$ or L are not injective. For $\mathcal{L}^{em}$, it is often argued that von Helmholtz already showed this fact (von Helmholtz, 1853). However, he showed that it is true assuming *arbitrary* current distributions j^{imp}. Current distributions originating from brain activity may not take arbitrary form but may rather show very characteristic features due to the properties of the underlying network dynamics. Whether $\mathcal{L}^{em}$ is still not injective on this subset $\mathcal{J}_{brain}$ is not known yet. For the discrete case, it is clear that L is not injective on $\mathbb{R}^n$. But the model of brain currents in Section 1.3 that led to (1.5) has to be considered a provisional one in lack of a more appropriate image of brain activity. It assumes that $\mathcal{J}$ is a linear vector space, which is probably false: The simplest fact to see this

[6] It is actually not the pure number of k which is important, but rather if all locations in the source space that have a characteristic gain vector are included. Adding new locations that have similar gain vectors to already included ones does not improve the solution.

is that the brain current amplitude is surely not unbounded. A more subtle reason is that the currents originate from highly non-linear processes which makes it unlikely that, for two currents j_1, j_2 that originate from this activity, $j_1 + j_2$ also represents a reasonable outcome of it. To summarize, if the equations are really under-determined on the appropriate domains cannot be stated so far.
Another aspect of the term information in this context is that one should be careful when using it in conjunction with the discretization of model expressed by n: The measurement and hence the information given by it is a-priori independent of our modeling of the situation and of our usage of the measurements. A source of confusion might be that a quantification of the information in terms of *Shannon entropy* is only possible with respect to a given model. Still, this only means that some models are more appropriate to describe the underlying reality and cam therefore take advantage of the measurements in a better way. Apparently, this topic also brings us back to the discussion of the discretization level from above. A continuous modeling of the whole situation might also be advantageous here.
We conclude here and postpone a more detailed and concrete discussion of the above topics to the work subsequent to this thesis. The intention of this rather abstract paragraph was solely to address overhasty judgments about hierarchical models in the context of EEG/MEG.

3 Algorithms and Implementation

This chapter outlines strategies to compute the two fully-Bayesian estimates, i.e., the Full-MAP and the Full-CM estimate. We will propose and discuss a class of algorithms that rely on *alternated conditional moves* through the pseudo source space. For this we will extend ideas from Nummenmaa et al. (2007a); Calvetti et al. (2009) who proposed specific members of this class of algorithms for the covariance set $\mathcal{C} = \{\boldsymbol{e}_i \boldsymbol{e}_i^t \otimes \mathrm{Id}_d, i = 1, \ldots, k\}$. Using this set, the real source vectors s_{i*} coincide with the pseudo sources $\tilde{s}_i$ ($A_i = \boldsymbol{e}_i \otimes \mathrm{Id}_d$ for all $i = 1, \ldots, k$, since $h = k$) and (2.8) already factorizes over γ_i. The following chapter gives a generalized description of these specific algorithms and extends them to handle arbitrary covariance sets by means of the pseudo source decomposition.

3.1 Motivation

None of the point estimation methods for HBM presented in Section 2.4.2 can be computed explicitly. Thus, numerical approximations of the points $(\tilde{s}_{\mathrm{CM}}, \boldsymbol{\gamma}_{\mathrm{CM}})$, $(\tilde{s}_{\mathrm{MAP}}, \boldsymbol{\gamma}_{\mathrm{MAP}}) \in \mathbb{R}^g \times \mathbb{R}^h$ have to be found[1]. Within this thesis we will rely on approximation schemes that step through $\mathbb{R}^g \times \mathbb{R}^h$ in a more or less directed fashion:

- ⋆ For approximating $(\tilde{s}_{\mathrm{MAP}}, \boldsymbol{\gamma}_{\mathrm{MAP}})$, we seek to find a sequence $(\tilde{s}_i, \boldsymbol{\gamma}_i)$, $i = 1, \ldots, M$ such that
$$\max_{i=1,\ldots,M} \{\tilde{p}_{post}(\tilde{s}_i, \boldsymbol{\gamma}_i | b)\} \quad \xrightarrow{M \to \infty} \quad \tilde{p}_{post}(\tilde{s}_{\mathrm{MAP}}, \boldsymbol{\gamma}_{\mathrm{MAP}} | b)$$
Another desirable property of the sequence would be that it actually converges to $(\tilde{s}_{\mathrm{MAP}}, \boldsymbol{\gamma}_{\mathrm{MAP}})$. In the following, we will examine sequences where the computation of the subsequent state $(\tilde{s}_{i+1}, \boldsymbol{\gamma}_{i+1})$ only relies on the current state $(\tilde{s}_i, \boldsymbol{\gamma}_i)$.

- ⋆ For approximating $(\tilde{s}_{\mathrm{CM}}, \boldsymbol{\gamma}_{\mathrm{CM}})$, one has to compute the integral
$$\int_{\mathbb{R}^g \times \mathbb{R}^h} (\tilde{s}, \boldsymbol{\gamma})\, \tilde{p}_{post}(\tilde{s}, \boldsymbol{\gamma} | b)\, d\tilde{s}\, d\boldsymbol{\gamma}$$
numerically. Due to the high dimensionality of the source space, this is intractable by means of traditional quadratures. Integration by *Monte Carlo* methods can avoid these difficulties, because the rate of convergence does, in principle, not depend on the dimension g. For our application, the best would be if one could find a sequence $(\tilde{s}_i, \boldsymbol{\gamma}_i)$, $i = 1, \ldots, M$ independently drawn from $\tilde{p}_{post}(\tilde{s}, \boldsymbol{\gamma} | b)$, because in this case, the *law of large numbers* would guarantee that
$$\frac{1}{M} \sum_{i=1}^{M} (\tilde{s}_i, \boldsymbol{\gamma}_i) \quad \xrightarrow{M \to \infty} \quad (\tilde{s}_{\mathrm{CM}}, \boldsymbol{\gamma}_{\mathrm{CM}}) = \int_{\mathbb{R}^g \times \mathbb{R}^h} (\tilde{s}, \boldsymbol{\gamma})\, \tilde{p}_{post}(\tilde{s}, \boldsymbol{\gamma} | b)\, d\tilde{s}\, d\boldsymbol{\gamma}$$
almost surely and in ℓ_1 with rate $O(M^{-1/2})$, i.e., the empirical mean of the sequence converges to the expected value of the posterior (Klenke, 2008). A difficulty in our setting is that the posterior is not given in a form that allows for drawing independent samples, since it is only known up to a normalizing constant (the model-evidence) and does not belong

[1] To stress that, we will speak of CM and MAP *approximation* instead of *estimation* in the following.

to a class of distributions for which such sampling schemes are known. However, due to the *strong ergodic theorem*, the above convergence and its rate still hold if the sequence is dependent, but originates from an *ergodic Markov chain* that has $\tilde{p}_{post}(\tilde{s}, \boldsymbol{\gamma}|b)$ as its *equilibrium distribution* (Klenke, 2008). Techniques to construct such chains are called *Markov chain Monte Carlo* (*MCMC*) methods. Some of them are able to sample the posterior without knowing the model-evidence. They either work with ratios of probabilities only, or sample along some coordinates in one step while keeping the others fixed, such that the posterior conditioned on the fixed coordinates takes a simple form. For details on the theory of Markov chains and MCMC techniques, we refer to MacKay (2003); Kaipio and Somersalo (2005); Klenke (2008). Practically, the dimension g affects the time to derive a new sample, which does not affect the rate of convergence, but can render the method too slow to be a practical alternative to MAP-estimation. The speed of convergence relies heavily on the method used to construct the chain, and its *mixing properties* (MacKay, 2003). Furthermore, the generation of each sample point usually requires one or more evaluations of the forward mapping.
In summary, we will compute an approximation of the CM estimate by generating a sequence $(\tilde{s}_i, \boldsymbol{\gamma}_i)$, $i = 1, \dots, M$ such that

$$\frac{1}{M}\sum_{i=1}^{M}(\tilde{s}_i, \boldsymbol{\gamma}_i) \quad \xrightarrow{M\to\infty} \quad (\tilde{s}_{\mathrm{CM}}, \boldsymbol{\gamma}_{\mathrm{CM}}) = \int_{\mathbb{R}^g\times\mathbb{R}^h} (\tilde{s}, \boldsymbol{\gamma})\, \tilde{p}_{post}(\tilde{s}, \boldsymbol{\gamma}|b)\, d\tilde{s}\, d\boldsymbol{\gamma}$$

Since we rely on a Markov chain for the generation of the sequence, the *Markov property* assures that within this scheme, the generation of the subsequent state $(\tilde{s}_{i+1}, \boldsymbol{\gamma}_{i+1})$ only relies on the current state $(\tilde{s}_i, \boldsymbol{\gamma}_i)$, alike to the scheme we will use for approximating the MAP estimate.

The practical challenges for constructing the sequences for both MAP and CM approximation are the high dimension of the pseudo source space, and the potentially large number of hyperparameters. The main concept we will use to tackle this challenge is to exploit the special structure of the HBM in the pseudo source framework: The transfered model (2.11) is quadratic with respect to $\tilde{s}$ and factorizes over γ_i. An efficient way to exploit this for both MAP and CM approximation is to rely on alternated conditional moves only.

3.2 Alternated Conditional Walks for HBM

Basic Conditional Moves

We need to define four basic conditional moves. For this, remember that a conditional density is always proportional to the corresponding joint density by a factor only dependent on the conditioned parameter (cf. Section 2.2):

Os-Step: For a given point $(\tilde{s}_i, \boldsymbol{\gamma}_i)$, set $(\tilde{s}_{i+1}, \boldsymbol{\gamma}_{i+i}) := (\tilde{s}_{\mathrm{CMAP}}(\boldsymbol{\gamma}_i), \boldsymbol{\gamma}_i)$, where $\tilde{s}_{\mathrm{CMAP}}(\boldsymbol{\gamma})$ is the MAP estimate of $\tilde{S}$ conditioned on both $\boldsymbol{\gamma}$ and b:

$$\tilde{s}_{\mathrm{CMAP}}(\boldsymbol{\gamma}) := \operatorname*{argmax}_{\tilde{s}}\{\tilde{p}_{post}(\tilde{s}|\boldsymbol{\gamma}, b)\} = \operatorname*{argmax}_{\tilde{s}}\{\tilde{p}_{post}(\tilde{s}, \boldsymbol{\gamma}|b)\} \tag{3.1}$$

Oγ-Step: For a given point $(\tilde{s}_i, \boldsymbol{\gamma}_i)$, set $(\tilde{s}_{i+1}, \boldsymbol{\gamma}_{i+1}) := (\tilde{s}_i, \boldsymbol{\gamma}_{\mathrm{CMAP}}(\tilde{s}_i))$, where $\boldsymbol{\gamma}_{\mathrm{CMAP}}$ is the MAP estimate of $\boldsymbol{\gamma}$ conditioned on both $\tilde{S}$ and b:

$$\boldsymbol{\gamma}_{\mathrm{CMAP}}(\tilde{s}) := \operatorname*{argmax}_{\boldsymbol{\gamma}}\{\tilde{p}_{post}(\boldsymbol{\gamma}|\tilde{s}, b)\} = \operatorname*{argmax}_{\boldsymbol{\gamma}}\{\tilde{p}_{post}(\tilde{s}, \boldsymbol{\gamma}|b)\} \tag{3.2}$$

Os and *O*γ thus optimize the posterior in the direction of only one component.

Ss-Step: For a given point $(\tilde{s}_i, \boldsymbol{\gamma}_i)$, set $(\tilde{s}_{i+1}, \boldsymbol{\gamma}_{i+1}) := (\tilde{s}_{\mathrm{C}}(\boldsymbol{\gamma}_i), \boldsymbol{\gamma}_i)$, where $\tilde{s}_{\mathrm{C}}(\boldsymbol{\gamma})$ is drawn from the density of $\tilde{S}$ conditioned on both $\boldsymbol{\gamma}$ and b:

$$\tilde{s}_{\mathrm{C}}(\boldsymbol{\gamma}) :\sim \tilde{p}_{post}(\,\cdot\,|\boldsymbol{\gamma}, b) \propto \tilde{p}_{post}(\,\cdot\,, \boldsymbol{\gamma}|b) \tag{3.3}$$

Sγ-Step: For a given point $(\tilde{s}_i, \boldsymbol{\gamma}_i)$, set $(\tilde{s}_{i+1}, \boldsymbol{\gamma}_{i+1}) := (\tilde{s}_i, \boldsymbol{\gamma}_{\mathrm{C}}(\tilde{s}_i))$, where $\boldsymbol{\gamma}_{\mathrm{C}}$ is drawn from the density of $\boldsymbol{\gamma}$ conditioned on both $\tilde{S}$ and b:

$$\boldsymbol{\gamma}_{\mathrm{C}}(\tilde{s}) :\sim \tilde{p}_{post}(\,\cdot\,|\tilde{s}, b) \propto \tilde{p}_{post}(\tilde{s}, \,\cdot\,|b) \tag{3.4}$$

Ss and $S\gamma$ thus sample the posterior for one component conditioned on the other.

Composite Conditional Walks

The four basic moves can be composed to a cyclic scheme to generate a sequence in $\mathbb{R}^g \times \mathbb{R}^h$:

Algorithm 1 (Alternated Weighted Walks)
Given $\boldsymbol{\gamma}_0 \in \mathbb{R}^h$, $w : \mathbb{N} \to \mathbb{R}$, $M \in \mathbb{N}$, *do*
Initialize $\tilde{s}_0 = 0$.
For $i = 1, \ldots, M$ *do*
Set $(\tilde{s}_i, \boldsymbol{\gamma}_*) = w(i)\, Os[(\tilde{s}_{i-1}, \boldsymbol{\gamma}_{i-1})] + (1 - w(i))\, Ss[(\tilde{s}_{i-1}, \boldsymbol{\gamma}_{i-1})]$
Set $(\tilde{s}_i, \boldsymbol{\gamma}_i) = w(i)\, O\gamma[(\tilde{s}_i, \boldsymbol{\gamma}_*)] + (1 - w(i))\, S\gamma[(\tilde{s}_i, \boldsymbol{\gamma}_*)]$
Output: $(\tilde{s}_i, \boldsymbol{\gamma}_i)$, $i = 1, \ldots, M$.

Note that, as intended, the computation of $(\tilde{s}_{i+1}, \boldsymbol{\gamma}_{i+1})$ only relies on the current state $(\tilde{s}_i, \boldsymbol{\gamma}_i)$. Within this thesis, we will mainly consider alternated weighted walks for two choices of w:

- ⋆ *Alternated sampling* (AS), i.e., $w := 0$.
- ⋆ *Alternated optimization* (AO), i.e., $w := 1$.

Subsequent to the thesis, hybrid schemes for MAP approximation, where w is not constant, will be examined, too. Note, that $w \notin [0, 1]$ was not excluded, in certain situations, even $w < 0$ can be suitable, e.g., to escape from local maxima of the posterior. In Figure 3.1 an example of a AS and a AO walk are sketched for a multimodal posterior distribution.

Computing Results of Sequences

Given a sequence $(\tilde{s}_i, \boldsymbol{\gamma}_i)$, $i = 1, \ldots, M$, three quantities are interesting for MAP and CM approximation:

- ⋆ The last point, i.e.,
$$\mathrm{End}\,[(\tilde{s}_i, \boldsymbol{\gamma}_i),\, i = 1, \ldots, M] := (\tilde{s}_M, \boldsymbol{\gamma}_M)$$
- ⋆ The point with the highest posterior probability, i.e.,
$$\mathrm{MaxP}\,[(\tilde{s}_i, \boldsymbol{\gamma}_i),\, i = 1, \ldots, M] := \operatorname*{argmax}_i \{\tilde{p}_{post}((\tilde{s}_i, \boldsymbol{\gamma}_i)|b)\}$$
- ⋆ The empirical mean of the sequence, i.e.,
$$\mathrm{EmM}\,[(\tilde{s}_i, \boldsymbol{\gamma}_i),\, i = 1, \ldots, M] := \frac{1}{M} \sum_{i=1}^{M} (\tilde{s}_i, \boldsymbol{\gamma}_i)$$

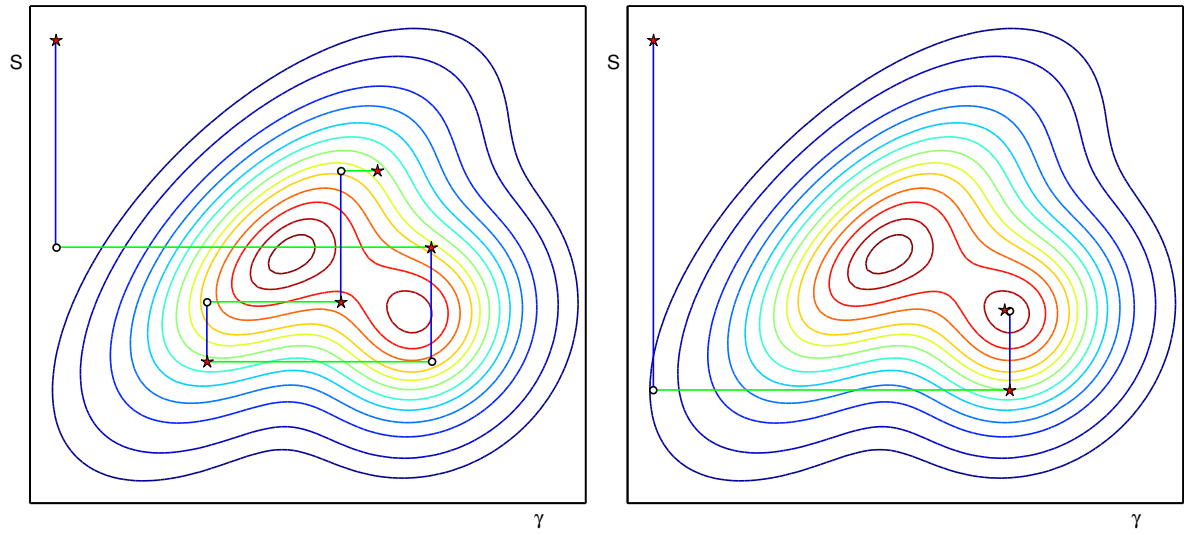


Figure 3.1: Sketch of alternated weighted walks for a multimodal posterior (plotted via contour lines). Red stars mark subsequent states, circles mark half steps. Left: AS; the blue lines correspond to Ss steps, the green lines to Sγ steps. Right: AO walk; the blue lines correspond to Os steps, the green lines to Oγ steps.

3.3 Alternated Conditional Algorithms for HBM

We are now ready to formulate the algorithms for MAP and CM approximation that we will use in our studies:

Algorithm 2 (Alternated Sampling for CM approximation (AS_CM))

Given a burn-in size Q and a sample size R do

Use AS with $M = Q$ and $\boldsymbol{\gamma}_0 = \mathbb{E}(\boldsymbol{\gamma})$ to generate the burn-in sequence $(\tilde{s}_i^b, \boldsymbol{\gamma}_i^b)$, $i = 1, \ldots, Q$.

Use AS with $M = R$ and $\boldsymbol{\gamma}_0 = \boldsymbol{\gamma}_Q^b$ to generate the main sequence.

Output: $(\tilde{s}_{AS\_CM}, \boldsymbol{\gamma}_{AS\_CM}) = EmM[(\tilde{s}_i, \boldsymbol{\gamma}_i),\ i = 1, \ldots, R]$.

This is a *blocked Gibbs sampling* method (MacKay, 2003; Gelman et al., 2003), and was used for CM approximation in Nummenmaa et al. (2007a); Calvetti et al. (2009) as well. Via a *balance condition*, one can show that the sequence is forming an ergodic Markov chain (see, e.g., MacKay, 2003; Gelman et al., 2003) which has the posterior as its equilibrium distribution. Therefore, its empirical mean converges to the CM estimate (cf. Section 3.1). This sampling technique is a very simple, but also very powerful one. A main advantage over other MCMC schemes is that it does not need any manual tuning of sampling parameters.

Concerning MAP approximation, it is important to stress that the basic AO scheme is apparently only locally convergent. If the posterior is *multimodal* (a situation we will face and illustrate in our studies in Chapter 4) the solution found by the AO scheme depends on the initialization of $\boldsymbol{\gamma}$ and may thus be suboptimal (cf. Figure 3.1). We propose three different strategies to initialize and compute a MAP approximation here, and examine their performance in the concrete studies. *AO_MAP* will abbreviate *alternated optimization for MAP approximation* in the following.

Algorithm 3 (Uniformly Initialized AO_MAP (uAO_MAP))

Given an iteration number T and an initialization rule $\boldsymbol{\gamma}_{ini}(p_{hyper}) \in \mathbb{R}$ do

Use AO with $M = T$ and $(\boldsymbol{\gamma}_0)_j = \boldsymbol{\gamma}_{ini}(p_{hyper}), \forall\, j = 1, \ldots, h$ to generate a sequence $(\tilde{s}_i, \boldsymbol{\gamma}_i)$, $i = 1, \ldots, T$.

Output: $(\tilde{s}_{uAO\_MAP}, \boldsymbol{\gamma}_{uAO\_MAP}) = End[(\tilde{s}_i, \boldsymbol{\gamma}_i),\ i = 1, \ldots, T]$.

As possible initialization rules, we will normally use the expectation or the mode of the single component hyperprior distributions $p_{hyper}(\gamma_i)$.

Algorithm 4 (Conditional Mean Initialized AO_MAP (cmAO_MAP))
Given a burn-in size Q, a sample size R and an iteration number T do
Use AS_CM with burn-in size Q and sample size R to generate a CM approximation $(\tilde{s}_{AS\_CM}, \boldsymbol{\gamma}_{AS\_CM})$.
Use AO with $M = T$ and $(\boldsymbol{\gamma}_0) = \boldsymbol{\gamma}_{AS\_CM}$ to generate a sequence $(\tilde{s}_i, \boldsymbol{\gamma}_i)$, $i = 1, \ldots,$.
Output: $(\tilde{s}_{cmAO\_MAP}, \boldsymbol{\gamma}_{cmAO\_MAP}) = End\,[(\tilde{s}_i, \boldsymbol{\gamma}_i),\ i = 1, \ldots, T]$.

Algorithm 5 (Multiple Conditional Mean Initialized AO_MAP (McmAO_MAP))
Given a number of seeds U, a burn-in size Q, a sample size R and an iteration number T do
For $l = 1, \ldots, U$, do
Use cmAO_MAP with burn-in size Q, sample size R and iteration number T to generate $(\tilde{s}_{cmAO\_MAP}, \boldsymbol{\gamma}_{cmAO\_MAP})$ and set $(\tilde{s}_l, \boldsymbol{\gamma}_l) = (\tilde{s}_{cmAO\_MAP}, \boldsymbol{\gamma}_{cmAO\_MAP})$.
Output: $(\tilde{s}, \boldsymbol{\gamma})_{McmAO\_MAP} = MaxP\,[(\tilde{s}_i, \boldsymbol{\gamma}_i),\ i = 1, \ldots, U]$.

The McmAO_MAP approach seems a bit heuristic at this point. However, it is intended to be used with very small values for Q and R compared to cmAO_MAP. In addition a blocked implementation of the for loop is used, i.e., all U cmAO_MAP results are computed simultaneously (see Section 3.6). It will turn out, that this way, Mcm_MAP yields the best MAP approximation of the above algorithms at moderate computation times.

3.4 Implementation for Gaussian Scale Mixture Models

In this section, the practical implementation of the four basic conditional moves (cf. Section 3.2) for the HBM given by (2.11) will be discussed.

3.4.1 Implementation of Os and Ss Steps

The fact that the energy of the posterior is quadratic with respect to $\tilde{s}$ allows that both the Os and the Ss step can be computed directly by solving systems of linear equations. We start with the computation of the Os step. Although not apparent at this point, we will then see that the Ss step can be implemented in a surprisingly similar fashion.
In (3.1) the values of the hyperparameters are fixed, and the minimization does not depend on the hyperprior. Thus we effectively compute the WMNE (cf. 1.3.2) for the diagonal weighting matrix $\mathrm{D}(\boldsymbol{\gamma}) := \tilde{\Sigma}_{\tilde{s}}(\boldsymbol{\gamma})$:

$$\tilde{s}_{\mathrm{CMAP}}(\boldsymbol{\gamma}) \overset{(2.11)}{=} \underset{\tilde{s}\in\mathbb{R}^g}{\operatorname{argmin}} \left\{ \|b - \tilde{\mathrm{L}}\tilde{s}\|^2 + \sigma^2 \|\mathrm{D}^{-1/2}\tilde{s}\|^2 \right\} \overset{(1.6)}{=} \mathrm{D}\tilde{\mathrm{L}}^t \left(\tilde{\mathrm{L}}\mathrm{D}\tilde{\mathrm{L}}^t + \sigma^2 \mathrm{Id}_m \right)^{-1} b \tag{3.5}$$

Nevertheless, concerning computation time and stability, it is preferable to solve the corresponding relaxed weighted least squares problem (cf. A.1.1) iteratively:

$$(3.5) \overset{A.1.1}{\Longleftrightarrow} \begin{bmatrix} \tilde{\mathrm{L}} \\ \sigma\,\mathrm{D}^{-1/2} \end{bmatrix} \tilde{s}_{\mathrm{CMAP}}(\boldsymbol{\gamma}) \overset{ls}{=} \begin{bmatrix} b \\ 0 \end{bmatrix} \tag{3.6}$$

$$\Longleftrightarrow \begin{bmatrix} \tilde{\mathrm{L}}\,\mathrm{D}^{1/2} \\ \sigma \mathrm{Id}_g \end{bmatrix} y \overset{ls}{=} \begin{bmatrix} b \\ 0 \end{bmatrix} \qquad \text{with} \qquad y = \mathrm{D}^{-1/2}\,\tilde{s}_{\mathrm{CMAP}}(\boldsymbol{\gamma}) \tag{3.7}$$

This can be done by using *Krylov subspace methods* such as the *conjugate gradient least squares* method (*CGLS*) (see Section 3.5) with the special preconditioning by $\mathrm{D}^{-1/2}(\boldsymbol{\gamma})$ as formulated

in (3.7). If the hyperparameters were fixed in our model, this would correspond to a *whitening transform* of the random variable $\tilde{S}$ (cf. A.1.5). Applied to iterative solvers for inverse problems, this technique is called *priorconditioning* (Calvetti and Somersalo, 2007b). In our hierarchical framework, the prior covariance itself is not fixed but relies on the fixation of the hyperparameters on their current values. The idea of using this present state of information, updated in every step of composite conditional walks is referred to as a *hyperpriorconditioning* (Calvetti et al., 2009). Note that the uAO_MAP algorithm is a generalization of the *Iterative Alternating Sequential* (*IAS*) algorithm which was introduced by Calvetti and Somersalo (Calvetti and Somersalo, 2007a, 2008a; Calvetti et al., 2009), inspired by a similar, more general algorithm called *half quadratic minimization* (Aubert and Kornprobst, 2006). The IAS algorithm relies on a a specific HBM, and uses formulation (3.7) with the CGLS method stopped after a few iterations.
The sampling of $\tilde{s}_{\mathrm{C}}$ in the Ss step is done indirectly: In principle, since the hyperparameters are known one can calculate the mean and the variance of the resulting Gaussian distribution via:

$$\mathbb{E}_{p(\tilde{s}|\boldsymbol{\gamma},b)}(\tilde{s}) = \mathrm{D}\tilde{\mathrm{L}}^t\left(\tilde{\mathrm{L}}\mathrm{D}\tilde{\mathrm{L}}^t + \sigma^2\mathrm{Id}_m\right)^{-1} b \tag{3.8}$$

$$\mathbb{C}\mathrm{ov}_{p(\tilde{s}|\boldsymbol{\gamma},b)}(\tilde{s}) = \mathrm{D} - \mathrm{D}\tilde{\mathrm{L}}^t\left(\tilde{\mathrm{L}}\mathrm{D}\tilde{\mathrm{L}}^t + \sigma^2\mathrm{Id}_m\right)^{-1}\tilde{\mathrm{L}}\mathrm{D} \overset{(A.8)}{=} \left(\mathrm{D}^{-1} + \frac{1}{\sigma^2}\tilde{\mathrm{L}}^t\tilde{\mathrm{L}}\right)^{-1} \tag{3.9}$$

and sample $\tilde{s}_{\mathrm{C}}$ from this distribution directly (see the second subsection of Section A.1.4 for the computation and the first for the sampling). Even so, similar to the Os step this is computationally expensive and especially the computation of the covariance matrix is highly unstable, since D is often almost singular. Instead, one can use a mixing strategy (cf. Section 2.4.3, A.1.4), i.e., we draw ω_m and ω_g from standard normal distributions of dimension m and g and solve:

$$\begin{bmatrix}\tilde{\mathrm{L}}\\ \sigma\,\mathrm{D}^{-1/2}\end{bmatrix}\tilde{s}_{\mathrm{C}}(\boldsymbol{\gamma}) \overset{ls}{=} \begin{bmatrix}b\\0\end{bmatrix} + \sigma\begin{bmatrix}\omega_m\\ \omega_g\end{bmatrix} \tag{3.10}$$

The proof that the solution $\tilde{s}_{\mathrm{C}}(\boldsymbol{\gamma})$ of (3.10) is really distributed according to (3.8) and (3.9) is done in the appendix (see A.1.4). Comparing (3.6) and (3.10) immediately shows that the computations for the Os and the Ss step can be carried out in a similar fashion, only the right hand side of the least squares problem is modified. Thus everything stated above also applies to the Ss step, especially for the preconditioning by $\mathrm{D}^{-1/2}(\boldsymbol{\gamma})$. For alternated weighted walks with $w \notin \{0,1\}$, the computations of both steps can also be combined relying on this formulation.
Using preconditioned iterative solvers for the problems (3.6) and (3.10) was proposed in Calvetti et al. (2009) and seems to be a canonical choice with regard to the high dimension of the problem. The advantage is that these schemes can easily be transferred to other fields of inverse problems, where the forward mapping is not given in explicit matrix form (Kaipio and Somersalo, 2005; Calvetti and Somersalo, 2007a,b, 2008a,b). In addition, we will see in Section 3.6 that iterative solvers allow for the construction of blocked inversion schemes, where multiple right hand sides are inverted simultaneously which results in a considerable gain in speed. Consequently, for the work on this thesis, much efforts were made to optimize these iterative approaches, and most of the results were computed with them. Recently, a very simple alternative implementation was developed that is competitive to the iterative approaches in terms of computation speed: Due to the small number of sensors (we usually use $m < 150$), the block-structure of (3.6) and (3.10) and the identity (A.8), the explicit solution of the systems can be computed very efficiently: Starting from (3.10) (the formula for (3.6) follows by setting $\omega_m = 0$, $\omega_g = 0$), we multiply by σ^{-1} and from A.1.1 and (A.8) it follows that

$$\begin{aligned}\tilde{s}_{\mathrm{C}} &\overset{A.1.1}{=} \left(\mathrm{D}^{-1} + \sigma^{-2}\tilde{\mathrm{L}}^t\tilde{\mathrm{L}}\right)^{-1}\left[\sigma^{-1}\tilde{\mathrm{L}} \quad \mathrm{D}^{-1/2}\right]\left(\begin{bmatrix}\sigma^{-1}b\\0\end{bmatrix} + \begin{bmatrix}\omega_m\\ \omega_g\end{bmatrix}\right)\\ &\overset{(A.8)}{=} \left(\mathrm{D} - \mathrm{D}\tilde{\mathrm{L}}^t\left(\tilde{\mathrm{L}}\mathrm{D}\tilde{\mathrm{L}}^t + \sigma^2\mathrm{Id}_m\right)^{-1}\tilde{\mathrm{L}}\mathrm{D}\right)\left(\tilde{\mathrm{L}}^t(\sigma^{-2}b + \sigma^{-1}\omega_m) + \mathrm{D}^{-1/2}\omega_g\right)\end{aligned}$$

This formula can be implemented in a straight forward manner:

Algorithm 6 (Analytical Os/Ss Solution)

1. *Set* $r = \left(\tilde{\mathrm{L}}^t(\sigma^{-2}b + \sigma^{-1}\omega_g) + \mathrm{D}^{-1/2}\omega_g\right)$;
2. *Set* $s_1 = \mathrm{D}\, r$;
3. *Set* $t = \tilde{\mathrm{L}}\, s_1$;
4. *Set* $\tilde{\Sigma}_b = \left(\tilde{\mathrm{L}}\mathrm{D}^{1/2}\right)\left(\tilde{\mathrm{L}}\mathrm{D}^{1/2}\right)^t + \sigma^2\mathrm{Id}_m$;
5. *Solve* $\tilde{\Sigma}_b\, x = t$;
6. *Set* $s_2 = \mathrm{D}\tilde{\mathrm{L}}^t\, x$;
7. *The solution is given by* $\tilde{s}_C = s_1 - s_2$;

Remember that the multiplication with D can be performed componentwise. The computation of the projected source covariance $\tilde{\mathrm{L}}\mathrm{D}\tilde{\mathrm{L}}^t$ within step 4. is the most computationally intensive part of the algorithm, solving the linear system in step 5. is far less demanding: The system is only of size $m \times m$ and is symmetric positive definite. A solution via Cholesky decomposition is still fast enough to be negligible in comparison to the matrix-matrix multiplication in step 4. The solution of (3.10) with this algorithm is considerably faster than with iterative solvers (see A.1.10), and finding an optimal implementation is less demanding. Furthermore, it yields the exact solution of (3.10) within the bounds posed by ill-condition and finite precision, and no stopping criteria have to be chosen ad hoc. Another advantage is that the computation time is effectively independent of the right hand side, which is not the case for the iterative solvers we applied: Empirically, it was observed that more complex source configurations also result in a slower convergence of the CGLS algorithm.

3.4.2 Implementation of Oγ and Sγ Steps

The fact that the posterior factorizes over γ_i enables that both Oγ and Sγ steps can be computed componentwise. Extracting the hyperparameter dependent part of (2.11) gives

$$\begin{aligned} \tilde{p}_{post}(\boldsymbol{\gamma}, \tilde{s}|b) &\propto \exp\left(-\frac{1}{2}\sum_{i=1}^{h}\left(\frac{\tilde{s}_i^t\tilde{s}_i}{\gamma_i} + \varrho_i \ln(\gamma_i) + f_i(\gamma_i)\right)\right) \\ \Longrightarrow \quad \tilde{p}_{post}(\gamma_i, \tilde{s}|b) &\propto \exp\left(-\frac{1}{2}\left(\frac{\tilde{s}_i^t\tilde{s}_i}{\gamma_i} + \varrho_i \ln(\gamma_i) + f_i(\gamma_i)\right)\right) \end{aligned} \tag{3.11}$$

Dependent on the specific hyperprior, maximizing (3.11) for γ_i to compute the Oγ step can be solved explicitly or has to be solved by a numerical approach, e.g., a Newton's method. Concerning the Sγ step, techniques like the *inverse cumulative distribution method* (Kaipio and Somersalo, 2005) or *slice sampling* (Neal, 2003) can be used to handle a wide range of possible hyperprior. However, the choice of the inverse gamma distribution as a hyperprior causes (3.11) to be inverse gamma distributed as well (due to the *conditional conjugacy* of prior and hyperprior), and once the corresponding distribution parameter are computed, standard sampling methods for gamma distributions can be used.
Concrete implementations will be discussed for the specific HBM introduced in Chapter 4.

3.5 Conjugate Gradient Method for Least Squares Problems

The solution of a linear least squares problem $\mathrm{G}x \overset{ls}{=} c$ with $\mathrm{G} \in \mathbb{R}^{M\times N}, M > N, \mathrm{rank}(\mathrm{G}) = N$ is given by the solution of the *normal equations* (cf. A.1.1):

$$\mathrm{G}^t\mathrm{G}x = \mathrm{G}^t c \tag{3.12}$$

But especially in many applications in inverse problems, it is not preferable to calculate and store $\mathrm{G}^t\mathrm{G}$ explicitly and solve (3.12) by some direct or iterative method: $\mathrm{G}^t\mathrm{G}$ is usually large and dense and its condition is even larger than the condition of G alone. Consider, e.g., (3.7), where $\mathrm{G}^t\mathrm{G} = (\mathrm{D}^{1/2}\tilde{\mathrm{L}}^t\tilde{\mathrm{L}}\mathrm{D}^{1/2} + \sigma^2\mathrm{Id}_g) \in \mathbb{R}^{g\times g}$. In addition, in many applications, the system matrix (which is part of G, cf. (3.6)) is not given explicitly or it is not preferable to compute it, but its action (and the action of G^t) on vectors can be computed with ease (e.g., in image deblurring, it is advantageous to compute the convolution with the point spread function by means of the *fast Fourier transformation* (*FFT*)).
For such cases, specific iterative solvers have been developed that work with matrix-vector products involving G and G^t alone. These solvers are problem specific implementations of iterative *Krylov subspace methods* (see Björck, 1996 for a general introduction). Define

$$\begin{aligned} r_l &:= \mathrm{G}^t(c - \mathrm{G}x_l) \\ \mathcal{K}_l^\dagger(\mathrm{G}, c) &:= \operatorname{span}\left\{\mathrm{G}^t c, (\mathrm{G}^t\mathrm{G})\mathrm{G}^t c, \ldots, (\mathrm{G}^t\mathrm{G})^{l-1}\mathrm{G}^t c\right\}, \end{aligned}$$

as l-th *residual error* and *Krylov subspace* of the normal equations. Krylov subspace methods seek to minimize $\|r_l\|$ over $\mathcal{K}_l^\dagger(\mathrm{G}, c)$:

$$x_l := \underset{x\in\mathcal{K}_l^\dagger(\mathrm{G},c)}{\operatorname{argmin}} \|\mathrm{G}^t(c - \mathrm{G}x)\|^2 \tag{3.13}$$

Mathematically, all methods would result in the same sequence of approximations x_l, and reach the exact solution after n steps. Still, as discussed above, due to finite precision and ill-conditioning, naive implementations might need more than n steps, might not converge to the exact solution or might not even converge at all. The concrete method we will use to solve (3.13) is called *conjugate gradient least squares* (*CGLS*) or *conjugate gradient normal residual* (*CGNR*) method. It can also be derived as a *line-search optimization* method for a quadratic functional:

Algorithm 7 (CGLS Algorithm)
Given the right hand side c, initialize:

$$\begin{aligned} x_0 &= 0; \\ d_0 &= c - \mathrm{G}x_0; \\ r_0 &= \mathrm{G}^t d_0; \\ p_0 &= r_0; \\ y_0 &= \mathrm{G}p_0; \end{aligned}$$

Iteration: For l = 1,2,... until a stopping criterion is satisfied

$$\begin{aligned} \alpha &= \frac{\|r_{l-1}\|}{\|y_{l-1}\|}; \\ x_l &= x_{l-1} + \alpha p_{l-1}; \\ d_l &= d_{l-1} - \alpha y_{l-1}; \\ r_l &= \mathrm{G}^t d_l; \\ \beta &= \frac{\|r_l\|^2}{\|r_{l-1}\|^2}; \\ p_l &= r_l + \beta p_{l-1}; \\ y_l &= \mathrm{G}p_l; \end{aligned}$$

This specific implementation of the ordinary conjugate gradient method needs one multiplication with G and one with G^t per iteration (or an equivalent implementation of this operation), but does not need the explicit formation of G^tG. Nevertheless, without good preconditioning at hand, more stable (but computationally more expensive) implementations like *LSQR* or *GMRES* should be used (Björck, 1996): Note that the right hand side c enters the algorithm only in the initialization, hence ill-conditioning and finite precision play an important role for the convergence. With the preconditioning introduced in Section 3.4 no problems have been encountered for CGLS in practice. To attain a good computation speed for systems like (3.10), it is crucial to exploit the special structure of G: Note that it consists of three matrices with different properties: $\tilde{L}$ is dense but small compared to the others, $D^{1/2}$ is a diagonal matrix and σId_g the scaled identity. Computing and storing G (necessarily as a sparse matrix) and using it directly to calculate its action on vectors results in unnecessary overhead. Decomposing this mapping, computing the subparts and reassembling the results afterwards leads to an enormous speed-up.

3.6 Single vs. Blocked Inversion Schemes

For our studies in Chapter 4, a large number of different measurements b have to be considered. Normally, one would invert each b separately using the approximation algorithms discussed in Section 3.3. This would result in a linear dependence of the computation time on the number of samples. However, using the pseudo source decomposition ensures that the action of both operators that depend on the right hand side in the Os and Ss step, i.e., $D^{1/2}$ and $\sigma Id_g{}^2$, can be computed by componentwise multiplication. This, and the special form of the CGLS algorithm allows us to design a blocked inversion scheme, where all measurements b are inverted simultaneously. The basic advantage of this is that the most time-consuming part of the CGLS algorithm, i.e., the matrix-vector multiplications of $\tilde{L}$ or $\tilde{L}^t$ with the corresponding iteration *vectors* are replaced by matrix-matrix multiplications of $\tilde{L}$ or $\tilde{L}^t$ with corresponding iteration *matrices*. The number of columns of these matrices is given by the number of right hand sides b. Especially when implemented on programming platforms that are optimized for array operations and used on modern multi-core CPU or even GPU devices (see Section 3.8) the computation time for matrix-matrix multiplications $A \cdot X$ is only a fraction of the time needed for a single matrix-vector multiplication $A \cdot x$ times the number of columns of X. This issue will be illustrated in more detail in the appendix in Section A.1.10. The implementation needs some care and especially for the AS scheme, a good memory management is needed as well.
Algorithm 6 to compute the analytical solution of the Os and Ss step cannot be formulated in a blocked form. As a consequence, using the blocked inversion scheme with CGLS outperforms using the single inversion scheme with the analytical solution for large studies (see Section A.1.10).

3.7 Computation of the Earth Mover's Distance

To compute the p^{th}-EMD (see Definition 8 in Section 1.3.3) between real and estimated source activity, both are transferred into discrete probability distributions: For the real source activity j_{real}, a suitable discretization by some localized basis functions $\psi_i(x)$ (not necessarily the $j_{i,l}$ used for the source space discretization) has to be chosen:

$$j_{real}(x) \approx \sum_{i=1}^{\tau} M_i \cdot \psi_i(x) \qquad \forall x \in \Omega$$

[2] In our studies we will assume a fixed noise level, i.e., σ will be computed from b for each b separately rather than being fixed for the whole study (cf. 4.4.1)

Now let $\breve{r}_i$ be the midpoint of $\mathrm{supp}(\psi_i)$ and define a discrete *signature* P by:

$$P = \{(p_1, w_{p_1}), \ldots, (p_\tau, w_{p_\tau})\} \quad \text{with} \quad p_i := \breve{r}_i; \quad w_{p_i} := \frac{|M_i|}{M_{tot}}; \quad M_{tot} = \sum_{i=1}^{\tau} |M_i|$$

For the estimated CDR, we define a signature Q by:

$$Q = \{(q_1, w_{q_1}), \ldots, (q_l, w_{q_k})\} \quad \text{with} \quad q_i := r_i; \quad w_{q_i} := \frac{\| s_{i*} \|_2}{a_{tot}}; \quad a_{tot} = \sum_{i=1}^{k} \| s_{i*} \|_2$$

Finally, define the distance matrix D^p by letting $\mathrm{D}^p_{(i,j)}$ be the p^{th} power of the 3D-Euclidean distance between p_i and q_j. Now we are ready to formulate the computation of the p^{th}-EMD between P and Q as a linear programming problem as formulated by Kantorovich (Kantorovich, 1942; Kantorovich and Gavurin, 1949). In our setting, computing (1.7) becomes:

Definition 9 (Reformulation of the EMD) *With the above definitions, find a transport plan* $\Gamma \in \mathbb{R}^{\tau \times k}$ *that minimizes the work*

$$\mathcal{W}(P, Q, \Gamma) = \sum_{i=1}^{\tau} \sum_{j=1}^{k} \mathrm{D}^p_{i,j} \cdot \Gamma_{i,j} \tag{3.14}$$

subject to the following constraints:

$$\Gamma_{i,j} \geq 0, \qquad 1 \leq i \leq \tau,\ 1 \leq j \leq l \tag{3.15}$$

$$\sum_{i=1}^{\tau} \Gamma_{i,j} = w_{p_i}, \qquad 1 \leq i \leq \tau \tag{3.16}$$

$$\sum_{j=1}^{k} \Gamma_{i,j} = w_{q_j}, \qquad 1 \leq j \leq l \tag{3.17}$$

$$\tag{3.18}$$

The minimal work resulting from this computation is the EMD between P and Q. The constraints (3.15) - (3.17) ensure that Γ is a valid transport plan:

(3.15) ensures that the mass is transfered from P to Q and not vice versa.

(3.16) determines the amount of mass that has to be transfered from one position.

(3.17) determines the amount of mass that has to be transfered into one position.

There are efficient algorithms to solve this linear programming problem exploiting its special structure. However, in the studies we are performing for this thesis, the size of P is usually very small, and the problem can be solved with standard linear programming toolboxes with negligible time costs. The transformation of (3.14) into standard form can be found in Section A.1.6.

3.8 Implementation

The main advantage of the hierarchical modeling is the flexible, modular, level-based construction. To retain these advantages, an object-oriented implementation of the whole scheme is desired. Matlab is an adequate environment for this intention, as it supports object oriented programming and is a good choice from both computational and practical point of view: The computationally intensive parts are the CGLS-iterations, which mainly consist of matrix-vector

multiplications in the single inversion scheme and matrix-matrix multiplications in the blocked inversion scheme. It is hard to beat Matlab's performance on these tasks with own handcrafted code, especially due to the implicit multi-threading capabilities. Furthermore, Matlab offers the possibility to easily test alternative solvers and sampling routines. Most of the existing toolboxes for EEG/MEG are written for Matlab, and after the thesis, we would like to compare our results with methods like VB-estimation, implemented, e.g., in SPM8[3]. Furthermore, from Matlab as a basis platform one can easily call other, command line driven, programs: The toolbox provides interfaces to SCIRun[4] for visualization and to SimBio[5] for the FEM forward simulation in an automated fashion. Its working title is *BayesNEMESIS* (***Bayes****ian* ***N****euro****E****lectro****M****agn****E****tic* ***S****ource* ***I****maging* ***S****oftware*). A detailed description of its structure and a discussion of particular implementations is not topic of this thesis, and we will thus only give a rough sketch of its contents here. The toolbox relies on three main classes: Model, prior and estimator:

- ⋆ The model class serves as:
  - ∗ An interface to create and modify forward models for EEG/MEG computations. In particular, three types of models are supported:
    1. Isotropic rectangular models: A homogeneous isotropic conductivity for the whole space is assumed, and source nodes and sensors are arranged in regular grids. The lead-field matrix can be calculated by an explicit formula in this case (see e.g., Calvetti et al. (2009)) and the visualization is carried out by Matlab (see, e.g., Figure 2.1).
    2. Anisotropic multi-layered sphere models: The volume conductor consists of multiple concentric spheres and each layer is given a homogeneous conductivity, where anisotropic conductivities are allowed. The lead-field matrix is calculated by an asymptotic series expansion formula Munck and Peters (1993); De Munck (1988) and the visualization is carried out by Matlab. This model is not used within this thesis.
    3. Tetrahedral Finite Element head models: The main class of models that are used within this thesis. Their use and creation will be illustrated in Section 4.3.1. A large part of the whole toolbox consists of functions to work with such models and to handle them in an efficient way. All visualization for these models is carried out by an interface to SCIRun.
  - ∗ An interface that gathers and provides all the information of the forward model that is needed to invert the data, e.g., the lead-field matrix and the noise model.
  - ∗ Functions to create several types of mesh-free source configurations, to generate the corresponding measurement data, and to validate the performance of inverse methods applied to this data by means of the methods introduced in Section 1.3.3.
- ⋆ The prior class contains and provides all the information about the HBM.
- ⋆ The estimator class defines a superclass for different inverse methods. These schemes have been implemented so far:
  - ∗ MNE, sLORETA and different WMNE schemes and methods for the choice of the regularization parameter.
  - ∗ All the methods introduced in Section 3.3 are implemented in a common interface that initiates and manages the alternated weighted walks needed for the computation.

[3]For information on SPM8, see: http://www.fil.ion.ucl.ac.uk/spm/software/spm8/
[4]For information on SCIRun, see: http://www.sci.utah.edu/software.html
[5]For information on SimBio, see: https://www.mrt.uni-jena.de/simbio/index.php/Main_Page

4 Simulation Studies

4.1 Motivation

The choice of the locations of the source nodes is a crucial point for CDRs, and will be discussed in the following. Since the neural generators of the EEG/MEG signal are located in parts of the gray matter (cf. Section 1.1) a fine volumetric discretization of this thin, layered compartment would be preferable. However, this approach is constrained by the available structural information: The cortex involves deep but thin sulci and is strongly folded. To attain a detailed volumetric representation of it, structural imaging scans (CT or MRI) with a high imaging resolution and sophisticated segmentation algorithms are needed. Instead, many approaches try to segment the cortical surface only. Furthermore smoothing is used frequently, which results in a flattened surface representation. In addition, deep-lying gray matter areas, or areas encased by white matter, e.g., the insular, the cingulate cortex, the hippocampus (see below) or the thalamus are often not represented. Working with such surface representations is reasonable and advantageous for a wide range of experimental designs, e.g., when the activity is known to occur only in superficial cortical areas and the location of these areas is important, whereas depth information is not of interest. The concrete source locations r_i are then restricted to the segmented surface and the normal constraint is used (i.e., $d = 1$, cf. Section 1.3). In the absence of structural information, a spherical or ellipsoidal surface is used (the parameters for these surfaces can, e.g., be obtained by maximizing the *free energy* of the model, see Sato et al., 2004 for details).
Nevertheless, often the active brain networks involve deep-lying components of the cortex as well. One example are networks involving the *hippocampus*, an archicortex component of the limbic system (see Figure 4.1). It plays an important role in episodic or autobiographical memory, spatial memory and navigation (Duvernoy, 2005; Andersen, 2007). Concerning its pathology, the hippocampus is often the focus of epileptic seizures: Hippocampal sclerosis is the most commonly visible type of tissue damage in temporal lobe epilepsy (Chang and Lowenstein, 2003; Stefan et al., 2009). Other pathologic implications include, e.g., Alzheimer's disease (Duvernoy, 2005; Andersen, 2007). As a result a number of networks which are interesting from clinical or scientific point of view include the hippocampus as an active component[1]. For these networks, a complete representation of the cortex is mandatory.

[1] At the moment, EEG is actually the main diagnostic toll for presurgical epilepsy diagnosis

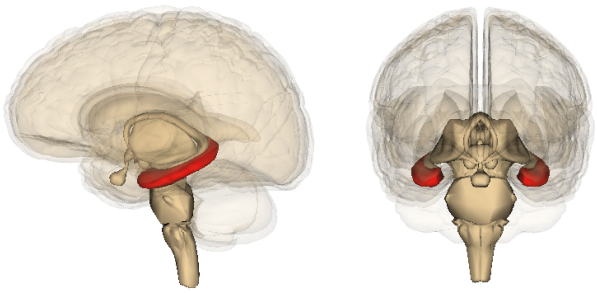


Figure 4.1: The hippocampus compartment (red)

Source: Wikimedia Commons, file: Hippocampus_image.png

When the complete gray matter of a high resolution MRI scan is segmented and its volume is discretized, many more deep-lying locations form the source space and a phenomena called "depth bias" gains fundamental importance for the correct localization of source activity: Many inverse methods fail to reconstruct deep-lying sources in the right depth, reconstructing them too close to the skull (cf. Figures A.18, A.19 and A.20). This is a well known systematic error (e.g., Ahlfors et al., 1992; Wang et al., 1992; Gencer and Williamson, 1998) and was subject to many studies (e.g., Ioannides et al., 1990; Pascual-Marqui, 1999b; Fuchs et al., 1999; Pascual-Marqui, 2002; Greenblatt et al., 2005; Sekihara et al., 2005; Lin et al., 2006; Grave de Peralta et al., 2009). The depth bias can be a crucial error, e.g., in the presurgical diagnosis for epilepsy patients, where the task is to determine the right location of the resection volume. Still, a deep mathematical analysis has not been undertaken, yet (to the best of my knowledge). There are inverse methods that do not show this error, but they suffer from other drawbacks instead (e.g., sLORETA, see Section 1.3.2). Again, it is not well understood yet, why those methods do not show a depth bias.
Another effect related to the depth bias is the masking of deep-lying sources by superficial ones: If the real source configuration consists of multiple, spatially separated sources with different depths, many inverse methods only recover the sources close to the skull. This effect can lead to crucial errors in the presurgical diagnosis for epilepsy patients suffering from multi focal epileptiform discharges: This form of epilepsy is correlated to a worse postoperative outcome regarding seizure freedom and complicates the presurgical diagnosis (Chang and Lowenstein, 2003). The correct detection and separation of multiple sources is hence of greatest importance to guide the presurgical diagnosis and operation planning.

For the reasons presented above, we want to examine if the Full-MAP and Full-CM approximation methods which were proposed in Section 3.3 and then computed for a specific HBM can improve upon commonly used inverse methods applied to the following scenarios:

- ⋆ Study 1: Localization properties and depth bias for single focal sources.
- ⋆ Study 2: Recovery and separation of multiple focal sources. Masking of deep-lying sources by superficial ones.

To ease the interpretation of the results of the main studies, a number of preliminary examinations are carried out in advance. Furthermore, we are interested in comparing different performance measures for our purpose, with a focus on the usage of Wasserstein distances for multiple-source scenarios (see Section 1.3.3 and 3.7). Although not apparent at this stage, the main motivation to examine HBMs in this thesis was one result from Calvetti et al. (2009): Within a simplified geometry, a single deep-lying source was reconstructed (cf. Figures 1-4 on page 894 in Calvetti et al., 2009). The CM approximation with an inverse gamma hyperprior yielded the best result, both in location and in extend of the estimated source. Moreover, it seemed to have no depth bias whereas MAP approximation by the IAS algorithm for the same HBM seemed to suffer from it. Within the publication the topic was not examined any further, and especially, only a single source configuration within a realistic geometry was considered. In this thesis these issues will be pursued with realistic head modeling and a large number of reconstructions. In the following, we will introduce a HBM for the recovery of source configurations consisting of a few and focal sources, and present the general setting in which our studies are carried out.

4.2 Hierarchical Modeling of Sparse Source Configurations

The HBM we use to represent focal source activity has been introduced in Sato et al. (2004) and has further been examined in Nummenmaa et al. (2007a,b); Calvetti et al. (2009). We will

rely on the introduction given by Calvetti et al. (2009) to illustrate it: The aim is to formulate a HBM for the instantaneous reconstruction of currents consisting of *few* and *focal* sources. This a-priori information is of *qualitative* nature:

1. Nearby source elements should a-priori not be mutually dependent, to favor focality.
2. No location preference for activity should be given a-priori.
3. Most of the dipole-like sources should be silent, while few could have a large amplitude.

To construct a HBM out of this information, a covariance set

$$\mathcal{C} = \left\{ \boldsymbol{e}_i \boldsymbol{e}_i^t \otimes \mathrm{Id}_d,\ i = 1, \cdots, k \right\}$$

is chosen and conditions 1.-3. are transcribed into the construction of the hyperprior:

$\xrightarrow{1.}$ The hyperparameters γ_i should be stochastically independent. This is assumed by default in our framework.

$\xrightarrow{2.}$ The variances γ_i should be equally distributed.

$\xrightarrow{3.}$ A possible choice to realize this is given by the *generalized gamma distribution*, a distribution often used to model random variables describing *scale variables* as the γ_i in our model. Its shape is determined by three parameters that can change the distribution from promoting a certain scale to being (nearly) scale invariant. The latter choice allows some γ_i to have a large amplitude (and thus the associated dipole-like source is allowed to show significant activity), while all others have a very small amplitude.

The choice of the generalized gamma distribution as a hyperprior leads to:

$$p_{hyper}(\gamma) \propto \prod_{i=1}^{h(=k)} \gamma_i^{\zeta\alpha-1} \exp\left(-\frac{\gamma_i^\zeta}{\beta^\zeta}\right) = \exp\left(-\sum_{i=1}^{h} \frac{\gamma_i^\zeta}{\beta^\zeta} + (\zeta\alpha - 1) \sum_{i=1}^{h} \ln \gamma_i\right) \tag{4.1}$$

The parameters $\alpha > 0$ and $\beta > 0$ determine *shape* and *scale* of the distribution, whereas ζ distinguishes between different classes of two-parameter scale distributions, e.g., $\zeta = -1$ yields the inverse gamma distribution, $\zeta = 1$, yields the (standard) gamma distribution. Although the methods should work for a broader range of ζ, we will restrict ourselves to these two cases for MAP estimation and to the inverse gamma case for the CM estimation. In Section A.1.7 the main properties of both distributions are discussed. The full posterior is now given by (cf. (2.11), $\varrho_i = 3, h = k$):

$$p_{post}(s, \boldsymbol{\gamma}|b) \propto \exp\left(-\frac{1}{2}\left(\frac{1}{\sigma^2}\|b - \mathrm{L}\, s\|_2^2 + \sum_{i=1}^{k} \frac{\|s_{i*}\|^2}{\gamma_i} + 2\sum_{i=1}^{k}\left(\frac{\gamma_i}{\beta}\right)^{\pm 1} - 2\left(\pm\alpha - \frac{5}{2}\right)\sum_{i=1}^{k} \ln \gamma_i\right)\right) \tag{4.2}$$

As discussed in the beginning of Chapter 3, for this choice of covariance set, real source vectors s_{i*} and pseudo sources $\tilde{s}_i$ coincide and accordingly (4.2) already factorizes over γ_i.

For this concrete HBM we are now able to complete the computation of the Oγ and Sγ step based on Section 3.4.2. The hyperparameter dependent single component part of the posterior now reads (cf. (3.11)):

$$p_{post}(\gamma_i|s, b) \propto \exp\left(-\frac{1}{2}\left(\frac{\|s_{i*}\|^2}{\gamma_i} + 2\left(\frac{\gamma_i}{\beta}\right)^{\pm 1} - 2\left(\pm\alpha - \frac{5}{2}\right)\ln \gamma_i\right)\right) \tag{4.3}$$

For the gamma hyperprior (i.e., "+" in the above formula), the conditional update $\gamma_{\mathrm{CMAP},i}$ is given by:

$$\gamma_{\mathrm{CMAP},i} = \underset{\gamma_i}{\operatorname{argmax}} \left\{ \exp\left(-\frac{1}{2} \left(\frac{\|s_{i*}\|^2}{\gamma_i} + 2\left(\frac{\gamma_i}{\beta}\right)^{\pm 1} - 2\left(\pm\alpha - \frac{5}{2}\right) \ln \gamma_i \right) \right) \right\} \tag{4.4}$$

$$= \underset{\gamma_i}{\operatorname{argmin}} \left\{ \frac{\|s_{i*}\|^2}{\gamma_i} + 2\left(\frac{\gamma_i}{\beta}\right)^{\pm 1} - 2\left(\pm\alpha - \frac{5}{2}\right) \ln \gamma_i \right\} \tag{4.5}$$

Computing the first and second order conditions yields

$$\text{First order:} \qquad 0 = \gamma_i^2 - \eta\beta\gamma_i - \frac{\|s_{i*}\|^2}{2}\beta \quad \text{where} \quad \eta := (\alpha - 5/2)$$

$$\text{Second order:} \qquad 0 \leqslant \frac{\|s_{i*}\|^2}{\gamma_i^3} + \frac{(\alpha - 3/2)}{\gamma_i^2}.$$

The second order condition is fulfilled if $\alpha \geqslant 1.5$. The first order condition yields two solutions, of which only the positive root fulfills the positivity constraint $\boldsymbol{\gamma} > 0$, and only if $\alpha \geqslant 2.5$. In this case, the solution is given by:

$$\gamma_{\mathrm{CMAP},i} = \frac{\beta}{2}\left(\eta + \sqrt{\eta^2 + \frac{2\|s_{i*}\|^2}{\beta}} \right) \tag{4.6}$$

For the inverse gamma hyperprior (i.e., "−" in (4.3)), a similar computation shows that all $\alpha \geqslant 0$ fulfill the second order condition and that the update rule is given by:

$$\gamma_{\mathrm{CMAP},i} = \frac{\frac{1}{2}\|s_{i*}\|^2 + \beta}{\kappa}, \quad \text{with} \quad \kappa = \alpha + 3/2 \tag{4.7}$$

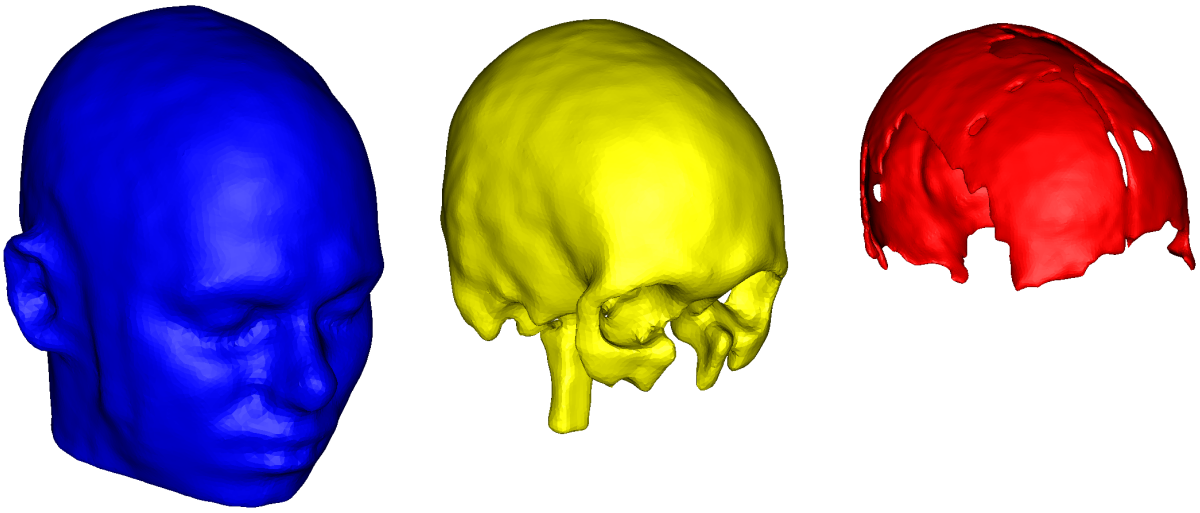
In principle, Os and Oγ steps within the AO scheme can now be combined explicitly, resulting in a fixed point iteration for s. This sheds some light on the motivation for introducing the cyclic AO scheme in combination with this special HBM: It turns out that through varying α, β and ζ, a variety of well known optimization schemes for regularization based approaches to CDR can be assessed naturally, i.e., by applying a simple procedure for MAP approximation for one common generative model. This is discussed in the appendix in Section A.1.8.
Concerning the Sγ step for the inverse gamma hyperprior, the conditional distribution $p_{post}(\gamma_i, s|b)$ given by (4.3) can be rearranged to:

$$p_{post}(\gamma_i, s|b) \propto \exp\left(-\frac{\frac{1}{2}\|s_{i*}\|^2 + \beta}{\gamma_i} + (-(\alpha + 3/2) - 1)\ln(\gamma_i) \right) \tag{4.8}$$

This is also an inverse gamma distribution, with parameters $\bar{\beta} = \frac{1}{2}\|s_{i*}\|^2 + \beta$ and $\bar{\alpha} = (\alpha + 3/2)$ (cf. (4.1)). This invariance property is called *conditional conjugacy* and simplifies the sampling scheme considerably. In principle, even the implicit prior on S, i.e., $p(s)$ (which is given by 2.5) can be computed explicitly. This is also done in the appendix, see Section A.1.9.

Remarks: (from Nummenmaa et al., 2007a)

- ⋆ Note that the currents are only assumed to be independent *a-priori*, and not *a-posteriori*.
- ⋆ Even though the assumptions of a-priori independent (implying also uncorrelated) sources, stationary noise distribution, and a data driven characterization of the source covariance resemble seemingly those of local, spatial scanning methods and beamforming, this hierarchical approach is a global, source space based method (cf. Section 1.3.1). That means all currents (and other parameters) are estimated simultaneously, rather than using a spatial filter methodology and projecting the data to each source point separately.



(a) Skin (b) Skull compacta (c) Skull spongiosa

Figure 4.2: Surfaces used for head model generation.

4.3 General Setting for the Studies

4.3.1 Head model

The head model we will use for our studies was created in the following way: T1 and T2 weighted MR images of a healthy proband were taken, and the T2 image was registered onto the T1 image (see Figure A.7 in the appendix). Different tissues were segmented and high resolution surface meshes were created from the voxel-based segmentation volumes. The surfaces were smoothed using Taubin smoothing (Taubin, 1995) to remove the blocky structure which results from the fine sampling of the voxels. For the aims of our specific studies only the surfaces of skin, eyes, skull compacta and skull spongiosa (see Figure 4.2) were used to create a high quality 3D Delaunay triangulation via TetGen[2]. Within both skull compartments, a higher mesh resolution is used. In total, the model consists of 512 394 FEM nodes and 3 176 162 tetrahedra (see Figure 4.3). The electrical conductivities of the different tissues are listed in Table 4.1.
The reason for using this model instead of a model including the inner brain compartments like gray matter and white matter is that we want to focus on the effect of depth bias separate from others, e.g., from the effects caused by the anisotropy of the white matter (which also makes the results comparable to those obtained using BEM models, which cannot capture the anisotropy). In addition, to facilitate the interpretation of the results, we need a homogeneous innermost compartment without holes and enclosures where we can place the test sources. Another important aspect for practical EEG/MEG studies is the effect of *insufficient sensor coverage*: For an optimal scan of the electromagnetic field pattern, the sensors should be placed uniformly distributed in every spatial direction. However, for practical reasons, this is not possible in realistic settings: The neck causes a semi shell like sensor distribution which is not able to record fields in the direction of the feet. Especially deep lying sources suffer from this insufficiency. The influence of insufficient sensor coverage should not be mixed with the effects of depth bias. Therefore we will use an artificial sensor configuration consisting of 134 EEG sensors distributed uniformly over the surface of the head model (see Figure A.8 on page XVI in the appendix).
Within the inner compartment, a source space consisting of 1 000 FEM nodes based on a regular grid is chosen, the grid size is 10.986 mm (for details and illustration, see Figures A.11 on page XVIII in the appendix). At each node, $d = 3$ orthogonal dipoles are placed, and the

[2]TetGen: A Quality Tetrahedral Mesh Generator and a 3D Delaunay Triangulator. http://tetgen.berlios.de/

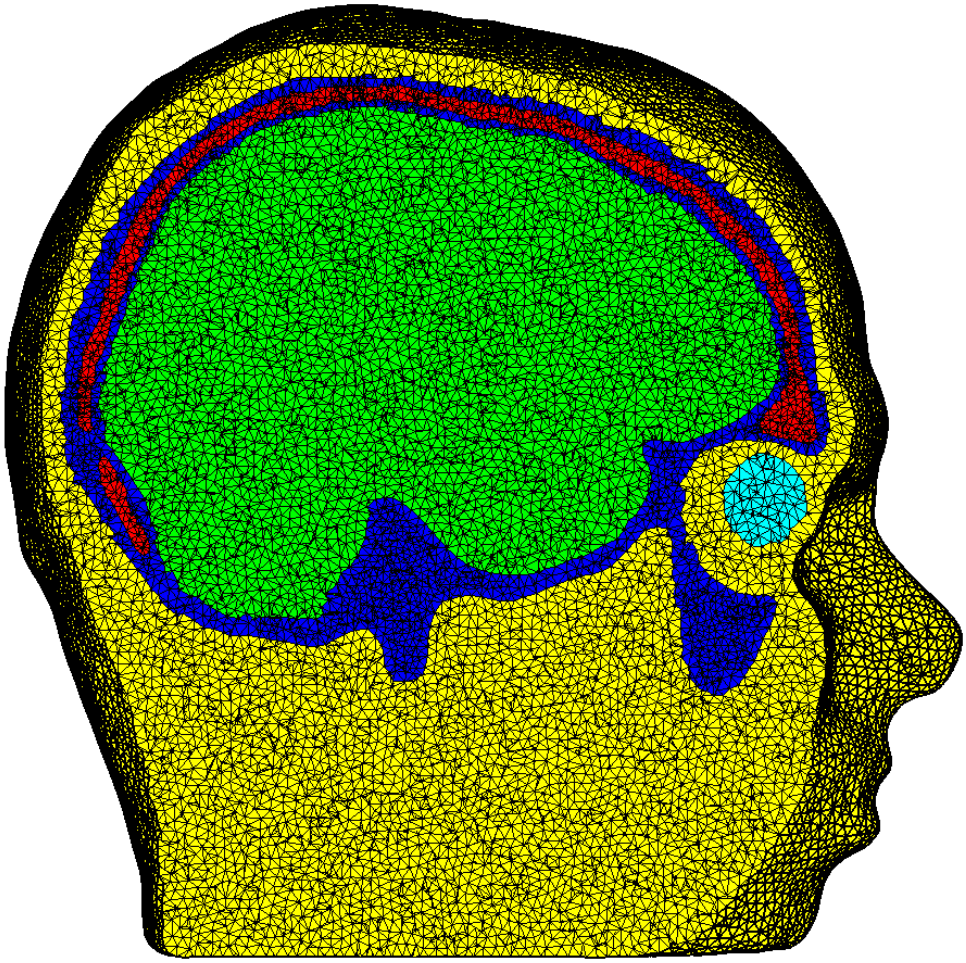


Figure 4.3: 5-compartment realistic head model used for the forward computation.

corresponding lead-field matrix is computed with SimBio[3] (with linear basis functions and the *Venant* approach for dipole modeling). In Figure A.9 in the appendix on page XVII, the sum of the ℓ_2-norms of the three gain-vectors is depicted. The complete model generation pipeline is depicted in Figure 4.4.

4.3.2 Inverse Methods

We will use the following methods for our studies:

- ⋆ CM approximation via the AS_CM algorithm as described in Section 3.3 for the HBM introduced in Section 4.2 with hyperpriors of the inverse gamma type. Parameters: α, β, Q, R.
- ⋆ MAP approximation via the uAO_MAP algorithm as described in Section 3.3 for the HBM introduced in Section 4.2 with hyperpriors of the gamma and inverse gamma type. Parameters: α, β, T. The convergence criterion for the CGLS iterations used in the Os steps will be that the relative residual of the normal equations falls below 10^{-6} (cf. Section 3.5).
- ⋆ MAP approximation via the cmAO_MAP algorithm as described in Section 3.3 for the HBM introduced in Section 4.2 with hyperpriors of the inverse gamma type. Parameters:

[3]For information on SimBio, see: https://www.mrt.uni-jena.de/simbio/index.php/Main_Page

Table 4.1: Isotropic tissue conductivities used for the different compartments.

Head tissue	Conductivity (S/m)
Skin	0.43
Eyes	0.505
Skull compacta	0.0064
Skull spongiosa	0.02865
Brain	0.33

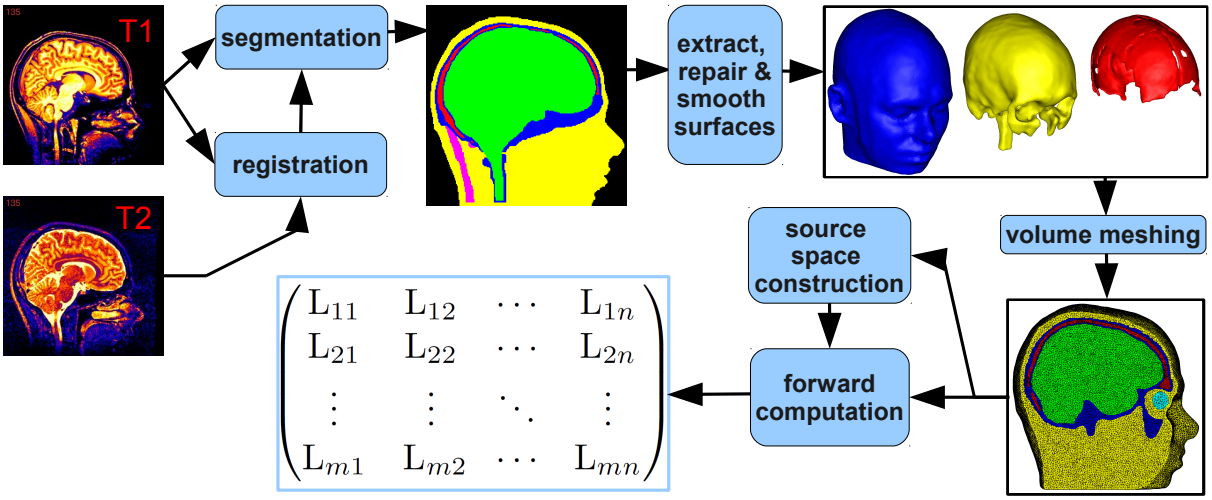


Figure 4.4: Model generation pipeline.

α, β, Q, R, T. The convergence criterion for the CGLS iterations used in the Os and Ss steps will be that the relative residual of the normal equations falls below 10^{-6}.

⋆ MAP approximation via the McmAO_MAP algorithm as described in Section 3.3 for the HBM introduced in Section 4.2 with hyperpriors of the inverse gamma type. Parameters: α, β, U, Q, R, T. The convergence criterion for the CGLS iterations used in the Os and Ss steps will be that the relative residual of the normal equations falls below 10^{-6}.

⋆ MNE as described in Section 1.3.2. Parameter: Regularization parameter λ.

⋆ WMNE as described in Section 1.3.2 with ℓ_2 weighting (Fuchs et al., 1999) and regularized ℓ_∞ weighting (Fuchs et al., 1999):

$$\Sigma_s^{\ell_2} = \operatorname*{diag}_{i=1,\ldots,n} \left((\|L_{(\cdot,i)}\|_2^2)^{-1} \right);$$

$$\Sigma_s^{\ell_\infty,reg} = \operatorname*{diag}_{i=1,\ldots,n} \left(\frac{\chi_i^2}{(\chi_i^2+\beta^2)^2} \right), \quad \text{with} \quad \chi_i = \|L_{(\cdot,i)}\|_\infty; \quad \beta = \max(\chi) \cdot \frac{m\,\sigma^2}{\|b\|_2^2}$$

Parameter: Regularization parameter λ.

⋆ sLORETA as described in Section 1.3.2. Parameter: Regularization parameter λ.

To get an initial visual impression of the different methods, their results for a single dipole source are depicted in Figures A.12 - A.20. The parameter setting used for these reconstructions is the result of the following sections where we will discuss the choice of the parameters, and the effects of adding measurement noise.

Table 4.2: Computation time (sec) for one AS_CM or uAO_MAP computation for different assumed noise levels x.

	100%	**10%**	**1%**	**0.1%**	**0.01%**	**0.001%**
AS_CM	95.41	217.23	397.73	841.92	1182.68	1604.16
uAO_MAP	0.13	0.26	0.60	1.26	1.07	1.28

Table 4.3: Influence of real and assumed noise level (nl) on MNE: For different real underlying noise level, different noise level for the choice of λ based on the discrepancy principle are assumed (but never less than the real noise level). The resulting mean DLE for 10 000 randomly placed dipoles is depicted (in mm).

	assumed noise levels →				
real noise level ↓	**0%**	**2.5%**	**5%**	**7.5%**	**10%**
0.0%	24.64	28.91	30.32	31.80	33.95
2.5%		28.57	30.10	31.62	33.76
5.0%			29.57	31.11	33.20
7.5%				30.50	32.22
10.0%					31.38

4.4 Preliminary Examinations

4.4.1 The Influence of Noise and the Noiseless Case

So far, our setting aims at separating the effects of depth bias from all other potentially arising effects that influence the estimation process. Thus we could try to avoid the effects of adding measurement noise as well, and add no noise to the simulated data, i.e., the *noiseless case*. On the other hand, this poses some methodical problems that we will discuss in this section.
First of all, a definition of *noise level* used in the following is fixed, as there is no commonly accepted one in the literature: In line with Calvetti et al. (2009) we will speak of a (relative) noise level of x if the standard deviation of the measurement noise (i.e., σ in our notation) fulfills $\sigma = x \cdot \|b_0\|_\infty$, where b_0 are the measurements in the noiseless case.
For HBMs, the noise variance σ^2 needs to be specified in some way (cf. (4.2)), even in the absence of real noise. Otherwise the likelihood will become singular, and the numerical implementation breaks down. One could hence choose a very small value of σ^2, but that increases the computation time dramatically, as depicted in Table 4.2. In addition, assuming noise enters a regularization term into the scheme (via the likelihood) that is needed due to the ill-condition of the problem, regardless whether the data really contains noise or not. These are the reasons, why Calvetti et al. (2009) do not actually add noise, but assume that σ is 5% of the maximum of the noiseless signal. Still, for our purpose and the error measures we use, this approach may discriminate the other inverse methods[4]: In Table 4.3 the mean DLE for the MNE of 10 000 randomly placed dipoles is depicted for different combinations of assumed and actually added noise. The regularization parameter is chosen according to the discrepancy principle (see Section 4.4.3) based on the assumed noise level. For a fixed assumed noise level, the DLE slightly decreases with increasing real noise level, which is a quite counterintuitive result. This phenomena was even more pronounced for a different head model which was used at the beginning of our studies (but was then replaced by the new head model described in Section 4.3.1). It occurs for all ℓ_2-norm based linear estimators, and for the uAO_MAP approximations. Its origin is related to realistic head modeling: In summary, realistic head modeling results in source locations with gain vectors that are "more unique" compared to the other vectors than it is the case in sim-

[4] Calvetti et al. (2009) do not consider other inverse methods in their studies

plified (more symmetric and homogeneous) volume conductors. Certain inverse methods prefer to place the maximal amplitude of their estimates on these locations. As they are often found in boundary areas of the source space, this tendency leads to large DLEs. Adding noise "cures" this problem to a certain extend. This issue will not be pursued any further within this thesis. Facing these problems, we will not consider the noiseless case in our studies at all, but always add noise to our data (at noise levels of 5% or 10%). Although it would be preferable to separate the effects of noise from those of depth bias, our considerations show that it might produce unwanted phenomena when using other inverse methods than HBM based types. This will complicate the comparison between the different methods, which is one of the central aims of our studies and is far more important than getting rid of the effects of sensor noise. The noiseless case is of theoretical interest, but omitting it does not weaken the value of the studies with regard to realistic applications, where noise is not avoidable (EEG/MEG recordings often suffer from rather low SNRs).

4.4.2 The Choice of the Parameters of the HBM Based Methods

In this section, we will discuss the general behavior of our CM and MAP approximation schemes for different settings of the parameters of the hyperprior, α and β (cf. (4.1)) and fix the internal parameters of the methods that have not been set yet. Note that from the Bayesian modeling paradigms, the choice of the hyperprior parameters should, in principle, not depend on the method we use for inversion (cf. Section 2.1), but only rely on our a-priori information about the value of the hyperparameters. Remember that they define the distributions of the γ_i which on their part determine the typical scale and spread of the source amplitudes. As the source amplitudes represent the macroscopic net current flow in the vicinity of the source location (cf. Section 1.2) the choice of α and β should ideally include our physiological a-priori knowledge about typical source current amplitudes found in the gray matter volume. Furthermore, it is in particular improper to use, e.g., different parameters for MAP and CM estimation. Nevertheless, since we do not have sufficient a-priori knowledge, and our main interest lies in the practical value of these new methods, we will commit this abuse (Sato et al., 2004; Nummenmaa et al., 2007a,b; Calvetti et al., 2009 all choose them ad hoc, as well).

Inverse Gamma Hyperprior: We will start with the inverse gamma hyperprior, which can be seen as a canonical choice for a hyperprior, and was thus used in many more studies until now[5]. As examined in Nummenmaa et al. (2007a,b), the choice of parameters is not a trivial issue, since the full posterior distribution is *multimodal*, i.e., it has multiple local maxima (modes). In the following we will illustrate this phenomena indirectly and sketch the consequences for our approximation methods. It is beyond the scope of this thesis to examine it deeply, especially since we would need to introduce a lot of technical terms and tools.
The multimodality is a result of the non-convexity of the energy of the inverse gamma hyperprior, i.e., the negative natural logarithm of its density function ($f_i(\cdot)$ in our notation, cf. (2.7)). This is discussed in Section A.1.7 in the appendix. The multimodality is always present to some extend, however, the concrete choice of the parameters and the interplay with the under-determinedness of the linear equation system determine whether it affects the estimation process practically. To understand this, it is crucial to distinguish between the theoretical and practical consequences of parameter changes for MAP and CM estimation and approximation when dealing with a multimodal distribution.
First, we use a simple toy model to sketch the theoretical consequences for MAP and CM estimation. Let $p(x;\theta)$ be a probability distribution on $\mathbb{R}^N$ that depends continuously on a parameter $\theta \in \Theta \subset \mathbb{R}^K$. Given that the posterior mean exists for all $\theta \in \Theta$, the CM estimate $\hat{x}_{\mathrm{CM}}$ by its very nature is a continuous mapping from $\{p(\,\cdot\,;\,\theta) \mid \theta \in \Theta\}$ to $\mathbb{R}^N$, and therefore

[5] I am actually only aware of Calvetti et al. (2009) using the gamma hyperprior.

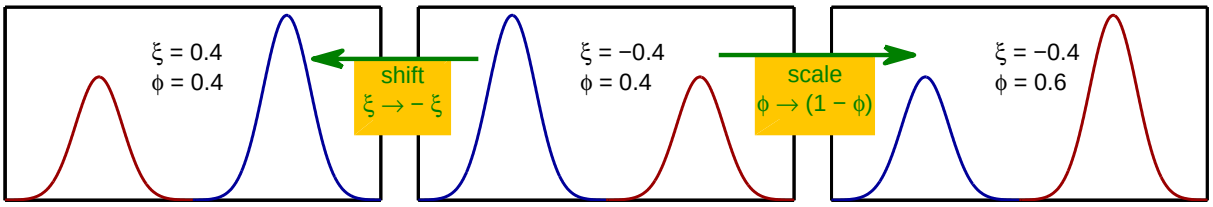


Figure 4.5: The impact of parameter changes for the toy model. From middle to left image: A change of ξ causes the modes of p_1 (red) and p_2 (blue) to switch position but does not affect the heights. The MAP estimate will hence follow this shift and changes continuously. From middle to right image: A change of ϕ will change the relative heights of the modes without changing their position. At $\phi = 0.5$ the MAP estimate will discontinuously jump from the blue to the red mode.

gives rise to a continuous mapping $\hat{x}_{\mathrm{CM}}(\theta)$ from Θ to $\mathbb{R}^N$: $\hat{x}_{\mathrm{CM}}(\theta) = \hat{x}_{\mathrm{CM}}(p(\,\cdot\,;\,\theta))$. If $p(x;\theta)$ is multimodal for some $\theta \in \Theta$ this does not affect $\hat{x}_{\mathrm{CM}}(\theta)$ theoretically. Practically, MCMC-based CM approximation methods may fail to reach every mode within a reasonable sample size, i.e., the chain is practically *reducible*.
Now let $\hat{x}_{\mathrm{MAP}}(\theta)$ be the mapping from Θ to $\mathbb{R}^N$ given by $\hat{x}_{\mathrm{MAP}}(\theta) = \hat{x}_{\mathrm{MAP}}(p(\,\cdot\,;\,\theta))$. The mapping $\hat{x}_{\mathrm{MAP}}(\theta)$ is very sensitive to certain changes of the modes of $p(\,\cdot\,;\theta)$ due to the discontinuous nature of the argmax function: A change in the location of the modes is not problematic, but a change of the relative heights of different modes of the distribution may cause a sudden jump of $\hat{x}_{\mathrm{MAP}}(\theta)$. To illustrate this and to prepare our further examinations, we consider a simple toy example of a bimodal distribution: For $\theta = (\phi, \xi) \in [0,1] \times [-1,1]$, let $p(x;\phi,\xi)$ be a mixture of two Gaussians with equal variances $\nu^2 = 0.05$ but mirrored means:

$$p(x;\phi,\xi) = (1-\phi)\cdot p_1(x;\xi) + \phi \cdot p_2(x;\xi)$$

with

$$p_1(x;\xi) = \left(2\pi\nu^2\right)^{-1/2} \exp\left(-\frac{1}{2\nu^2}\left(x - \tfrac{\xi}{2}\right)^2\right)$$
$$p_2(x;\xi) = \left(2\pi\nu^2\right)^{-1/2} \exp\left(-\frac{1}{2\nu^2}\left(x + \tfrac{\xi}{2}\right)^2\right)$$

Now a change in ξ causes a shift of the modes, whereas a change in ϕ scales the relative heights of the modes. In Figure 4.5 the different impact of these changes is illustrated. In Figure 4.6 the values of $\hat{x}_{\mathrm{MAP}}(\theta)$ and $\hat{x}_{\mathrm{CM}}(\theta)$ are plotted as a function of ξ and ϕ. The discontinuity of the MAP estimate at $\phi = 0.5$ is a clearly visible hint to the multimodality. We will exploit this phenomenon to indirectly illustrate the multimodality of the posterior (4.2) when using an inverse gamma hyperprior: The true number and location of the modes of $p_{post}(s,\boldsymbol{\gamma}|b)$ is hard to detect, but as we have seen, varying the parameters of the hyperprior can change height and location of the modes. A MAP approximation scheme may thus end up in different modes, dependent on the parameter values (not only due to a change of the real MAP estimate, but also due to a change of the mode that attracts a locally convergent MAP approximation scheme like, e.g., AO_MAP). For a continuous scalar property computed from a MAP approximation and plotted for different parameter values as in Figure 4.6, edges mark parameter sets, across which the MAP approximation jumps from one mode to another, and therefore indicate that the distribution is in fact multimodal. Still, this will only give a vague impression of the complexity of the posterior, as only those modes that are locally attractive to the approximation scheme for some set of parameters will be revealed.
For a single dipole (see Figure A.10 in the appendix) the uAO_MAP result has been computed for different values of α and β, using $T = 100$. The measures spatial dispersion, relative residual

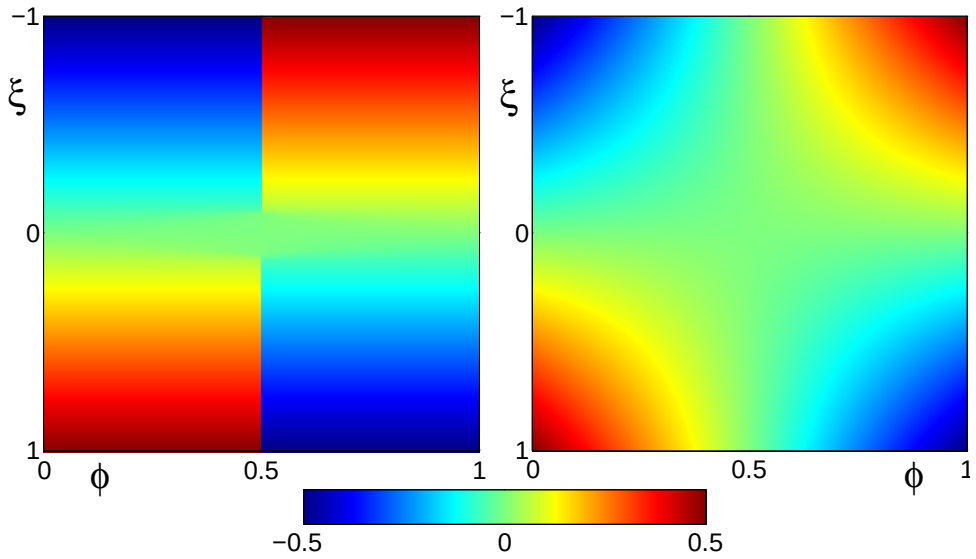


Figure 4.6: The impact of parameter changes for MAP estimate (left) and CM estimate (right).

$\|b - \mathrm{L}s_{\mathrm{uAO\_MAP}}\|/\|b\|$, DLE, 8^{th}-COME , 1^{st}-COME and EMD were computed for the resulting MAP approximation. Figure 4.7 shows the results. From the spatial dispersion result we can see that for a fixed b, the HBM proposed here comprises focal MAP approximations as well as spatially smoother ones sharply separated in the parameter space. This separating line marks the region where the influence of the hyperprior in the posterior (4.2) (which prefers focal solutions) overcomes the influence of the likelihood (which prefers smooth solutions). In the region below the line, the non-convexity of the hyperpriors energy (cf. Section A.1.8) causing the multimodality becomes visible: The continuous measures relative residual, COME and EMD show a clear fragmentation into different areas representing single modes of the posterior. Within these areas, the measures look quite smooth again. Facing this complex variety of the characteristic features of the MAP approximation by the uAO_MAP algorithm for different parameter choices even for one single dipole source, the task of finding an optimal parameter set for a whole study with different single dipole sources and for subsequent studies with more complex source pattern seems rather hopeless, or in gentle words, is beyond the scope of this thesis. For our purpose, the expectation of the outcome of such an examination is very limited: The CM approximation via AS_CM with a parameter set found by visual inspection outperforms every score achieved by a uAO_MAP approximation encountered within the search over all 156 791 parameter sets that were sampled to attain the above results[6]. The MAP approximations that are based on CM approximation, i.e., cmAO_MAP and McmAO_MAP, even improve up on the results achieved by AS_CM. Calvetti et al. (2009) use $\alpha = 1.55$ and $\beta = 10^{-7}$ for both gamma and inverse gamma hyperprior without further justification. However, we cannot simply take over the value of α as we use a slightly different parameterization. Unfortunately, no other publication we are aware of examines Full-MAP estimation for this HBM in EEG/MEG.

To proceed nevertheless, we fix T to 50 iterations and the values of α and β will be fixed by averaging the EMD for 1 000 single unit-strength dipole sources for a set of possible parameters. The range of the parameters within this set has been chosen on the basis of preliminary studies with smaller sample sizes on a wider range of parameters. The choice of the EMD over the more common DLE will be justified in Section 4.7. In summary, the EMD is the best single measure that captures both our modeling assumptions of focal source activity and right localization. In Table 4.4 the results are depicted. From this we choose $\alpha = 0.5$ and $\beta = 5 \cdot 10^{-6}$. We did not test smaller values of α, since other findings suggest that the uAO_MAP scheme gets unstable

[6]CM result to best uAO_MAP result: 4.45 to 23.74 in DLE; 5.27 to 31.70 in EMD; 0.0048 to 0.0086 in relative residuum; 3.11 to 4.24 in 1^{st}-COME.

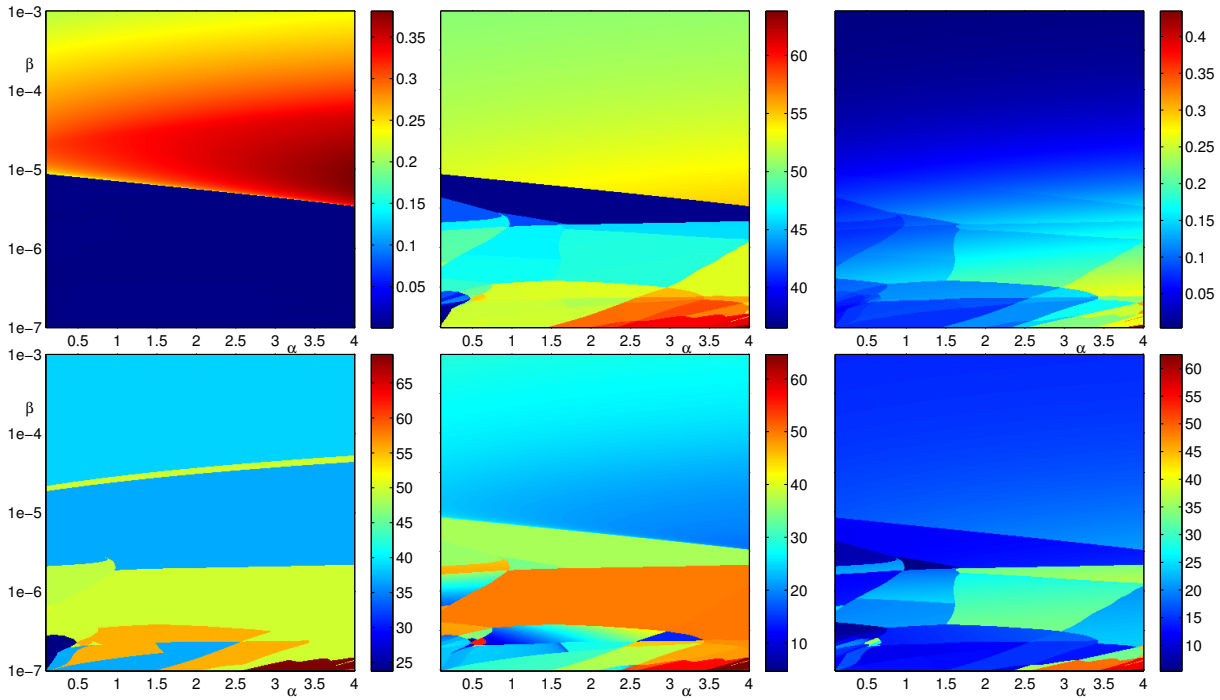


Figure 4.7: Features of the uAO_MAP result with an inverse gamma hyperprior for a single dipole source using different hyperprior parameter sets (α, β). Upper row from left to right: SD, EMD, relative Residuum. Bottom row from left to right: DLE, 8^{st}-COME, 1^{th}-COME.

Table 4.4: EMD of IAS result for different parameters of the inverse gamma hyperprior, averaged over 1 000 single unit-strength dipole sources.

	$\alpha \rightarrow$								
$\beta \downarrow$	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**	**1.0**	**1.2**	**1,5**	**2.0**
10^{-4}	38.08	38.31	38.48	38.58	38.73	38.81	39.19	39.80	40.75
$5 \cdot 10^{-5}$	34.17	34.31	34.46	34.69	34.92	35.03	35.28	35.78	36.75
10^{-5}	28.82	28.94	29.04	29.17	29.22	29.35	29.66	30.08	31.21
$5 \cdot 10^{-6}$	28.18	28.28	28.36	28.49	28.66	28.88	29.29	29.98	31.67
10^{-6}	32.90	33.50	33.90	34.44	35.03	35.70	37.34	39.36	41.54

in these regions. This might not be visible in the averaged values and might therefore be overlooked. Furthermore, we will use the same value for α for all other methods based on the HBM with an inverse gamma hyperprior which will ease the comparison.

As discussed above, concerning CM approximation, the multimodality of the posterior should only affect the practical performance of the approximation. It would thus be interesting to repeat the above examinations and see how Figure 4.7 looks like for AS_CM and based on this, for cmAO_MAP and McmAO_MAP. Unfortunately, the computation time for AS_CM is not only way larger than for uAO_MAP by default, it even increases considerably for larger values of β. The parameters which we will use later (i.e., $\alpha = 0.5$, $\beta = 5 \cdot 10^{-8}$) yield the best results while exhibiting moderate computation times. Still we decided to omit the comparison in this thesis. The concrete parameters used for AS_CM for our studies are chosen in the following order: First, the burn-in size is fixed to $Q = 1\,000$ for all of our studies. This is certainly sufficient, since the chain is reported to have a rapid mixing (Nummenmaa et al., 2007a). Furthermore, Calvetti et al. (2009); Nummenmaa et al. (2007a) do not report to use burn-in steps at all. Second, the real sampling steps R used will be 10 000 for preliminary studies, 1 000, 5 000 and 50 000 for the first study, and 5 000, 50 000 and 200 000 for the second. In addition, we will use the

Table 4.5: EMD of the AS_CM result for different parameters of the inverse gamma hyperprior, averaged over 100 single unit-strength dipole sources.

	$\alpha \rightarrow$			
$\beta \downarrow$	**0.50**	**0.90**	**1.25**	**2.00**
10^{-6}	23.39	9.10	14.62	24.22
$5 \cdot 10^{-7}$	15.24	10.50	16.52	26.72
10^{-7}	7.97	14.46	22.59	30.88
$5 \cdot 10^{-8}$	7.68	16.17	24.75	32.70
10^{-8}	8.27	21.70	29.90	38.62

blocked inversion scheme for all studies, which means that we solve the WLS-problem in the Ss step with the CGLS algorithm (cf. Section 3.4.1). The values of α and β will be chosen similar as for the uAO_MAP method, except that only 100 single unit-strength dipole sources are used for testing. In Table 4.5 the results are depicted. From this, we choose $\alpha = 0.5$ and $\beta = 5 \cdot 10^{-8}$ for our further studies. The values of α and β for cmAO_MAP and McmAO_MAP are set to those used for AS_CM. This choice is not only to reduce the computational burden, but also relies on the observation that both methods mainly refine the result given by AS_CM. For both methods T is set to 50. For cmAO_MAP, Q and R are also set to the same values as for AS_CM. For McmAO_MAP the choice of U, Q and R is more complicated: The intention of the McmAO_MAP method is to have a number of different starting points for the cmAO_MAP scheme in such a way that different modes are found and can be compared. If R is chosen very large, the convergence of the AS_CM scheme will cause the starting points to be almost equal, and hence, the subsequent AO_MAP scheme should end up in the same mode for all seeds. On the other hand, if R is too small, the initial $\boldsymbol{\gamma}_0$'s will be too "non-sparse" and the modes found by the AO_MAP scheme starting at these seeds will be quite similar to the one found by the uAO_MAP method. For U the situation is more simple: The performance can only increase by an increase of U. However, that will of course increase the computational burden. For these reasons, a preliminary study is carried out: 100 single unit-strength source dipoles are reconstructed using different combinations of U, Q and R. For each dipole, a ranking of the methods is computed by comparing the (rounded) probabilities of the MAP approximations found by the different methods. The method that found the approximation with the highest probability is ranked at the first place. Methods that found an approximation with the same probability are ranked at the same place. Subsequently the mean rank of each method is computed over all 100 dipoles. The results are listed in Table 4.6. From this we choose $U = 64$, $Q = 25$ and $R = 200$.

Gamma Hyperprior: For the gamma hyperprior, the situation for MAP estimation and approximation is way less complicated, since the energy of the hyperprior is always convex (cf. Section A.1.7), and so is the complete objective function (4.2). Thus we have a unique minimizer and the uAO_MAP algorithm should converge globally to the MAP estimate. It is shown in the appendix in Section A.1.8 that the uAO_MAP algorithm in fact turns out to work like a fixed point scheme to minimize a functional involving a relaxed ℓ_1-norm penalty on the source amplitudes. Just for comparison and illustration, the same examinations as in the last paragraph are carried out for the gamma hyperprior, and the results are shown in Figure 4.8. The comparison of the plots for the relative residual shows in particular the difference of a convex objective function to a non-convex one. The values of α and β are also chosen by a preliminary study, Table 4.7 shows the corresponding results. The results indicate to use $\alpha = 2.6$ and $\beta = 10^{-5}$.

Table 4.6: Mean ranking of different McmAO_MAP methods.

	$M = 5$	$R = 10$	$R = 25$	$R = 50$	$R = 100$	$R = 200$
$U = 4,\ Q = 5$	16.41	16.04	14.88	13.44	11.46	8.16
$U = 4,\ Q = 10$		16.00	15.54	11.97	9.95	7.79
$U = 4,\ Q = 25$			13.39	12.06	9.80	7.70
$U = 4,\ Q = 50$				10.95	9.56	7.73
$U = 16,\ Q = 5$	9.25	9.31	6.88	5.22	4.43	2.59
$U = 16,\ Q = 10$		8.50	6.57	4.91	3.91	2.77
$U = 16,\ Q = 25$			5.43	4.47	3.65	2.14
$U = 16,\ Q = 50$				3.96	3.12	2.40
$U = 32,\ Q = 5$	5.91	5.47	3.88	2.79	2.55	1.39
$U = 32,\ Q = 10$		5.11	4.19	2.93	2.28	1.45
$U = 32,\ Q = 25$			3.09	2.45	2.09	1.41
$U = 32,\ Q = 50$				2.12	1.67	1.49
$U = 64,\ Q = 5$	3.24	3.14	2.04	1.73	1.62	1.13
$U = 64,\ Q = 5$		3.16	2.13	1.71	1.39	1.17
$U = 64,\ Q = 25$			1.67	1.57	1.30	1.12
$U = 64,\ Q = 50$				1.38	1.21	1.19

Table 4.7: EMD of the uAO_MAP result for different parameters of the gamma hyperprior, averaged over 1 000 single unit-strength dipole sources.

	$\alpha \rightarrow$								
$\beta \downarrow$	**2.5**	**2.6**	**2.7**	**2.8**	**2.9**	**3**	**3.1**	**3.2**	**3.3**
10^{-4}	39.56	35.35	41.00	45.25	48.31	50.53	52.14	53.33	54.27
$5 \cdot 10^{-5}$	38.24	33.04	37.50	41.39	44.43	46.79	48.61	50.02	51.10
10^{-5}	36.24	30.95	32.68	35.47	38.20	40.64	42.71	44.44	45.88
$5 \cdot 10^{-6}$	35.80	31.14	31.69	33.81	36.32	38.79	41.01	42.94	44.56
10^{-6}	36.03	33.53	32.67	33.02	34.48	36.58	38.92	41.20	43.25

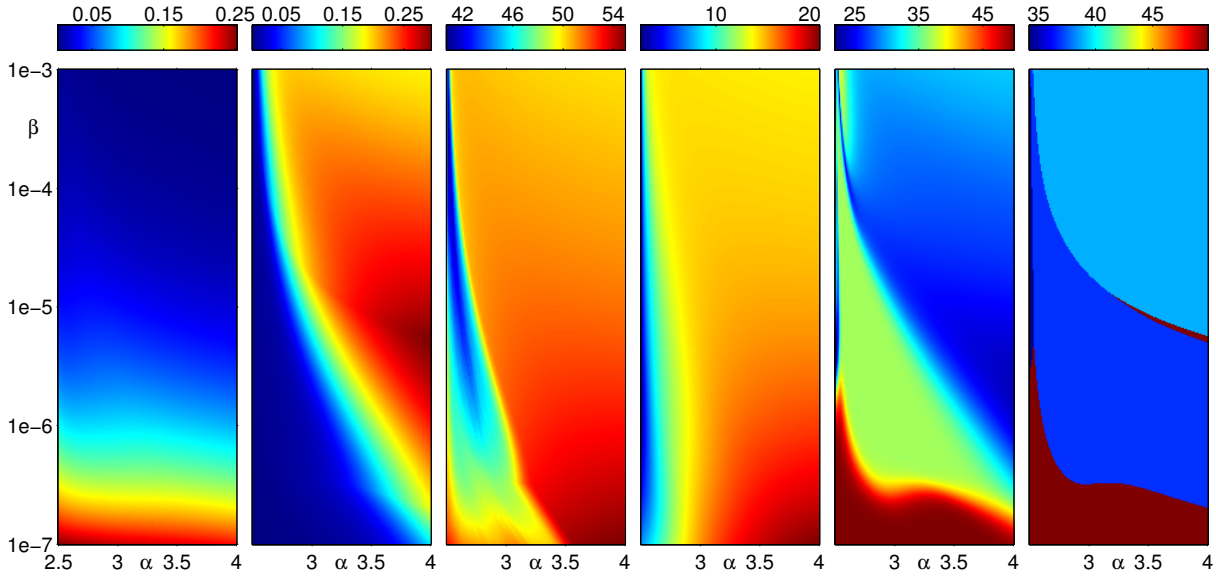


Figure 4.8: Features of the uAO_MAP approximation with a gamma hyperprior for a single unit-strength dipole source using different hyperprior parameter sets (α, β). From left to right: Rel. res., SD, EMD, 1^{st}-COME, 8^{th}-COME, DLE

Summary Parameters used in the following:

- ⋆ Hyperprior parameter:
  - ∗ uAO_MAP method with gamma hyperprior: $\alpha = 2.6$, $\beta = 10^{-5}$
  - ∗ uAO_MAP method with inverse gamma hyperprior: $\alpha = 0.5$, $\beta = 5 \cdot 10^{-6}$
  - ∗ AS_CM, cmAO_MAP and McmAO_MAP methods with inverse gamma hyperprior: $\alpha = 0.5$, $\beta = 5 \cdot 10^{-8}$
- ⋆ Algorithm parameter:
  - ∗ uAO_MAP: Number of iterations T: 50.
  - ∗ AS_CM: Burn-in steps Q: 1 000. Sample sizes R: 1 000, 5 000 and 50 000 for the first study, 5 000, 50 000 and 200 000 for the second.
  - ∗ cmAO_MAP: Burn-in steps Q: 1 000. Sample sizes R: 1 000, 5 000 and 50 000 for the first study, 200 000 for the second. Number of iterations T: 50.
  - ∗ McmAO_MAP: $U = 64$, $Q = 25$ and $R = 200$. Number of iterations T: 50.

4.4.3 The Choice of the Regularization Parameter

For MNE, WMNE and sLORETA, rules to choose the regularization parameters have to be determined. There are a variety of approaches, which can broadly be separated into approaches where the noise level is known, or a good estimate of it is available, and approaches where it is not known. In the toolbox, the *discrepancy principle* (e.g., Engl et al., 1996; Kaipio and Somersalo, 2005) as a member of the first class and the *cross-validation technique* (e.g., Pascual-Marqui, 1999a) as a member of the second class are implemented[7]. Since we assume to know the noise level we will apply the discrepancy principle. In addition, it is a simple and robust scheme allowing for an easy interpretation of the results without making further assumptions on the problem.

4.5 Study 1: Localization of Single Dipoles

4.5.1 Setting

In the following, the *depth* of a location within our model is defined as the minimal distance to one of the sensors. For the study, 750 single unit-strength source dipoles with random location and orientation were placed in the inner compartment (not necessarily on the source space nodes to avoid an obvious inverse crime, cf. Section 1.3.3). The following restriction on their depth was posed: First, the nearest sensor is searched. For that sensor, the nearest source space node is searched. The position for the dipole is only accepted if its depth is larger than the depth of the source space node plus 10 mm. This way, dipoles that are closer to the sensors than any source space node are avoided, which facilitates the interpretation of the results (for dipoles that are closer to the surface than any source space node, no depth bias can occur).
Measurement data is generated using the same forward computation procedure used for the lead-field generation, and noise at the noise levels of 5% and 10% is added.
The inverse methods listed in Section 4.3.2 are used to invert the data with the parameter setting discussed in the proceeding sections.

4.5.2 Results

General properties The mean distance from the current dipoles to the next source space node was 5.24 mm, which is the lower bound for DLE and EMD for all methods. Table 4.8 shows EMD, DLE and SD, averaged over all dipoles. Interestingly, for the WMNE with regularized ℓ_∞

[7] *L-curve* approaches did not show convincing results.

Table 4.8: Different validation measures averaged over 750 single unit-strength dipoles

Method	EMD		DLE		SD	
	5% nl	10% nl	5% nl	10% nl	5% nl	10% nl
AS_CM, $R = 1\,000$	10.61	13.02	9.10	10.82	1.08e-03	1.34e-03
AS_CM, $R = 5\,000$	8.65	11.35	7.40	9.14	1.25e-03	2.03e-03
AS_CM, $R = 50\,000$	7.27	10.21	6.24	7.54	1.24e-03	2.67e-03
cmAO_MAP, $R = 1\,000$	9.43	10.29	8.52	10.06	3.00e-04	1.05e-04
cmAO_MAP, $R = 5\,000$	7.11	8.43	6.73	8.39	2.61e-04	4.35e-05
cmAO_MAP, $R = 50\,000$	6.06	7.27	5.82	7.28	2.31e-04	8.41e-06
McmAO_MAP, $U = 64$	5.88	6.67	5.77	6.67	1.33e-05	8.62e-07
uAO_MAP (inv. gamma)	28.26	53.72	27.10	37.65	1.36e-02	2.92e-01
uAO_MAP (gamma)	31.07	35.74	27.19	32.65	8.73e-03	1.06e-02
MNE	53.22	55.93	29.55	30.63	2.36e-01	2.95e-01
WMNE ℓ_2	52.15	54.88	30.39	34.27	2.54e-01	3.12e-01
WMNE $\ell_{\infty,reg}$	49.53	52.41	29.37	25.04	2.17e-01	2.92e-01
sLORETA	40.55	44.92	6.02	7.33	1.86e-01	2.45e-01

Table 4.9: Mean ranking of different MAP approximation methods in the first study.

Method	5% nl	10% nl
cmAO_MAP, $R = 1\,000$	1.96	1.99
cmAO_MAP, $R = 5\,000$	1.79	1.68
cmAO_MAP, $R = 50\,000$	1.62	1.37
McmAO_MAP, $U = 64$	1.02	1.01

weighting adding noise again leads to a counterintuitive decrease of the DLE (cf. Section 4.4.1). This confirms that the effect of adding noise is a non trivial issue and should be examined in more detail.

MAP approximations We briefly compare the different MAP approximation methods concerning the posterior probability of their results. For this, only methods that rely on the same parameter set can be compared. This limits the comparison to the cmAO_MAP and the McmAO_MAP methods which is still interesting, since both showed the best performance concerning localization (cf. Table 4.8). In Table 4.9 the average rank within a ranking similar to the one performed in Section 4.4.2 (cf. Table 4.6) is depicted.

Depth bias There is no concrete definition of depth bias yet that we could use for a direct evaluation of the results. In general, in frequentist statistics, a statistical *bias* of an estimator measures whether the estimator produces systematic errors. If an estimator aims to estimate a scalar value, it could over- or underestimate it *on average*. The Bayesian analogue to this property is called *calibration* (Gelman, 2006). However, there are technical problems which complicate both the application of these concepts to our situation and the subsequent interpretation of the results. For instance, the parameter space of the real sources and the estimated sources are intentionally non-conforming in order to avoid an inverse crime. Therefore we will rely on a visual presentation of the results in this thesis. Figure 4.9 gives an explanation of the scatter plots we are using for that purpose: On the horizontal axis, the depth of the real source is plotted. On the vertical axis, the depth of the source space node with the largest source estimate amplitude is plotted. A mark within the area underneath the y = x line indicates that the dipole has been reconstructed too close to the surface, whereas a mark above the line indicates

the opposite. By q_{bl} we denote the percentage of marks below the y = x line. If a method shows a clear tendency to favor the lower area and q_{bl} is considerably below 0.5, it suffers from depth bias. A method performs well if its marks in this type of scatter plot are tightly distributed around the y = x line as this does usually not only indicate a localization in the right depth but also in total. Figures 4.10(a) - 4.11(c) show all data for every single method separately. Since we found that the addition of noise to the data has no systematic impact on the phenomenon of depth bias (although it affects the total performance, cf. Table 4.8) we omitted the plots for the noisy case for all methods, and solely demonstrate it for the cmAO_MAP method with $R =$ 50 000 in Figure 4.11(d).

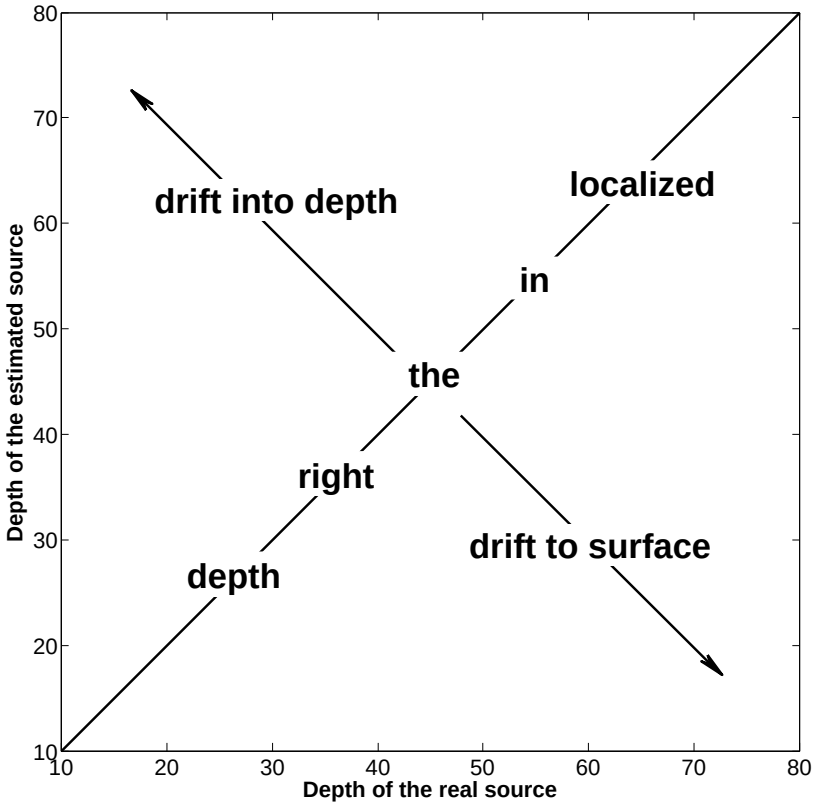
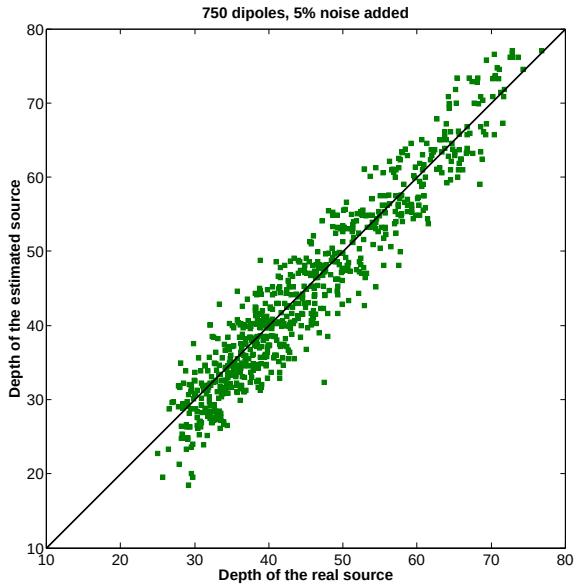
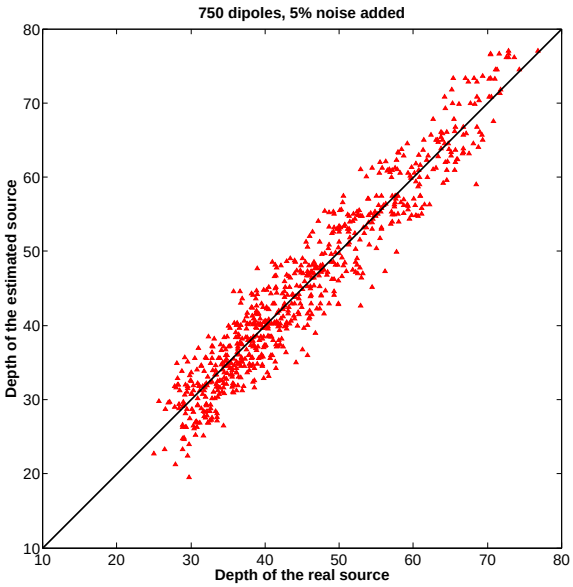


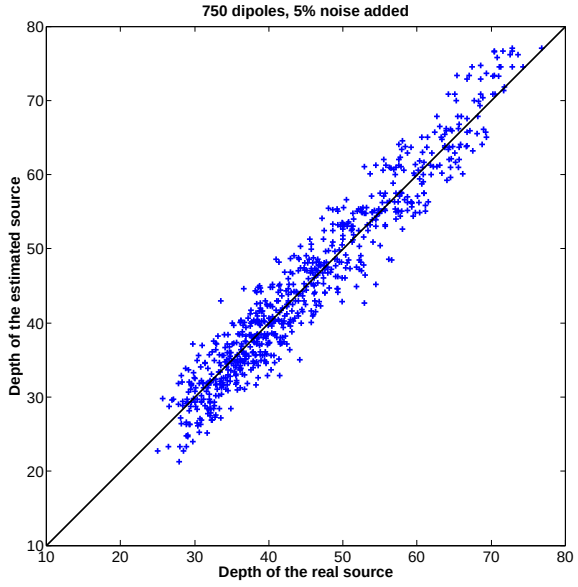
Figure 4.9: Explanation of the scatter plots



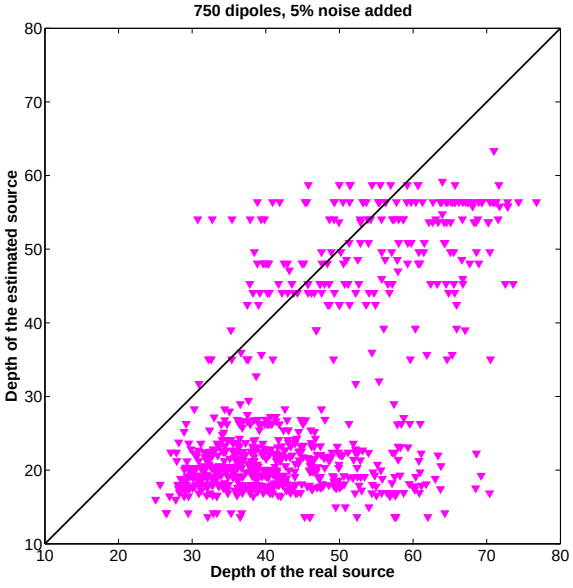
(a) AS_CM, $R = 50\,000$: $q_{bl} = 0.44$.



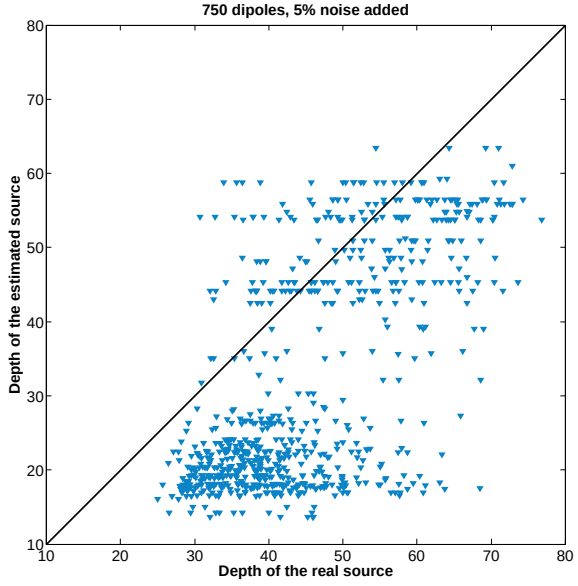
(b) cmAO_MAP, $R = 50\,000$: $q_{bl} = 0.49$.



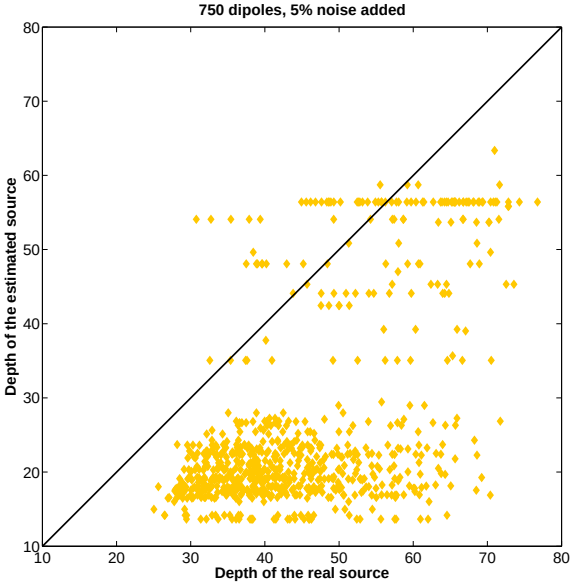
(c) McmAO_MAP, $U = 64$: $q_{bl} = 0.48$.



(d) uAO_MAP (inv. gamma): $q_{bl} = 0.09$.

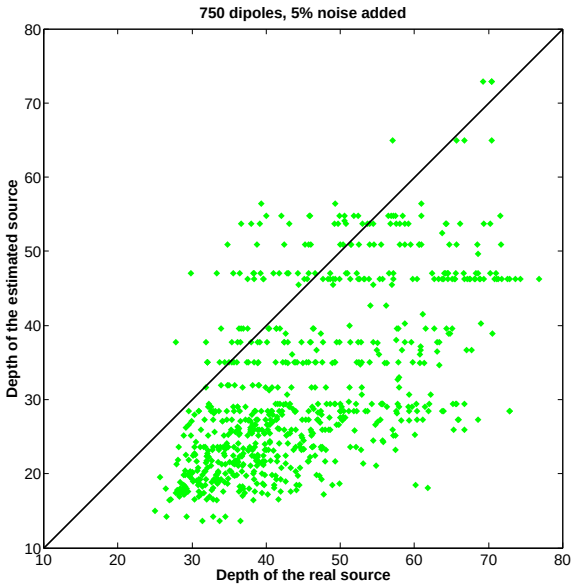


(e) uAO_MAP (gamma): $q_{bl} = 0.13$.

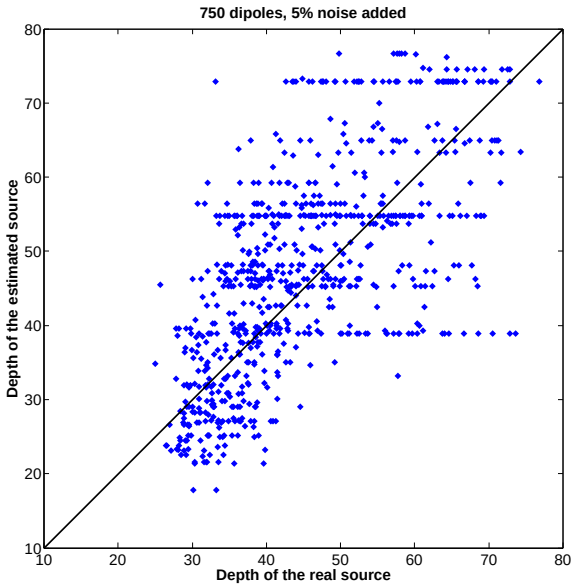


(f) MNS: $q_{bl} = 0.05$.

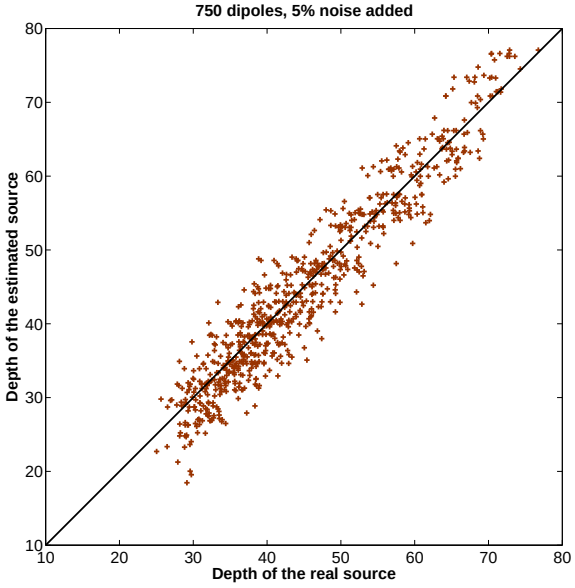
Figure 4.10: First Study: Scatter plots part 1



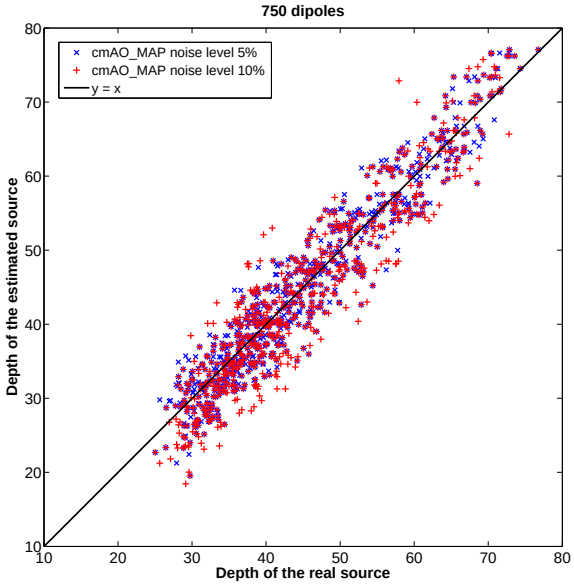
(a) WMNE-ℓ_2: $q_{bl} = 0.09$.



(b) WMNE-$\ell_{\infty,reg}$: $q_{bl} = 0.56$.



(c) sLORETA: $q_{bl} = 0.45$.



(d) cmAO_MAP, both noise level

Figure 4.11: First Study: Scatter plots part 2

4.5.3 Discussion

McmAO_MAP: The McmAO_MAP method showed the best results in EMD, DLE and SD (cf. Table 4.8). Compared to the second best method, i.e., the cmAO_MAP method with $R = 50\,000$, its computation is about three times faster. Compared to other MAP approximation schemes, it attains the highest posterior probability (cf. Table 4.11), which suggests that it should be seen as the best approximation to the real MAP estimate examined here. Furthermore, the method does not seem to suffer from depth bias (cf. Figures 4.10(c)).

AS_CM and cmAO_MAP: Both methods showed promising results for the specific source scenario examined here. They clearly benefit from increasing R (cf. Table 4.8), which of course also increases the computation time. Compared to each other, the cmAO_MAP result outperforms the AS_CM result for the same value of R. Since the additional computation time is

negligible this result suggests to always perform a subsequent AO run after an initial AS_CM scheme. Compared to established methods like MNE and sLORETA both methods clearly show better results concerning EMD and SD (cf. Table 4.8) and the visual impression is more convincing as well (cf. Figures A.12, A.13, A.17 and A.20). This even holds for small sizes of R. Concerning DLE, cmAO_MAP outperforms sLORETA for $R = 50\,000$, while for AS_CM it is to examine whether a further increase of R could accomplish this.
In addition they do not seem to suffer from depth bias (cf. Figures 4.10(a) -4.10(b)).
uAO_MAP: The uAO_MAP scheme with the inverse gamma hyperprior showed the worst results concerning localization of all methods based on the inverse gamma hyperprior (cf. Table 4.8). As described above, AO schemes based on other initialization rules showed much better results. Using the uAO_MAP scheme could then only be justified by the faster computation time. To see if this argument really holds, it would thus be interesting to examine how small the values of R, Q for the cmAO_MAP scheme or additionally U for the McmAO_MAP scheme can be chosen to improve upon the uAO_MAP scheme in a significant way. The gamma hyperprior combination with the uAO_MAP scheme attains similar results (cf. Table 4.8). Both methods seem to suffer from depth bias (cf. Figures 4.10(e) and 4.10(d)). However, for practical applications, both improve upon the MNE in for low noise level, especially with regard to the extent of the estimated source. Since they can also be computed very fast, they are a good alternative to the MNE if the a-priori information suggests a focal source configuration.
MNE and WMNE: The WMNE schemes used in this study are modifications of the original MNE explicitly aiming to improve the depth localization. Figures 4.10(f) - 4.11(b) clearly show that they succeed in this aspect. Concerning EMD, DLE and SD, the conclusion is less clear (cf. Table 4.8). The visualizations in Figures A.18 - A.20 do not yield a clear impression on the different characteristics of the estimates either. Hence more detailed examinations are needed.
sLORETA: The sLORETA estimate performs well concerning DLE and depth bias (cf. Table 4.8 and Figure 4.11(c)). Yet, Figure A.17 suggests that the sLORETA result overestimates the spatial extent of the source considerably. The average EMD and SP of sLORETA clearly confirm this (cf. Table 4.8).

It is important to stress that the above results were only attained for the specific source scenario examined in this study. Without further examinations, their significance might be very limited, since the ability to localize single dipoles is a rather trivial and largely uninformative property, as shown by Grave de Peralta et al. (2009). It has been overrated for long due to a fundamental misconception. Nevertheless, reconstructing single dipoles it is a starting test for every inverse method for CDR, and the results for the methods based on HBM clearly motivate to examine their use in more complex scenarios as well.

4.6 Study 2: Masking of Deep-lying Sources

4.6.1 Setting

The single dipoles that we used in the first study (see Section 4.5 for the constraint imposed on their locations) are now combined to form source configurations consisting of a deep-lying and a near-surface dipole: The dipoles are evenly divided into three parts by their depth (i.e., the minimal distance to one of the sensors). For each of the 250 source configurations used in the study, one dipole from the part with the largest, and one from the part with the smallest depth are randomly picked. Noise at a noise level of 5% is added to the measurements. The inverse methods listed in Section 4.3.2 are used to invert the data with the parameter setting discussed in the proceeding sections. In comparison to the first study, a larger number of samples R is used for AS_CM and cmAO_MAP (5 000, 50 000 and 200 000).

4.6.2 Results

Initial example The results for a source configuration for which the masking effect is very pronounced are shown (It was chosen by visual inspection after viewing the results for the first five source configurations of the study.). Figure A.21(a) shows the source configuration. In the left image, the left one of the sources is the near-surface one, located very close to the eye-nerve hole in the skull. The other source is located deep in the brain, distant to all sensors. Figures A.21(b) and A.21(c) show the MNE and sLORETA results. Even a careful successive thresholding of the estimated source amplitudes does not reveal any evidence for the presence of a second source. In practice, these results would probably not provoke a user to try out other inverse methods in addition. Hence the deep-lying source is most likely overlooked. Figures A.21(d) - A.21(f) show the AS_CM, cmAO_MAP and McmAO_MAP results. Here the cmAO_MAP result clearly improves upon the AS_CM result. The AS_CM result seems only capable of marking an ambiguous region around and in-between the support of the true sources. The cmAO_MAP scheme subsequently uses this "region of interest" to draw a clearer picture of the source activity[8].

General properties The right validation of inverse methods for multiple source scenarios and a large number of different source configurations pose some problems. To generalize the commonly used DLE, an algorithm to reliably detect local maxima of the estimated source amplitude would be needed. Within the work for this thesis no such scheme was found. One problem arises from the outermost nodes. The source grid is constructed by choosing all nodes of a regular grid within a convex surface. Still, there are always points outside the convex hull of the source space nodes for which the two nearest source space nodes are not connected within the normal 26-neighborhood (cf. Figure A.11). Now assume that the estimated source amplitude originates, e.g., from the discretization of a Gaussian function centered on one of these outer points. This estimate will hence have two local amplitude maxima on the discrete graph formed by the source space nodes with a 26-neighborhood. For inverse methods suffering from depth bias often estimates are encountered that look like they originated from the discretization of a continuous distribution of which the local maxima lie on the boundary of the source compartment (cf. Figures A.18 and A.21(b)). As explained above, normal graph-based approaches for detecting local maxima will fail in such situations. Another problem is the right choice of thresholds. In particular the AS_CM scheme with a small number of steps R produces very non-smooth estimates, with many small source amplitudes as a result of the finite averaging. From the literature we are aware of, the only study that examines multiple source scenarios systematically is Phillips et al. (2002b) who also use the two source scenario. However, their approach was based on clustering the thresholded amplitudes. This should lead to the problems with the outer nodes again, and in addition the extension of their approach to more than two sources did not seem to lead to a feasible measure.
For these reasons, only EMD and SD are considered here. Table 4.10 shows both measures, averaged over all source configurations.

MAP approximations The different MAP approximations are compared in a similar way to the first study. Table 4.11 lists the results.

4.6.3 Discussion

The initial example showed that the source scenario examined in this study is a very challenging one for inverse methods. The methods that performed best in the first study, i.e., the Mc-

[8] The AS_CM result actually looks as if the MCMC method has not converged yet. To clarify this, $R = 20\,000\,000$ was used as well. The results look very similar, and are therefore not depicted. It is still possible that the markov chain is not ergodic for practical reasons.

Table 4.10: EMD and SD for the masking study, averaged over 250 source configurations.

Method	EMD	SD
AS_CM, $R = 5\,000$	15.96	2.54e-03
AS_CM, $R = 50\,000$	14.82	3.17e-03
AS_CM, $R = 200\,000$	14.70	3.32e-03
cmAO_MAP, $R = 5\,000$	13.69	8.72e-04
cmAO_MAP, $R = 50\,000$	12.54	8.42e-04
cmAO_MAP, $R = 200\,000$	12.15	8.22e-04
McmAO_MAP, $U = 64$	13.62	7.55e-04
uAO_MAP (inv. gamma)	42.93	1.5e-03
uAO_MAP (gamma)	36.51	7.03e-03
MNE	44.54	2.19e-01
WMNE ℓ_2	43.76	2.54e-01
WMNE $\ell_{\infty,reg}$	41.77	2.37e-01
sLORETA	36.34	1.94e-01

Table 4.11: Mean ranking of different MAP approximation methods in the second study.

Method	5% nl
cmAO_MAP, $R = 5\,000$	2.37
cmAO_MAP, $R = 50\,000$	1.80
cmAO_MAP, $R = 200\,000$	1.72
McmAO_MAP, $U = 64$	1.62

mAO_MAP and the cmAO_MAP scheme, also performed best in this study (cf. Table 4.10 and Figures A.21(e) and A.21(f)). Compared to each other the McmAO_MAP scheme still outperforms the cmAO_MAP scheme with regard to the posterior probability (cf. Table 4.11), but no longer concerning the EMD. This needs to be examined in more detail. In particular the tuning of U, Q and R for the McmAO_MAP scheme needs to be repeated for this scenario and extended for larger values of U. Similar to the first study, Table 4.10 shows that both AS_CM and cmAO_MAP benefit from a larger value for R and that the cmAO_MAP result improves upon the corresponding AS_CM result.
The results also suggest that the posterior distribution for these scenarios is more complex than for single sources. More comprehensive studies on this topic are therefore needed.

4.7 The Value of Wasserstein Metrics as Performance Measures

We briefly discuss some features of the use of Wasserstein metrics as a for the reconstruction performance of inverse methods. Remember that we introduced the earth mover's distance (EMD) as a particular example of a Wasserstein metric in order to have a measure that is both sensitive to localization and spatial extent of estimate (cf. Section 1.3.3).
In Figure 4.7 comparing the images for SD, EMD, 8^{th}-COME and 1^{st}-COME suggests that the EMD shares features of the localization measures (8^{th}-COME and 1^{st}-COME) as well as of the spatial extent measures (SP). Table 4.8 confirms this impression. However, with regard to the sLORETA estimate, it would be preferable if more weight is on the right localization. Even though the sLORETA method has a small DLE and is commonly used due to its localization properties, its EMD is much larger than for methods that produce focal estimates but mislocalize considerably (e.g., the uAO_MAP scheme with the gamma hyperprior).
The big advantage of the EMD is that it is applicable to more complex source scenarios just

as well. In contrast, the extension of other localization measures like the DLE is not straight forward, neither for the implementation nor from the interpretation of the results (cf. Section 4.6). In multiple source scenarios, more emphasis on the separation properties of the inverse methods would be preferable. In Table 4.10, sLORETA clearly outperforms MNE with regard to EMD. This is not fully consistent with the visual impression of the results in such source scenarios. When different single sources are active, sLORETA is likely to cluster the main activity in the center of mass of the true source locations and does not reveal the true number of the active sources. In contrast, the MNE result often shows the right number of local maxima, but only shifted due to the depth bias. Nonetheless, the EMD in this situation is usually smaller for sLORETA than for MNE. This relies on the fact that for sLORETA, the mass of the estimate has to be transported from the inside to the outside, where it is vice versa for the MNE. The difference is that in case of MNE, the local clusters on the outside need to have the exact mass that has to be transported to the next source location in order to attain a low EMD. In contrast, in case of sLORETA this is not important, since all the mass that has to be transported is clustered in a similar position. Hence it does not matter that much to which of the source locations it has to be transported.
In summary, Wasserstein metrics, in particular the EMD, are promising tools to investigate the performance of inverse methods. Especially for multiple source scenarios, where other measures are not easily available. Still more research on it has to be done to examine and improve certain features.

5 Conclusion

5.1 Summary

This thesis aimed at four main topics:

1. An elementary but consistent introduction to the mathematical modeling of bioelectromagnetism and the specific properties and the development of the field of EEG/MEG current density reconstruction (Chapter 1).

2. A more comprehensive illustration of the methodology underlying a specific branch of CDR methods, namely Bayesian statistics, and the introduction of an unifying theoretical framework called hierarchical Bayesian modeling which comprises many well established methods but also offers new ways of inference (Chapter 2).

3. The practical implementation of new estimation methods derived from the framework of hierarchical Bayesian modeling (Chapter 3).

4. The examination and comparison of these methods to established methods for specific source scenarios (Chapter 4).

In addition, a minor focus was on the connections between Bayesian inference and regularization approaches.

5.2 Discussion

- ⋆ The inverse methods introduced in this thesis, i.e., the cmAO_MAP and the McmAO_MAP scheme outperformed all other methods for the source scenarios examined (cf. Section 4.5 and 4.6). This confirms the big potential of the hierarchical modeling approach and justifies further research on that topic.

- ⋆ As functional brain imaging is a very interdisciplinary and relatively young field of research, concepts from many different fields entered this development at certain stages, different approaches were followed in parallel, and the diversity of methods used in practice and branches of research increased quite soon. Within the publications of the last five years, a clear trend is visible to search for unifying frameworks to clarify the basic properties and relations between the different methods for CDR. The hierarchical Bayesian modeling described in this thesis is a very promising candidate for this task (cf. Section 2.3 and A.1.8).

- ⋆ These new methods and concepts may also help to shed new light on the inverse problem of EEG/MEG in general and to clarify some common beliefs about it (cf. Section 2.4.3). For instance, a common belief is that "the measurements simply do not contain enough depth-information". However, the first study shows that there are actually even multiple methods that are able to localize within the right depth (cf. Section 4.5). The second study shows that the presence of a surface-near source does not necessarily conceal all information about a second, deep-lying source (cf. Section 4.6 and Figures A.21(a) - A.21(f)).

6 Outlook

Enhancement of the HBM-based Methods The thesis motivated further research on the particular hierarchical model and the methods based on it. Especially the properties of the real MAP estimate remain unknown. Within the thesis, only improvements concerning local modes could be achieved. It remains unclear how well theses modes already approximate the global mode. For this task, alternative optimization schemes need to be considered. The cmAO_MAP and the McmAO_MAP scheme performed best within this thesis. Since their computation takes quite long compared to other methods that are currently used or developed, faster implementations have to be found.

Comparison to other HBM-based Estimation Methods Only two of the possible estimation methods that the HBM offers were examined in this thesis (cf. Section 2.4.2). Especially for CM estimation we are only aware of Nummenmaa et al. (2007a,b); Calvetti et al. (2009) dealing with CM estimation as well (and Nummenmaa et al., 2007a only for a theoretical examination of the model). Most other publications using HBM deal with Variational Bayesian inference methods. As the CM estimate is different in nature from optimization based methods, a repetition of our studies involving all other HBM-based estimators would be desirable.

Validation with Real Data The significance of the simulation studies performed in this thesis could be increased considerably if they are partly confirmed by results for real data, even if only few data is available that is appropriate for validation.

Multimodal Integration *Multimodal integration* is a current focus of interest in medical imaging: To overcome the limitations of single modality recordings, techniques are developed to simultaneously or separately record and fuse different types of information by different imaging devices. As outlined in Section 2.4.3 HBM is more flexible for the inclusion of different types of information compared to classical WMNE schemes, and might thus be a promising framework for multimodal integration. The different possibilities for this and the impact on the different possible estimation methods have not yet been examined intensely enough, especially for real data.

Spatio-Temporal Extensions In this thesis, only the instantaneous inverse problem of CDR was addressed. A simple extension into the temporal domain would be to perform an instantaneous CDR to each time slice separately. However, due to the ill-condition and varying SNRs, the result is often unsatisfactory: The time course of the reconstruction is very unsmooth. Spatio-temporal CDRs aim to invert all time slices simultaneously by incorporating a-priori information about the spatio-temporal properties of the source activity. There is a variety of different approaches, and we will not go into details here. For the HBM approach presented in this thesis, temporal extensions have already been proposed as well, see, e.g., Trujillo-Barreto et al. (2008). We would suggest to extend the HBM on the level of the hyperparameters rather than on the level of sources: Temporal information is more of qualitative nature, more suitable to be incorporated at the higher stages of the model (cf. Section 2.4.3). An inclusion on the level of sources over temporal source covariance components would need precise information on strength, delay and decay of the temporal correlation of the source activity, which is not available at the moment. Furthermore, invalid information can lead to rigid schemes, likely to lose rapid, instationary

activity. Furthermore, the algorithmic complexity of the estimation process would be increased at the stage where most computational effort is spend even now. On the level of hyperparameters, temporal information would be embedded in a soft way, retaining flexibility at the source level. The algorithmic complexity would be increased at a stage where negligible time is spend by now. As discussed in the next paragraph, a re-parametrization of the HBM might make it possible to model temporally delayed inhibition and excitation between brain areas as well.

Modeling Inhibition and Excitation In section 2.4.3 we sketched the problem that within our model, modeling inhibition and excitation between different brain areas is not possible on level of the sources, i.e., over covariance components. On the level of hyperparameters, it is possible, and it would even be a better stage to incorporate such information for similar reasons as in the last paragraph: Such information is more of qualitative type, as precise information on strength, temporal delay and duration of inhibition and excitation processes is usually not available, and invalid information may lead to considerable errors. In addition, as noted in the previous paragraph, the algorithmic complexity of the estimation process would be increased at the stage where less time is spend right now.
The standard definition of statistical correlation aimes at length variables that are best described on a linear scale. To model statistical correlation between scale variables like our hyperparameters, a different parametrization of the HBM is advantageous: In the parametrization we chose to construct the HBM, the hyperparameters γ_i describe the scale of the corresponding covariance components C_i. Another possible parametrization is to let the hyperparameters determine the logarithm of this scale:

$$\Sigma_s(\boldsymbol{\gamma}) = \sum_{i=1}^{h} \exp(\gamma_i) C_i$$

See, e.g., Friston et al. (2008). The log-space is a more natural space to describe a scale variable like $\boldsymbol{\gamma}$ (many practical algorithms for estimators formulated in our setting actually operate in log-space, see Wipf and Nagarajan, 2009). Now a positive correlation between two hyperparameters encourages them to simultaneously take values that are larger or smaller than their mean, which means that the scale of the variance of the corresponding source locations is simultaneously larger or smaller than average. This way, activation in one brain area can trigger activation in another. A negative correlation between two hyperparameters leads to the opposite: If one is positive, which means that the corresponding source location has a large variance compared to the average, the other is likely to be negative, which means that source activity in the corresponding location is inhibited.

Effects of Realistic Head Modeling Within this thesis only a realistically shaped high resolution FEM model without inner brain compartments and very simplified model was used. As implied in Section 4.4.1 and 4.5, the interplay between forward model and inverse method may be non-trivial as well and cause counterintuitive phenomena.

Analytical Treatment of the Depth Bias The results of the first study remain unsatisfactory in the aspect that it was not clarified where the actual cause for the depth bias lies, and why some methods suffer from it, while others do not. This has to be examined also from a theoretical perspective.

A Appendix

A.1 Miscellaneous

A.1.1 Normal, Relaxed and Weighted Least Squares Problems

Here we briefly summarize some facts about least squares problems that frequently occur in the framework of ℓ_2-based regularization as well as in Bayesian modeling with Gaussian priors (for references, see, e.g., Ben-Israel and Greville, 2003; Kaipio and Somersalo, 2005; Hastie et al., 2009)
Consider a underdetermined, but full rank matrix equation:

$$\mathrm{A}x = b, \quad \text{where} \quad \mathrm{A} \in \mathbb{R}^{m \times n}, \quad m < n, \quad \operatorname{rank}(\mathrm{A}) = m$$

Since there are infinitely many solutions to this system, we are interested in the one with the smallest ℓ_2 norm, which can be formulated as

$$x = \operatorname{argmin} \left\{ \|x\|_2^2 \right\}, \quad \text{such that} \quad \|\mathrm{A}x - b\|_2^2 = 0 \tag{A.1}$$

The solution of this problem is given by the solution of the *normal equations*, which can be used to define the pseudo inverse of a non-square matrix A:

$$\mathrm{A}^t\mathrm{A}x = \mathrm{A}^t b \quad \Rightarrow \quad x = \mathrm{A}^+ b \quad \text{where} \quad \mathrm{A}^+ := (\mathrm{A}^t\mathrm{A})^{-1}\mathrm{A}^t \tag{A.2}$$

Especially when the condition of A is bad and b might contain measurement errors, it is helpful to add some regularization to the problem. This can, e.g., be done by relaxing the problem (A.1), i.e., substituting the hard constraint $\|\mathrm{A}x - b\|_2^2 = 0$ by a softened version:

$$\begin{aligned} x &= \operatorname{argmin} \left\{ \|\mathrm{A}x - b\|_2^2 + \lambda \|x\|_2^2 \right\}, \quad \text{where} \quad \lambda > 0 \\ &= \operatorname{argmin} \left\{ \left\| \begin{bmatrix} \mathrm{A} \\ \sqrt{\lambda}\mathrm{Id}_n \end{bmatrix} x - \begin{bmatrix} b \\ 0 \end{bmatrix} \right\|_2^2 \right\} \\ \Leftrightarrow: \begin{bmatrix} \mathrm{A} \\ \sqrt{\lambda}\mathrm{Id}_n \end{bmatrix} x &\overset{ls}{=} \begin{bmatrix} b \\ 0 \end{bmatrix} \end{aligned} \tag{A.3}$$

This is an overdetermined linear least squares problem, whose solution is also given by the corresponding normal equations:

$$\left(\mathrm{A}^t\mathrm{A} + \lambda \mathrm{Id}_n\right) x = \mathrm{A}^t b \quad \Rightarrow \quad x = \left(\mathrm{A}^t\mathrm{A} + \lambda \mathrm{Id}_n\right)^{-1} \mathrm{A}^t b$$

The normal equations for the relaxed problem are referred to as *relaxed normal equations* for the original problem $\mathrm{A}x = b$. For different reasons, one might introduce some weighting into either the relaxation, the data fit or both, i.e., in the domain of A, its range or in both. Let $\mathrm{W}_r \in \mathbb{R}^{m \times m}$ and $\mathrm{W}_d \in \mathbb{R}^{n \times n}$ be non singular[1]:

$$\begin{aligned} x &= \operatorname{argmin} \left\{ \|\mathrm{W}_r(\mathrm{A}x - b)\|_2^2 + \lambda \|\mathrm{W}_d x\|_2^2 \right\} \\ &= \operatorname{argmin} \left\{ \left\| \begin{bmatrix} \mathrm{W}_r\mathrm{A} \\ \sqrt{\lambda}\mathrm{W}_d \end{bmatrix} x - \begin{bmatrix} \mathrm{W}_r b \\ 0 \end{bmatrix} \right\|_2^2 \right\} \\ \overset{(A.2)}{\Longleftrightarrow} \left(\mathrm{A}^t\mathrm{W}_r^t\mathrm{W}_r\mathrm{A} + \lambda \mathrm{W}_d^t\mathrm{W}_d\right) x &= \mathrm{A}^t\mathrm{W}_r^t\mathrm{W}_r b \\ \Longleftrightarrow x &= \left(\mathrm{A}^t\mathrm{W}_r^t\mathrm{W}_r\mathrm{A} + \lambda \mathrm{W}_d^t\mathrm{W}_d\right)^{-1} \mathrm{A}^t\mathrm{W}_r^t\mathrm{W}_r b \end{aligned} \tag{A.4}$$

[1] This scheme easily extents to non-square weightings, see e.g., Calvetti and Somersalo, 2008a.

One speaks of a *weighted* or *weighted relaxed* least squares problem, and of *weighted* or *weighted relaxed* normal equations, depending on whether relaxation is used or not (set $\lambda = 0$ in the equations above).

A.1.2 Matrix Calculus

In this thesis we use some results from the theory of block partitioned matrices, which we will list in the following. Proofs can, e.g., be found in Bernstein (2009).
Let

$$\Gamma = \begin{bmatrix} \Gamma_{11} & \Gamma_{12} \\ \Gamma_{21} & \Gamma_{22} \end{bmatrix} \in \mathbb{R}^{(n+m)\times(n+m)}$$

where $\Gamma_{11} \in \mathbb{R}^{n\times n}$, $\Gamma_{12} \in \mathbb{R}^{n\times m}$, $\Gamma_{21} \in \mathbb{R}^{m\times n}$ and $\Gamma_{22} \in \mathbb{R}^{m\times m}$ are matrices, such that Γ_{11} and Γ_{22} are not singular.

Definition 10 (Schur complements) *The Schur complements $\tilde{\Gamma_{jj}}$ of Γ_{jj}, $j = 1, 2$, are defined by:*

$$\tilde{\Gamma_{22}} = \Gamma_{11} - \Gamma_{12}\Gamma_{22}^{-1}\Gamma_{21}, \qquad \tilde{\Gamma_{11}} = \Gamma_{22} - \Gamma_{21}\Gamma_{11}^{-1}\Gamma_{12}$$

The Schur complements play an important role in calculations involving Γ:

Lemma 1 (Schur identity)

$$|\Gamma| = |\Gamma_{11}|\,|\tilde{\Gamma}_{11}| = |\Gamma_{22}|\,|\tilde{\Gamma}_{22}|$$

Lemma 2 (Block Matrix Inversion)

$$\begin{aligned}\Gamma^{-1} &= \begin{bmatrix} \tilde{\Gamma}_{22}^{-1} & -\tilde{\Gamma}_{22}^{-1}\Gamma_{12}\Gamma_{22}^{-1} \\ -\tilde{\Gamma}_{11}^{-1}\Gamma_{21}\Gamma_{11}^{-1} & \tilde{\Gamma}_{11}^{-1} \end{bmatrix} \\ &= \begin{bmatrix} \Gamma_{11}^{-1} + \Gamma_{11}^{-1}\Gamma_{21}\tilde{\Gamma}_{11}^{-1}\Gamma_{21}\Gamma_{11}^{-1} & -\Gamma_{11}^{-1}\Gamma_{12}\tilde{\Gamma}_{11}^{-1} \\ -\Gamma_{22}^{-1}\Gamma_{21}\tilde{\Gamma}_{22}^{-1} & \Gamma_{22}^{-1} + \Gamma_{22}^{-1}\Gamma_{21}\tilde{\Gamma}_{22}^{-1}\Gamma_{12}\Gamma_{22}^{-1} \end{bmatrix}\end{aligned}$$

By equating the blocks and re-substituting the Schur complements the above formulas yield some useful identities (sometimes referred to as "matrix inversion lemma"):

$$\begin{aligned}\left(\Gamma_{11} - \Gamma_{12}\Gamma_{22}^{-1}\Gamma_{21}\right)^{-1} &= \Gamma_{11}^{-1} + \Gamma_{11}^{-1}\Gamma_{12}\left(\Gamma_{22} - \Gamma_{21}\Gamma_{11}^{-1}\Gamma_{12}\right)^{-1}\Gamma_{21}\Gamma_{11}^{-1} \\ \left(\Gamma_{11} - \Gamma_{12}\Gamma_{22}^{-1}\Gamma_{21}\right)^{-1}\Gamma_{12}\Gamma_{22}^{-1} &= \Gamma_{11}^{-1}\Gamma_{12}\left(\Gamma_{22} - \Gamma_{21}\Gamma_{11}^{-1}\Gamma_{12}\right)^{-1} \\ \left(\Gamma_{22} - \Gamma_{21}\Gamma_{11}^{-1}\Gamma_{12}\right)^{-1}\Gamma_{21}\Gamma_{11}^{-1} &= \Gamma_{22}^{-1}\Gamma_{21}\left(\Gamma_{11} - \Gamma_{12}\Gamma_{22}^{-1}\Gamma_{21}\right)^{-1} \qquad \text{(A.5)} \\ \left(\Gamma_{22} - \Gamma_{21}\Gamma_{11}^{-1}\Gamma_{12}\right)^{-1} &= \Gamma_{22}^{-1} + \Gamma_{22}^{-1}\Gamma_{21}\left(\Gamma_{11} - \Gamma_{12}\Gamma_{22}^{-1}\Gamma_{21}\right)^{-1}\Gamma_{12}\Gamma_{22}^{-1} \qquad \text{(A.6)}\end{aligned}$$

From (A.5) we can deduce an identity that can be used for the inverse computations in Chapter 1.3.2: Choosing $\Gamma_{11} := \Sigma_\varepsilon$, $\Gamma_{12} := -\mathrm{L}$, $\Gamma_{21} := \mathrm{L}^t$ and $\Gamma_{22} := \Sigma_s^{-1}$ it follows that

$$\begin{aligned}\Sigma_s\mathrm{L}^t(\mathrm{L}\Sigma_s\mathrm{L}^t + \Sigma_\varepsilon)^{-1} &= \Gamma_{22}^{-1}\Gamma_{21}\left(\Gamma_{11} - \Gamma_{12}\Gamma_{22}^{-1}\Gamma_{21}\right)^{-1} \\ &\overset{(A.5)}{=} \left(\Gamma_{22} - \Gamma_{21}\Gamma_{11}^{-1}\Gamma_{12}\right)^{-1}\Gamma_{21}\Gamma_{11}^{-1} = \left(\Sigma_s^{-1} + \mathrm{L}^t\Sigma_\varepsilon^{-1}\mathrm{L}\right)^{-1}\mathrm{L}^t\Sigma_\varepsilon^{-1} \qquad \text{(A.7)}\end{aligned}$$

Note that on the left hand side, a $m \times m$ matrix has to be inverted, whereas it is a $n \times n$ matrix on the right hand side.
From (A.6), an identity that is helpful for the computation of the conditional covariance can be derived (see (3.9)): Using the same definitions as above, it follows that

$$\begin{aligned}\Sigma_s - \Sigma_s\mathrm{L}^t\left(\Sigma_\varepsilon + \mathrm{L}\Sigma_s\mathrm{L}^t\right)^{-1}\mathrm{L}\Sigma_s &= \Gamma_{22}^{-1} + \Gamma_{22}^{-1}\Gamma_{21}\left(\Gamma_{11} - \Gamma_{12}\Gamma_{22}^{-1}\Gamma_{21}\right)^{-1}\Gamma_{12}\Gamma_{22}^{-1} \\ &\overset{(A.6)}{=} \left(\Gamma_{22} - \Gamma_{21}\Gamma_{11}^{-1}\Gamma_{12}\right)^{-1} = \left(\Sigma_s^{-1} + \mathrm{L}^t\Sigma_\varepsilon^{-1}\mathrm{L}\right)^{-1} \qquad \text{(A.8)}\end{aligned}$$

A.1.3 Theoretical Comparison of Statistical Estimators

Remind that in our framework, every variable is modeled as a random variable. Thus an estimator $\hat{s}(b)$ for the random realization s can also be considered as a random variable, as it takes different values depending on the realization of the noisy measurement B. A main concern of statistical estimation theory is to answer the question, how this estimator $\hat{S} = \hat{s}(B)$, now seen as a random variable, behaves *in general* rather than just for one given single measurement b. It might be that it gives good estimations of s for some realizations of B while giving catastrophic results for others. A natural tool to examine this behavior quantitatively is a *cost function* (or termed *loss* function in statistics) $\Psi : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ so that $\Psi(s, \hat{s})$ gives a measure for the desired and undesired properties of $\hat{s}$. The *Bayes cost* is then defined as:

$$\begin{aligned} BC(\hat{s}) = \mathbb{E}\left[\Psi(S, \hat{s}(B))\right] &= \iint \Psi(s, \hat{s}(b))\, p(s, b)\, ds\, db \\ &= \iint \Psi(s, \hat{s}(b))\, p_{like}(b|s)\, db\, p_{prior}(s)\, ds \\ &= \int BC(\hat{s}|s)\, p_{prior}(s)\, ds = \mathbb{E}\left[BC(\hat{s}|s)\right] \end{aligned}$$

where

$$BC(\hat{s}|s) = \int \Psi(s, \hat{s}(b))\, p_{like}(b|s)\, db$$

is the *conditional Bayes cost*.
For a given cost function, one theoretical approach to find an estimator, called *Bayes cost method* is to choose the one that minimizes the Bayes cost:

$$BC(\hat{s}_{BC}) \leq BC(\hat{s}) \qquad \forall\, \hat{s} : \mathbb{R}^m \to \mathbb{R}^n$$

This is called the *Bayes estimator*. Note that this estimator is the one that performs best for this cost function *on average*, it may not be optimal over other criteria. By using Bayes formula, we can write the Bayes cost in the form:

$$BC(\hat{s}) = \iint \Psi(s, \hat{s}(b))\, p_{post}(s|b)\, ds\, p(b) db$$

Since the marginal density $p(b)$ satisfies $p(b) \geq 0$ and $\hat{s}(b)$ does only depend on b, the minimizer of the Bayes cost is found by solving

$$\hat{s}_{BC}(b) = \underset{\hat{s}}{\operatorname{argmin}} \left\{ \int \Psi(s, \hat{s}(b))\, p_{post}(s|b)\, ds \right\} = \underset{\hat{s}}{\operatorname{argmin}} \left\{ \mathbb{E}\left[\Psi(s, \hat{s}(b))|b\right] \right\} \tag{A.9}$$

In the following, we will show how to derive the two estimators used in this thesis and an additional one within this framework.

CM-estimation: The most common choice for the cost function is $\Psi(s, \hat{s}(b)) = \|s - \hat{s}\|_2^2$ which leads to the *mean square error criterion*, as the Bayes cost takes the form:

$$BC(\hat{s}) = \mathbb{E}\left[\|S - \hat{S}\|_2^2\right]$$

The corresponding Bayes estimator is called *mean square estimator* and from (A.9) it turns out that:

$$\hat{s}_{MS} := \hat{s}_{BC} = \underset{\hat{s}}{\operatorname{argmin}} \left\{ \mathbb{E}\left[\|S - \hat{s}\|_2^2\right] \right\} = \int s\, p_{post}(s|b)\, ds = \hat{s}_{CM}$$

That is, our previously defined conditional mean estimator is the mean square estimator. One can further show that the CM-estimated is *unbiased* and that it is also the *minimum error variance* estimator.

Geometric median-estimation: Alongside the mean and the mode of a distribution which represent center of mass and maximal mass, the *median* is another characteristic point: For one dimensional distributions, it divides the support of the density in two halves in such a way that on both sides equal probability mass is located, i.e., 50 %. Since the median is very robust against outliers, it is used in a wide range of applications from clustering to noise removal in signal processing. However, its theoretical treatment as well as its generalization to high dimensional problems like ours is not trivial, and thus its is seldom examined. One generalization is given by the *geometric median* which is the Bayes estimator for the cost function $\Psi(s, \hat{s}(b)) = \|s - \hat{s}\|_1$.

MAP-estimation: The MAP-estimate is just *asymptotically* a Bayes estimator. One way to infer it within our framework is to define it as the limit $\epsilon \to 0$ of the solution to the Bayes cost optimization for

$$\Psi_\epsilon(s, \hat{s}) = \begin{cases} 0, & \text{if} \quad \|s_k - \hat{s}_k\| < \epsilon \quad \text{for all } k = 1, \ldots, n \\ 1 & \text{otherwise,} \end{cases}$$

because in this case, we have to solve

$$\begin{aligned} \hat{s}_{BC}(b) &= \operatorname{argmin} \left\{ \int \Psi_\epsilon(s, \hat{s}(b))\, p_{post}(s|b)\, ds \right\} \\ &= \operatorname{argmin} \left\{ \int_{|s_k - \hat{s}_k| > \epsilon} p_{post}(s|b)\, ds \right\} \\ &= \operatorname{argmin} \left\{ 1 - \prod_{k=1}^{n} \int_{\hat{s}_k - \epsilon}^{\hat{s}_k + \epsilon} p_{post}(s|b)\, ds_k \right\} \\ &\approx \operatorname{argmin} \{1 - (2\epsilon)^n p_{post}(\hat{s}|b)\} \\ &= \operatorname{argmax} \{p_{post}(\hat{s}|b)\} \\ &= \hat{s}_{MAP}(b) \end{aligned}$$

The limit of Ψ_ϵ for $\epsilon \to 0$ is called *uniform cost* or *0-1 loss*, as it penalizes every deviation of $\hat{s}$ from s equally.
So in summary, while the CM-estimator penalizes large errors heavily, while neglecting small ones, the MAP-estimate treats small and large errors equally. Their statistical estimation conception is therefore completely different, and from the Bayes cost formalism, it is clear, why it follows that their statistical properties are also completely different: They aim to minimize completely different error-measures. Nevertheless, both estimators are optimal over their own criteria. In EEG/MEG source-reconstruction, most of the existing methods aim to find a MAP-estimator, but that is for practical reasons rather than for theoretical. As mentioned above, the CM-estimator has some very attractive properties, too, and in other fields of parameter estimation, it is the standard estimator for these reasons.
As a final remark, comparing with A.3 one could have used the scaled minimum support stabilizer as a cost function to derive the MAP estimate as well, which poses the question, whether it is possible to derive the MAP estimate as a limit $p \searrow 0$ of an suitably scaled ℓ_p-based Bayes cost function. This would ease the comparison between the different estimators.

A.1.4 Gaussian Densities

In this section, some basic properties of Gaussian random variables are summarized. For proofs and further references, see, e.g., Kaipio and Somersalo (2005); Klenke (2008).

Affine Transformation and Mixing

The following rules are used in many situations in this thesis:

Lemma 3 *Let $X_1 \sim \mathcal{N}_n(\mu_1, \Gamma_1)$, $X_2 \sim \mathcal{N}_n(\mu_2, \Gamma_2)$ be two mutually independent Gaussian random variables, and $c \in \mathbb{R}^m$, $\mathrm{A} \in \mathbb{R}^{m \times n}$.*

(i) $Z_{aff} = \mathrm{A}X_1 + c$ is a m dimensional Gaussian random variable with mean $\mathrm{A}\mu_1 + c$ and covariance matrix $\mathrm{A}\Gamma_1\mathrm{A}^t$, i.e., $Z_{aff} \sim \mathcal{N}_m(\mathrm{A}\mu_1 + c, \mathrm{A}\Gamma_1\mathrm{A}^t)$.

(ii) $Z_{sum} = X_1 + X_2$ is a n dimensional Gaussian random variable with mean $\mu_1 + \mu_2$ and covariance matrix $\Gamma_1 + \Gamma_2$, i.e., $Z_{sum} \sim \mathcal{N}_n(\mu_1 + \mu_1, \Gamma_1 + \Gamma_2)$.

Rule (i) immediately yields an effective scheme to generate M samples x_i from a Gaussian random variable X_1 defined as above:

Algorithm 8 (Sampling from Gaussian Densities)

1. *Find any real matrix C such that $\mathrm{C}\,\mathrm{C}^t = \Gamma_1$ (Usually via a normal cholesky decomposition given that Γ_1 is positive-definite, and a variant based on an eigenvalue decomposition in the degenerate case).*
2. *For $i = 1, \ldots, M$ draw ω_i from $\otimes_{k=1}^{n} \mathcal{N}_1(0,1)$, i.e., componentwise from a one dimensional standard normal distribution (e.g., by using the Box–Muller transform).*
3. *Set $x_i = \mathrm{C}\omega_i + \mu_1$ for all i.*

However, such a scheme is only useful if a large number of samples M have to be drawn from the same distribution. For the Ss step used in the AS_CM algorithm for CM approximation (cf. Section 3.2) the conditional density that is sampled changes in every step. Computing both mean and covariance matrix over (3.8) and (3.9) and a cholesky decomposition of the covariance matrix would result in unnecessary overhead, and is numerically not stable, as discussed in the corresponding section. We demonstrate in the following that the proposed alternative scheme is valid for this task, i.e., to generate a sample of the g dimensional $\tilde{s}$ from the conditional density given by (3.8) and (3.9) by an affine mixing of two m and g dimensional standard normal distributed random variables ω_m and ω_g via solving the least squares problem (3.10):

$$
\begin{aligned}
(3.10) &\iff \begin{bmatrix} \frac{1}{\sigma}\tilde{\mathrm{L}} \\ \mathrm{D}^{-1/2} \end{bmatrix} s \overset{ls}{=} \begin{bmatrix} \sigma^{-1} b \\ 0 \end{bmatrix} + \begin{bmatrix} \omega_m \\ \omega_g \end{bmatrix} := \begin{bmatrix} \sigma^{-1} b \\ 0 \end{bmatrix} + \omega \\
&\overset{(A.1.1)}{\iff} s = \left(\begin{bmatrix} \frac{1}{\sigma}\tilde{\mathrm{L}}^t & \mathrm{D}^{-1/2} \end{bmatrix} \begin{bmatrix} \frac{1}{\sigma}\tilde{\mathrm{L}}^t \\ \mathrm{D}^{-1/2} \end{bmatrix} \right)^{-1} \begin{bmatrix} \frac{1}{\sigma}\tilde{\mathrm{L}}^t & \mathrm{D}^{-1/2} \end{bmatrix} \left(\begin{bmatrix} \frac{1}{\sigma} b \\ 0 \end{bmatrix} + \omega \right) \\
&\iff s = \left(\tfrac{1}{\sigma^2}\tilde{\mathrm{L}}^t\tilde{\mathrm{L}} + \mathrm{D}^{-1} \right)^{-1} \begin{bmatrix} \frac{1}{\sigma}\tilde{\mathrm{L}}^t & \mathrm{D}^{-1/2} \end{bmatrix} \omega + \left(\tfrac{1}{\sigma^2}\tilde{\mathrm{L}}^t\tilde{\mathrm{L}} + \mathrm{D}^{-1} \right)^{-1} \tfrac{1}{\sigma^2}\tilde{\mathrm{L}}^t b
\end{aligned}
$$

From Lemma 3 it follows that since $\omega \sim \mathcal{N}(0, \mathrm{Id}_{m+g})$, $\tilde{s}$ also follows a Gaussian distribution with mean and covariance matrix given by:

$$
\begin{aligned}
\mathbb{E}[s] &= \left(\tfrac{1}{\sigma^2}\tilde{\mathrm{L}}^t\tilde{\mathrm{L}} + \mathrm{D}^{-1} \right)^{-1} \tfrac{1}{\sigma^2}\tilde{\mathrm{L}}^t b \overset{(A.7)}{=} \mathrm{D}\tilde{\mathrm{L}}^t \left(\tilde{\mathrm{L}}\mathrm{D}\tilde{\mathrm{L}}^t + \sigma^2 \mathrm{Id}_m \right)^{-1} b = (3.8) \\
\mathbb{C}\mathrm{ov}[s] &= \underbrace{\left(\tfrac{1}{\sigma^2}\tilde{\mathrm{L}}^t\tilde{\mathrm{L}} + \mathrm{D}^{-1} \right)^{-1} \begin{bmatrix} \frac{1}{\sigma}\tilde{\mathrm{L}}^t & \mathrm{D}^{-1/2} \end{bmatrix} \begin{bmatrix} \frac{1}{\sigma}\tilde{\mathrm{L}} \\ \mathrm{D}^{-1/2} \end{bmatrix}}_{\mathrm{Id}} \left(\left(\tfrac{1}{\sigma^2}\tilde{\mathrm{L}}^t\tilde{\mathrm{L}} + \mathrm{D}^{-1} \right)^{-1} \right)^t \\
&= \left(\mathrm{D}^{-1} + \frac{1}{\sigma^2}\tilde{\mathrm{L}}^t\tilde{\mathrm{L}} \right)^{-1} = (3.9)
\end{aligned}
$$

Joint, Conditional and Marginal Distributions

Theorem 1 *Let X_1 be a n dimensional and X_2 a m dimensional Gaussian random variable whose joint density is of the form*

$$p(x_1, x_2) \propto \exp\left(-\frac{1}{2}\begin{bmatrix} x_1-\mu_1 \\ x_2-\mu_2 \end{bmatrix}^t \begin{bmatrix} \Gamma_{11} & \Gamma_{12} \\ \Gamma_{21} & \Gamma_{22} \end{bmatrix}^{-1} \begin{bmatrix} x_1-\mu_1 \\ x_2-\mu_2 \end{bmatrix}\right)$$

Then the marginal densities of X_1 and X_2 are given by:

$$p(x_1) = \int_{\mathbb{R}^m} p(x_1, x_2)\, dx_2 = \mathcal{N}_n(x, \mu_1, \Sigma_{11})$$
$$p(x_2) = \int_{\mathbb{R}^n} p(x_1, x_2)\, dx_1 = \mathcal{N}_n(x, \mu_2, \Sigma_{22})$$

Furthermore, the probability distribution of X_1 conditioned on $X_2 = x_2$, i.e., $p(x_1|x_2)$ is of the form:

$$p(x_1|x_2) \propto \exp\left(-\frac{1}{2}(x_1-\bar{\mu}_1)^t \tilde{\Gamma}_{22}^{-1}(x_1-\bar{\mu}_1)\right) \quad \textit{where} \quad \bar{\mu}_1 = \mu_1 + \Gamma_{12}\Gamma_{22}^{-1}(x_2-\mu_2)$$

That means, that $p(x_1|x_2) = \mathcal{N}_n(\mu_1 + \Gamma_{12}\Gamma_{22}^{-1}(x_2-\mu_2), \tilde{\Gamma}_{22})$ where $\tilde{\Gamma}_{22}$ is the Schur complement of Γ_{22} as defined in A.1.2.

Now we apply this theorem to a linear model with additive noise, i.e.

$$Y = \mathrm{A}X + \mathcal{E}, \quad \text{where} \quad \mathrm{A} \in \mathbb{R}^{m\times n}$$

We further assume that X and $\mathcal{E}$ are mutually independent and that $X \sim \mathcal{N}(\mu_x, \Gamma_{pr})$ and $\mathcal{E} \sim \mathcal{N}(\mu_\varepsilon, \Gamma_{noise})$. We could compute the posterior distribution directly by Bayes rule (2.3), however, the above formulas yield a more gentle way to do so. From A.1.4, it follows that $Y \sim \mathcal{N}(\mu_y, \mathrm{A}\Gamma_{pr}\mathrm{A}^t + \Gamma_{noise})$, with $\mu_y := \mathrm{A}\mu_x + \mu_\mathcal{E}$. To set up the joint distribution of X and Y, we have to compute $\mathbb{E}[(X-\mu_x)(Y-\mu_y)^t]$ and $\mathbb{E}[(Y-\mu_y)(X-\mu_x)^t]$:
Substituting Y by X and $\mathcal{E}$ and using the mutual independence of both it follows directly that

$$\mathbb{E}\left[(X-\mu_x)(Y-\mu_y)^t\right] = \mathbb{E}\left[(X-\mu_x)(\mathrm{A}(X-\mu_x) - (\mathcal{E}-\mu_\varepsilon)^t)\right] = \Gamma_{pr}\mathrm{A}^t$$
$$\mathbb{E}\left[(Y-\mu_y)(X-\mu_x)^t\right] = \mathbb{E}\left[(\mathrm{A}(X-\mu_x) - (\mathcal{E}-\mu_\varepsilon)(X-\mu_x)^t)\right] = \mathrm{A}\Gamma_{pr}$$

Thus the joint distribution is given by

$$p(x, y) \propto \exp\left(-\frac{1}{2}\begin{bmatrix} x-\mu_x \\ y-\mu_y \end{bmatrix}^t \begin{bmatrix} \Gamma_{pr} & \Gamma_{pr}\mathrm{A}^t \\ \mathrm{A}\Gamma_{pr} & \mathrm{A}\Gamma_{pr}\mathrm{A}^t + \Gamma_{noise} \end{bmatrix}^{-1} \begin{bmatrix} x-\mu_x \\ y-\mu_y \end{bmatrix}\right)$$

Then by theorem 1, the posterior density of X given Y is

$$p_{post}(x|y) \;=\; \mathcal{N}(\hat{x}, \Gamma_{post})$$

where

$$\begin{aligned} \hat{x} &= \mu_x + \Gamma_{pr}\mathrm{A}^t(\mathrm{A}\Gamma_{pr}\mathrm{A}^t + \Gamma_{noise})^{-1}(y - \mathrm{A}\mu_x - \mu_\varepsilon) \\ &\overset{(A.5)}{=} (\Gamma_{pr}^{-1} + \mathrm{A}^t\Gamma_{noise}^{-1}\mathrm{A})^{-1}(\mathrm{A}^t\Gamma_{noise}^{-1}(y-\mu_\varepsilon) + \Gamma_{pr}^{-1}\mu_x) \end{aligned}$$

and

$$\Gamma_{post} \;=\; \Gamma_{pr} - \Gamma_{pr}\mathrm{A}^t(\mathrm{A}\Gamma_{pr}\mathrm{A} + \Gamma_{noise})^{-1}\mathrm{A}\Gamma_{pr} \overset{(A.6)}{=} (\Gamma_{pr}^{-1} + \mathrm{A}^t\Gamma_{noise}^{-1}\mathrm{A})^{-1}$$

The last line states that in the sense of quadratic forms, $\Gamma_{post} \leqslant \Gamma_{pr}$. As the covariance expresses the width of the density this means, that *a measurement can never increase the uncertainty.*

A.1.5 Relation Between WMNE and Gaussian Prior Models

In Section 2.3 a direct link between the weighting matrix of the WMNE regularization scheme and the covariance matrix of a Gaussian prior in the statistical framework is shown. Up to a scaling by $\frac{\sqrt{\lambda}}{\sigma}$, W is an inverse square root of the covariance matrix Σ_s. This connection can help to understand both approaches better:

- ⋆ First, we start off with a WMNE scheme, with a regular W. Since $\|\mathrm{W}\, s\|_2 = \|\mathrm{U}\,\mathrm{W}\, s\|_2$ for any unitary U, not W but $\mathrm{V} := \mathrm{W}^t\,\mathrm{W}$ is the central ingredient that enters the method: $\|\mathrm{W}\, s\|_2^2 = (\mathrm{W}\, s)^t(\mathrm{W}\, s)\, s = s^t\,(\mathrm{W}^t\mathrm{W})\, s = s^t\,\mathrm{V}\, s$. Given some possible solutions s (this means that the data deviation is not very large), those having large components in the eigenspaces of V belonging to large eigenvalues will be penalized heavily and will consequently be rejected. Possible solutions s, mainly consisting of components in the eigenspaces of V belonging to small eigenvalues, will be favored. This implicit pruning of undesired solution components finds its explicitly observable counterpart in the statistical framework if we compute the inverse of V and use it as a covariance matrix C for a zero-mean Gaussian prior on s. The eigenspaces of C are the same as those of V, but to the reversed eigenvalues. In the statistical framework an eigenspace of the covariance matrix belonging to a large eigenvalue means that it is very likely that the solution has a large component in that eigenspace.
  For instance, if we choose W to be a discretized Laplacian operator, the solutions given by the WMNE are usually spatially very smooth. This is not surprising, because if we proceed as described above and calculate the corresponding covariance matrix it is a dense matrix imposing very strong correlations between all locations, only decreasing slowly with the distance between them. Thus a spatially smooth solution is extremely likely.

- ⋆ Now we start in the statistical framework and want to compute the MAP estimate for a Gaussian prior on s having a covariance matrix C and zero mean. If we reverse the above procedure, we end up with a WMNE scheme with a weighting matrix $\mathrm{C}^{-1/2}$, with $\mathrm{C}^{-1/2}$ being any of the square roots of $\mathrm{C}^{-1} = \mathrm{V}$. The penalty functional takes the form $\|\mathrm{C}^{-1/2}\, s\|_2^2$. In the field of signal processing, this transformation of s via $\mathrm{C}^{-1/2}$ is called *whitening transformation* as it decorrelates the entries of s, i.e., if $S \sim \mathcal{N}(0, \mathrm{C})$ and $Y = \mathrm{C}^{-1/2}\, S$ then $Y \sim \mathcal{N}(0, \mathrm{Id})$. Applied to data coming from different distributions, this transform has a probabilistic selecting effect: For a set of samples $\{x_i\}_i$ drawn from $\mathcal{N}(0, \mathrm{C})$ and a set of samples $\{\bar{x}_i\}_i$ drawn from $\mathcal{N}(0, \bar{\mathrm{C}})$ with the same average signal power ($|\mathrm{C}| = |\tilde{\mathrm{C}}|$), the set $\{\|\mathrm{C}^{-1/2} x_i\|_2\}_i$ will *on average* contain smaller values than $\{\|\mathrm{C}^{-1/2}\bar{x}_i\|_2\}_i$. That means that estimates s that are likely to come from our prior distribution, which on the other hand means that they fulfill our modeling assumptions, will be punished less hard by the corresponding penalty functional in the WMNE scheme.

A.1.6 Recast of the EMD Problem into Standard Form

In this section, we show how to recast the minimization problem in Section 3.7 into the standard formulation used in linear programming:

$$\min_x \left(c^t \cdot x\right) \quad \text{such that} \quad \begin{cases} \mathrm{A}\, x \leqslant b, \\ \mathrm{A}_{eq}\, x = b_{eq}, \\ lb \leqslant x \leqslant ub \end{cases} \tag{A.10}$$

For this purpose, let:

$$\begin{aligned} c &:= [\mathrm{D}^t_{(\cdot,1)}, \dots, \mathrm{D}^t_{(\cdot,k)}]^t \\ \mathrm{A}_{eq} &:= \begin{bmatrix} \mathbb{1}_k^t \otimes \mathrm{Id}_\tau \\ \mathrm{Id}_k \otimes \mathbb{1}_\tau^t \end{bmatrix} \quad \text{where} \quad \mathbb{1}_l \in \mathbb{R}^l \quad \text{is given by} \quad (\mathbb{1}_l^t)_i = 1 \, \forall i = 1, \dots, l \\ b_{eq} &:= (w_{p_1}, \dots, w_{p_\tau}, w_{q_1}, \dots, w_{q_k})^t \\ lb &:= 0 \in \mathbb{R}^{k \cdot \tau} \end{aligned}$$

We do not need any inequality constraints, and the upper bounds are automatically reflected in the equality constraints. Hence A, b and ub do not need to be specified. Now if x is the solution of this linear programming problem, it is easy to see that Γ given by $\Gamma_{i,j} = x_{(j-1)l+i}$ solves the EMD problem (3.14) subject to the constraints (3.15) - (3.17).

A.1.7 Gamma and Inverse Gamma Distributions

In this section, basic properties of the distributions used as hyperpriors in the concrete studies in Chapter 4 are summarized.
Both distributions are a two-parameter family of continuous probability distributions on the positive real line. Their density function is determined by a *shape* parameter $\alpha > 0$ and a *scale* parameter $\beta > 0$:

$$\begin{aligned} &\text{Gamma distribution:} && p(x; \alpha, \beta) = \frac{\beta^{-\alpha}}{\Gamma(\alpha)} \, x^{\alpha-1} \, \exp\left(-\frac{x}{\beta}\right) \\ &\text{Inverse gamma distribution:} && p(x; \alpha, \beta) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} \, x^{-\alpha-1} \exp\left(-\frac{\beta}{x}\right) \\ &\text{where} && \Gamma(z) = \int_0^\infty t^{z-1} e^{-t} \, dt \qquad \text{is the Gamma function.} \end{aligned}$$

In Figure A.1 both densities are plotted for different parameter values, in Table A.1 the main characteristics are listed. Gamma and inverse gamma distributions are commonly used to model scale variables such as the variance of a random process. They are closely connected, namely, if X follows a gamma distribution with shape α and scale β, then $Y := X^{-1}$ follows an inverse gamma distribution with shape α and scale β^{-1}, which explains the name of the inverse gamma distribution. While both seem quite similar at first glance, there are certain differences, which can lead to important differences in the behavior of HBM equipped with either one or the other:

- ⋆ Energy: The energy of the hyperprior (see Section 2.3) is one summand of the whole energy of the posterior (cf. (2.8)). The shape of the posterior energy is the central component which determines the properties of MAP and CM estimation. In Figure A.2 the energies of gamma and inverse gamma distribution are plotted for different parameter values. It reveals a central difference between both distributions: For the gamma distribution, a short calculation shows that the energy is either convex ($\alpha > 1$) or concave ($\alpha < 1$) on its whole domain. For the inverse gamma distribution, it is convex on the left and concave on the right side of its mode ($\beta/(\alpha + 1)$).

- ⋆ Outlier: The behavior of the limits $x \to \infty$ and $x \to 0$ are switched for gamma and inverse gamma distribution. As a result, the occurrence of outliers (realizations of a random variable that are numerically distant from the other realizations within a sample of finite size) substantially differs: The limit $x \to \infty$ is dominated by an exponential decay in the gamma distribution, and a power law in the inverse gamma distribution. As a result, the inverse gamma distribution bears way more probability weight in its right tail[2] and admits more outliers (see, e.g., Calvetti and Somersalo, 2008a for an illustration).

[2] One speaks of *heavy-tailed distributions*, as the tail is not exponentially bounded.

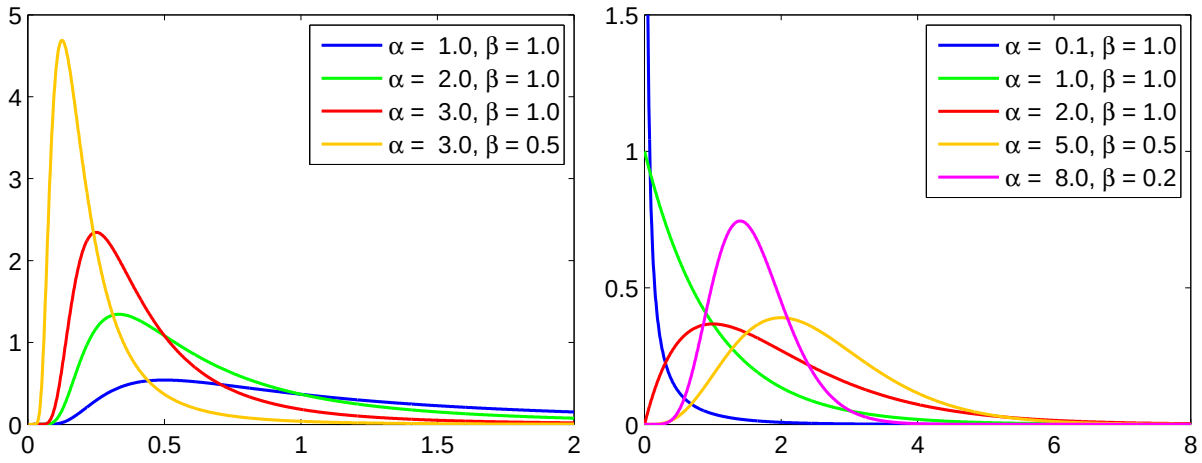


Figure A.1: Plots of the pdfs of inverse gamma (left) and gamma distribution (right).

Table A.1: Different characteristics of the gamma and the inverse gamma distribution

Type	Mean	Mode	Variance
Gam.	$\alpha\,\beta$	$(\alpha-1)\beta$ (for $\alpha \geqslant 1$)	$\alpha\,\beta^2$
Inv. Gam.	$\frac{\beta}{\alpha-1}$ (for $\alpha > 1$)	$\frac{\beta}{\alpha+1}$	$\frac{\beta^2}{(\alpha-1)^2(\alpha-2)}$ (for $\alpha > 2$)

⋆ Limits: Both distributions are often used to approximate a non-informative hyperprior of the form $p(x) \propto x^{-1}$. The limits $\alpha \to 0$, $\beta \to \infty$ for the gamma, and $\alpha, \beta \to 0$ for the inverse gamma distribution lead to this limiting distribution. As discussed in Gelman (2006), this might lead into a dilemma: The non-informative hyperprior is chosen to let the data determine the scale variable automatically and without requiring a-priori knowledge on it, especially to prevent that the estimation is sensitive to a scale manually predefined by the analysist. Still, the non-informative hyperprior is improper, and in contrast to some other improper hyperpriors, it leads to an improper posterior (Gelman, 2006; Nummenmaa et al., 2007a). When approximated by proper hyperpriors, this limiting improperness of the posterior leads to the phenomena that the estimates become very sensitive to the parameters of the approximating hyperpriors (Gelman, 2006). This is of course contradictory to intention of introducing the non-informative hyperprior which was to avoid a sensitivity to the manual choice of the hyperpriors parameters. Yet, Gelman (2006) only examined the inverse gamma hyperprior. It might be, that the gamma hyperprior yields a less sensitive approximation, since the way it approximates x^{-1} is substantially different from the way the inverse hyperprior does: The value of the inverse gamma hyperprior at the singularity $x = 0$ of the limiting distribution is always finite, while the gamma hyperprior is also singular for $\alpha < 1$. However, we cannot purse this issue any further here.

A.1.8 The Functionals behind AO-based MAP Approximation

In this section we demonstrate that in principle, a variety of well known optimization schemes for regularization based approaches to CDR can be assessed naturally by applying the AO scheme for MAP approximation to the HBM given in Section 4.2:

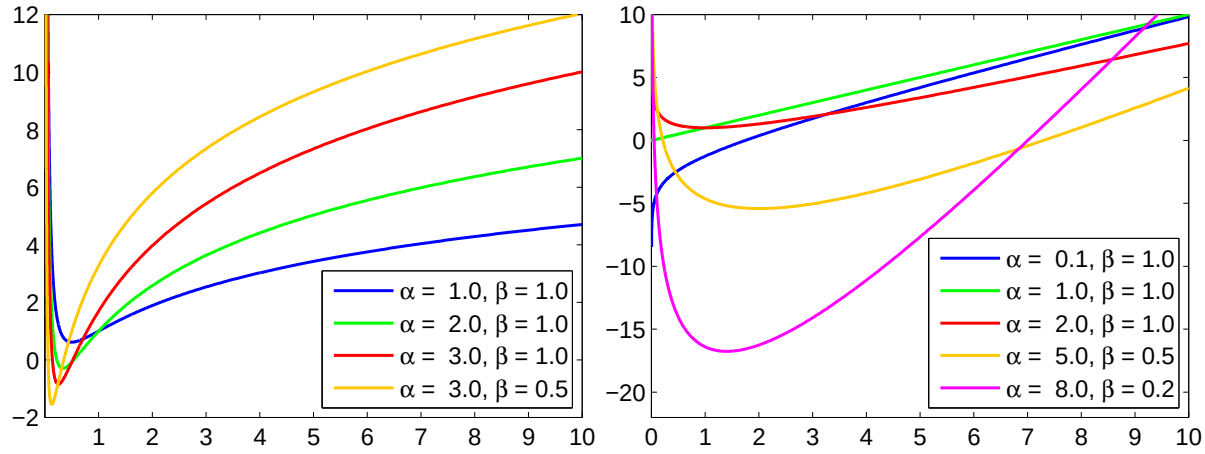


Figure A.2: Plots of the energy of inverse gamma (left) and gamma distributions (right).

Gamma Hyperprior: If we insert the particular update rule (4.6) into (3.5) we get a fixed point iteration for this problem:

$$\begin{aligned} s &= \operatorname*{argmin}_{s\in\mathbb{R}^n} \left\{ \|b - \mathrm{L}s\|^2 + \sigma^2 \sum_{i=1}^{k} \frac{\|s_{i*}\|^2}{\gamma_i} \right\} \\ &\overset{(4.6)}{=} \operatorname*{argmin}_{s\in\mathbb{R}^n} \left\{ \|b - \mathrm{L}s\|^2 + \frac{2\sigma^2}{\beta} \sum_{i=1}^{k} \frac{\|s_{i*}\|^2}{\eta + \sqrt{\eta^2 + \frac{\|s_{i*}\|^2}{\beta}}} \right\}, \quad \text{where} \quad \eta = (\alpha - 2.5) \end{aligned}$$

For the choice of $\alpha \searrow 2.5$, i.e., $\eta \searrow 0$, and $\lambda = \sigma^2\sqrt{2\beta^{-1}}$ this simplifies to:

$$s = \operatorname*{argmin}_{s\in\mathbb{R}^n} \left\{ \|b - \mathrm{L}s\|_2^2 + \sigma^2\sqrt{2\beta^{-1}} \sum_{i=1}^{k} \|s_{i*}\|_2 \right\} = \operatorname*{argmin}_{s\in\mathbb{R}^n} \left\{ \|b - \mathrm{L}s\|_2^2 + \lambda \|s_*\|_1 \right\}$$

where $\|s_*\|_1$ is the ℓ_1-norm of the source amplitudes $\|s_{i*}\|_2$. This type of regularization is called *minimum current estimate* (*MCE*, see Matsuura and Okabe, 1995; Uutela et al., 1999) in the context of EEG/MEG. The AO scheme is now essentially similar to a special variant of the FOCUSS algorithm (Gorodnitsky and Rao, 1997). In Wipf et al. (2007) this connection of MCE and FOCUSS solution has also been pointed out from the *empirical Bayesian* point of view. By choosing $\alpha > 2.5$ one avoids the problem of dividing by components s_i near to zero, and obtains a regularized algorithm for finding the MCE.

Generalized Gamma Hyperprior: Remind that the gamma distribution corresponded to the case $\zeta = 1$ for the generalized gamma distribution (cf. (4.1)). For p with $0 < p < 2$ a similar reasoning as above with the generalized gamma distribution with $\zeta = p/(2-p)$ will result in:

$$s = \operatorname*{argmin}_{s\in\mathbb{R}^n} \left\{ \|b - \mathrm{L}s\|_2^2 + \lambda \|s_*\|_p^p \right\}, \quad \text{with} \quad \lambda = \sigma^2 \left(\frac{2\zeta}{\beta\zeta} \right)^{1/(\zeta+1)}$$

For details, see Calvetti et al. (2009).

Inverse Gamma Hyperprior: If we insert the particular update rule (4.7) into (3.5) we get a fixed point iteration for this problem:

$$s = \operatorname*{argmin}_{s\in\mathbb{R}^n} \left\{ \|b - \mathrm{L}s\|^2 + 2\sigma^2(\alpha + 1.5) \sum_{i=1}^{k} \frac{\|s_{i*}\|^2}{\|s_{i*}\|^2 + 2\beta} \right\} \tag{A.11}$$

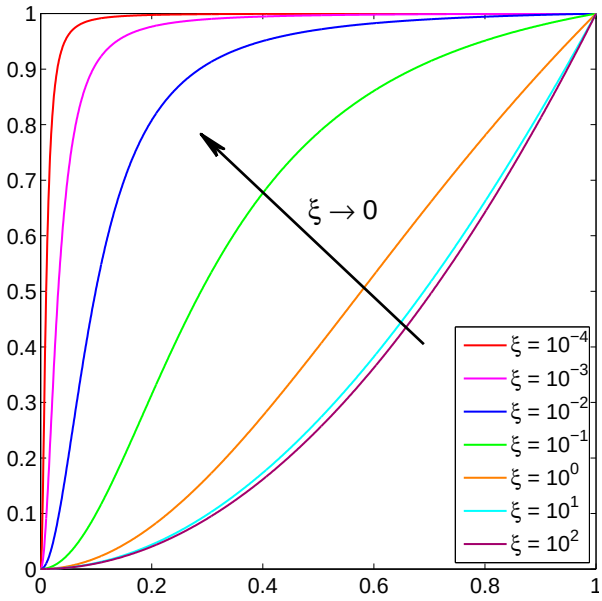


Figure A.3: Scaled minimum support stabilizer functional $(\xi+1)x^2/(x^2+\xi)$ for different values of ξ.

The regularization term $x^2/(x^2+\xi)$ is called minimum support stabilizer as it converges to $\|x\|_0$ for $\xi \searrow 0$ and was used in EEG/MEG for the definition of the *minimum support estimate* (*MSE*) (Nagarajan et al., 2006). Depending on the choice of β, it takes a different shape (and for a fixed β, α determines the relative weight of the regularizer). To illustrate this, Figure A.3 shows the shape of the function $(\xi+1)x^2/(x^2+\xi)$ for different values of β (the factor $(\xi+1)$ was just included for scaling). The function always starts with a quadratic-like convex part, until it passes to a concave part in which it ends up in a plateau. Varying α and β determines, which part of the minimum support stabilizer is of practical relevance for the problem (A.11): If the convex dominates for reasonable values of s_i, the solution is alike to those obtained by MNE, and (practically) unique. If the concave part dominates, the solution will be focal but non-unique, and the mode found by the AO scheme will be very sensitive to the initialization. Further relations between estimates in hierarchical models and functionals and optimization schemes have been shown: See Calvetti and Somersalo (2008a) for the relation of Total Variation (Rudin et al., 1992) and Perona–Malik (Perona and Malik, 1990) penalties to MAP approximation with Gaussian noise and Bardsley et al. (2010) for an extension to Poisson noise. Wipf and Nagarajan (2009) outline how to derive FOCUSS, MCE and sLORETA (Pascual-Marqui, 2002) within another estimation framework than the full-MAP scheme discussed here, namely the framework of γ-MAP and S-MAP estimation (see Section 2.4.2). Note however that there are estimators whose approximation methods cannot be recovered from another framework, like the CM-estimator for the model considered here.

A.1.9 The Student's T-distribution as an Implicit Prior on the Source Amplitudes

In this section we show that using an inverse gamma hyperprior in the HBM given in Section 4.2 leads to the Student's t-distribution as an implicit prior on the (scaled) source amplitudes $\|s_{i*}\|$: We start off by using 2.5 to compute the implicit prior on S:

$$p(s) = \int_{\mathbb{R}^h} p(s|\boldsymbol{\gamma})\, p(\boldsymbol{\gamma})\, d\boldsymbol{\gamma} \overset{4.8}{\propto} \prod_i^h \int_{\mathbb{R}} \exp\left(-\frac{\frac{1}{2}\|s_{i*}\|^2 + \beta}{\gamma_i} + (-(\alpha+3/2)-1)\ln(\gamma_i) \right)$$

As noted in Section 4.2, the integrand is proportional to an inverse gamma distribution with parameters $\bar{\beta}_i = \frac{1}{2}\|s_{i*}\|^2 + \beta$ and $\bar{\alpha}_i = (\alpha + 3/2)$, and therefore the integral has to be given

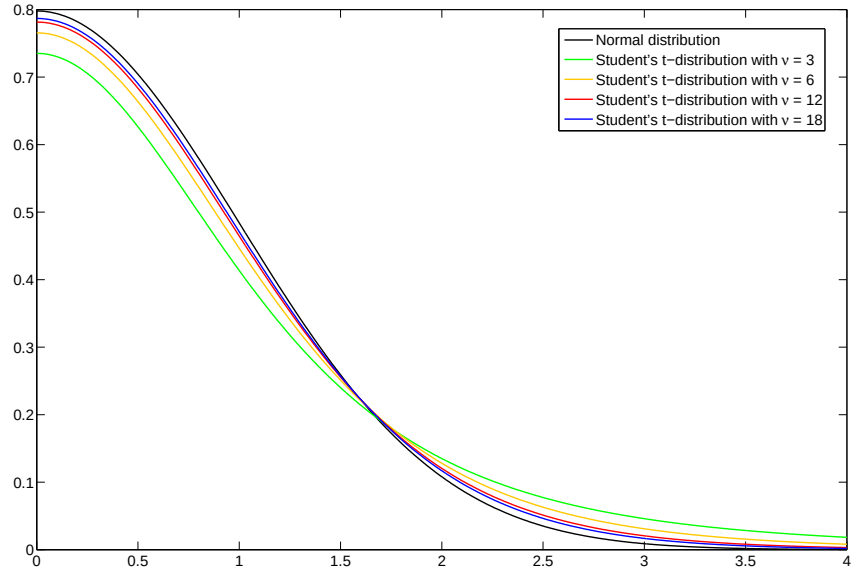


Figure A.4: Plots of the pdfs of the (one-sided) Student's t-distribution vs. the normal distribution for different values of ν.

by the normalization of the corresponding distribution. Comparing the expression with the corresponding terms in Section A.1.7, this leads to

$$\prod_i^h \int_{\mathbb{R}} \exp\left(-\frac{\frac{1}{2}\|s_{i*}\|^2+\beta}{\gamma_i} + (-(\alpha+3/2)-1)\ln(\gamma_i)\right) = \prod_i^h \frac{\Gamma(\bar{\alpha}_i)}{\bar{\beta}_i^{\bar{\alpha}_i}} \propto \prod_i^h \left(\tfrac{1}{2}\|s_{i*}\|^2+\beta\right)^{-(\alpha+3/2)}$$

$$\propto \prod_i^h \left(1+\frac{t^2}{\nu}\right)^{-\frac{1}{2}(\nu+1)} \quad \text{with} \quad \nu := 2(\alpha+1), \quad t := \frac{\|s_{i*}\|}{\sqrt{\gamma_{mode}}}, \quad \gamma_{mode} := \frac{\beta}{\alpha+1} \quad \text{(cf. A.1.7)}$$

This is a Student's t-distribution for the scaled source amplitudes t with the *degrees of freedom* ν (Gelman et al., 2003). It commonly arises in Bayesian inference when assuming normally distributed parameters of interest with unknown variance that is inverse gamma distributed (Gelman et al., 2003). More precisely, it is the positive part of it, since $t \geqslant 0$ by definition. Figure A.4 shows plots of its pdf for different values of α in comparison to a one-sided standard normal distribution. For $\alpha \to \infty$, i.e., $\nu \to \infty$ it approaches the standard normal distribution. Even so, for small α the scaling of both distributions differs considerably: Since the normal distribution decays exponentially for $t \to \infty$ it prohibits large outliers and imposes a typical length scale for the scaled source amplitude t. A standard normal prior on t hence promotes non-focal source activity, which is apparent, since it is equivalent to a uniform diagonal Gaussian prior on S and thus to a MNE with a certain regularization parameter λ (cf. Section 2.3). The Student's t-distribution on the other hand is a heavy-tailed distribution, which decays like a power law for $t \to \infty$ (cf. Section A.1.7). As a consequence, it allows for large outliers and promotes focal source activity.

A.1.10 Computation Time

In this section, the computation times for different implementations of the AS_CM scheme are compared. Since the other methods rely on the same schemes, the results can be transferred. For all testing scenarios, the burn-in size Q was set to 50 and the sample size R to 300. The same parameters as in the main studies in Chapter 4 were used. A noise level of 5% was assumed. The effect of using the implicit multi-threading capabilities of Matlab was also examined by using 1,2 or 4 cores of a 4 core CPU system (Intel Core 2 Quad @ 2.83 GHz)

Table A.2: Mean computation times (sec) of the AS_CM scheme for different implementations of the Ss step.

	$\bar{N}_s$	**1 core**	**2 cores**	**4 cores**
Analytical	100	4.23 ± 0.02	2.93 ± 0.09	2.36 ± 0.02
Adapted CGLS	100	9.31 ± 1.06	5.36 ± 0.80	5.88 ± 0.87
Standard CGLS	20	21.28 ± 1.60	21.71 ± 1.56	21.82 ± 1.94
LSQR	20	22.86 ± 1.82	22.84 ± 2.31	22.80 ± 1.92
mldivide	2	208.08 ± 0.01	151.36 ± 0.01	123.70 ± 0.05

Table A.3: Mean computation times (sec) *per right hand side* using the blocked inversion scheme.

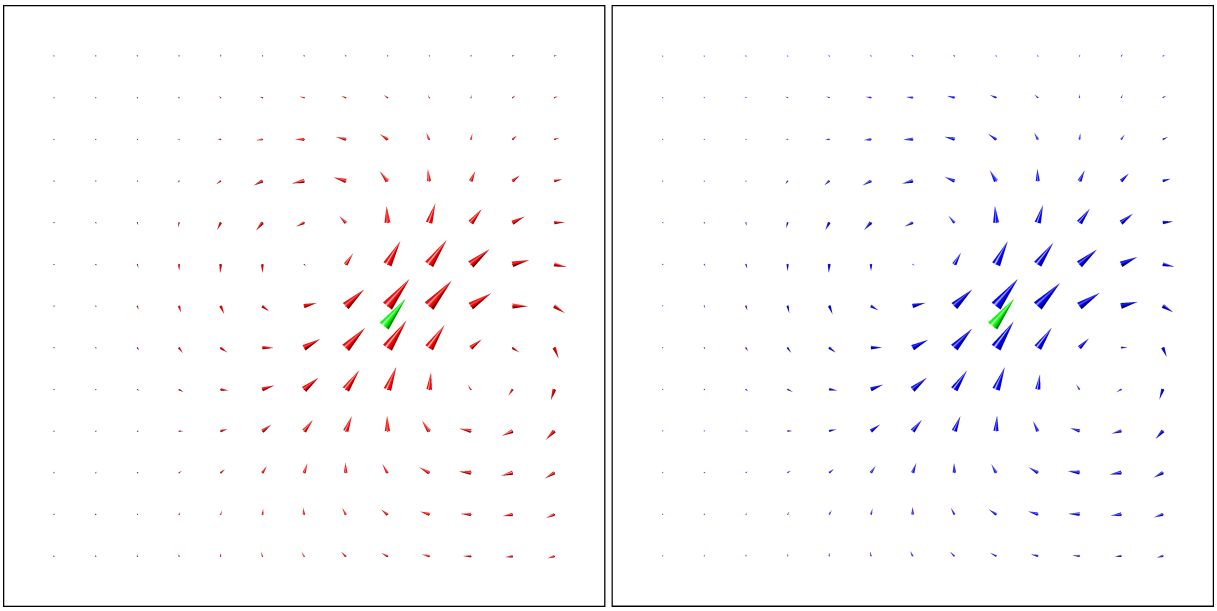
N_b	$\bar{N}_{rep}$	**1 core**	**2 cores**	**4 cores**
2	300	8.78	6.70	8.04
4	200	6.73	4.83	4.84
16	100	3.79	2.83	2.65
64	25	3.28	2.58	2.34
256	10	3.59	2.90	2.66
1024	5	3.71	2.98	2.75

First, the single inversion scheme was tested with different implementations of the Ss step (cf. 3.4.1):

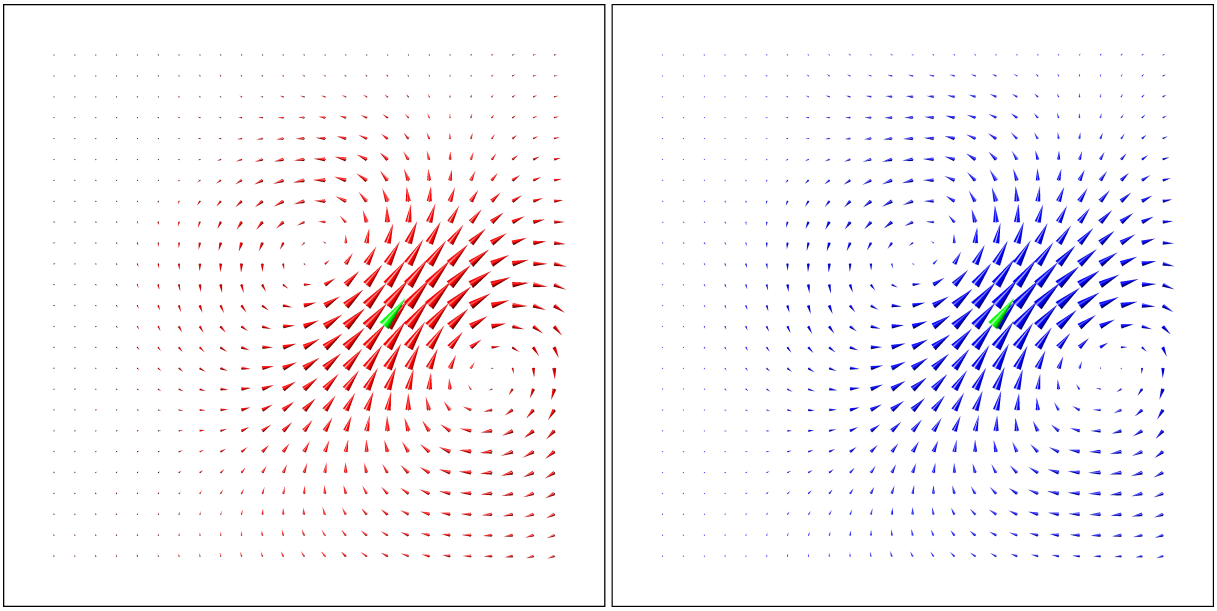
1. Algorithm 6 to compute the analytical solution.
2. An adapted implementation of the CGLS algorithm (cf. Section 3.5).
3. A standard implementation of the CGLS algorithm (cf. Section 3.5).
4. The LSQR algorithm implemented by Matlab (cf. Section 3.5).
5. The back slash operator of Matlab (mldivide).

For the implementations 2.-5. the preconditioning in the form of (3.7) and the sparse matrix format by Matlab are used. Table A.2 lists the mean computation times, averaged over an implementation-dependent number of single dipole sources $\bar{N}_s$ and the corresponding standard deviations. As expected, implementations based iterative solvers (CGLS and LSQR) show a larger variation. The results also given an impression how long a typical inversion with AS_CM based method like cmAO_MAP takes: If $Q = 1\,000$ and $M = 50\,000$ are used, Algorithm 6 takes 5 - 6 minutes on a modern 4 core CPU architecture. As a second study, the performance of the blocked inversion scheme is evaluated (cf. Section 3.6) for a different number of measurements N_b that are inverted simultaneously. Note that here only the adapted implementation of the CGLS algorithm can be used to compute the (blocked) Ss step. Table A.3 lists the mean computation times over $\bar{N}_{rep}$ repetitions and divided by N_b.
The implementations that have been developed and optimized within the work for this thesis (i.e., the analytical approach and the single and blocked adapted CGLS algorithm) exploit the characteristics of the problem and hence clearly outperform the others. The results also suggest not all implementations benefit from parallelization in the same way, and that this has to be examined more carefully.

A.2 Figures



(a) Left: MNE on grid 2 (k = 169). Right: Cubic Interpolation of MNE on grid 1 (k = 49) to grid 2. The maximal error in a single component between the (normalized) MNE and the (normalized) interpolation is 0.034.



(b) Left: MNE on grid 3 (k = 625). Right: Cubic Interpolation of MNE on grid 1 (k = 49) to grid 3. The maxmial error in a single component between the (normalized) MNE and the (normalized) interpolation is 0.002.

Figure A.5: Increasing source grid resolution vs. interpolation for MNE. The spatial distribution of the interpolation error suggests that the main cause for the error are boundary effects.

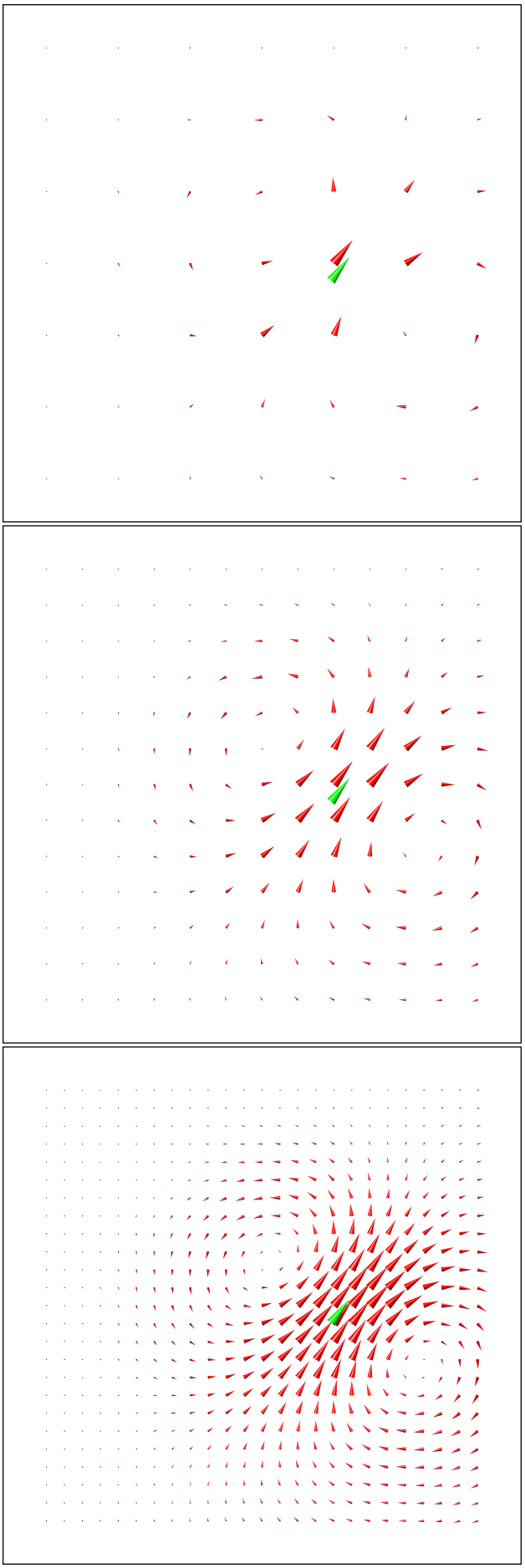


Figure A.6: Figure 2.1 in higher resolution

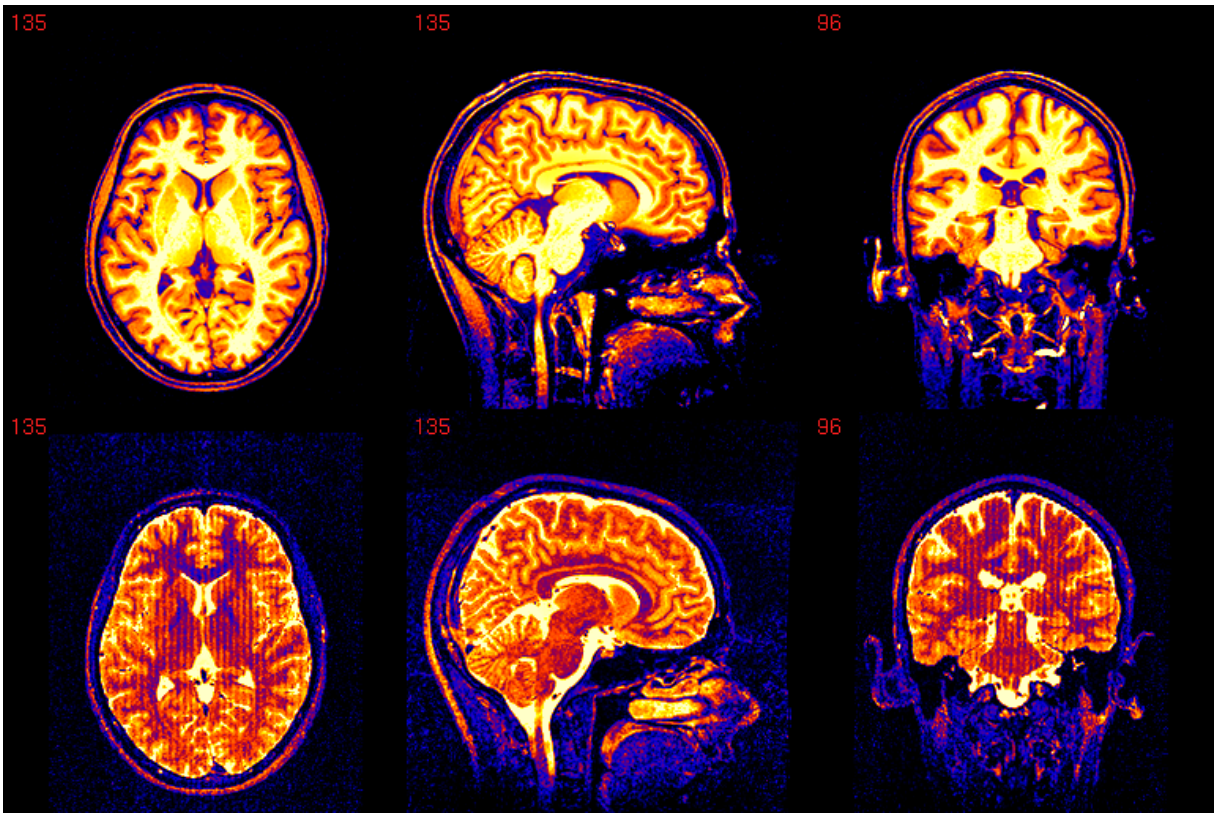


Figure A.7: Transversal, sagittal and coronal slices from the T1 (upper row) and T2 weighted (bottom row) MRI images (resolution: 1 mm, 256x256x256 voxel)

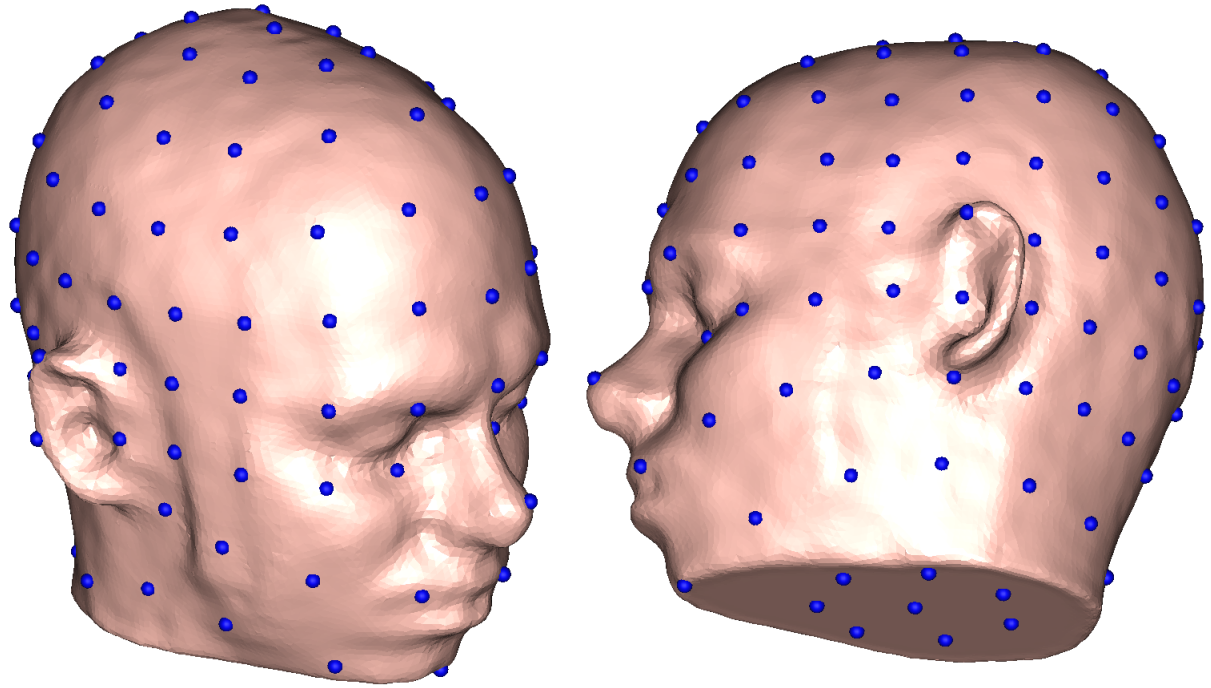


Figure A.8: Artificial full coverage EEG sensor configuration consisting of 134 EEG sensors: The sensors were placed uniformly on the surface of a sphere around the center of the model and were then projected onto the head surface

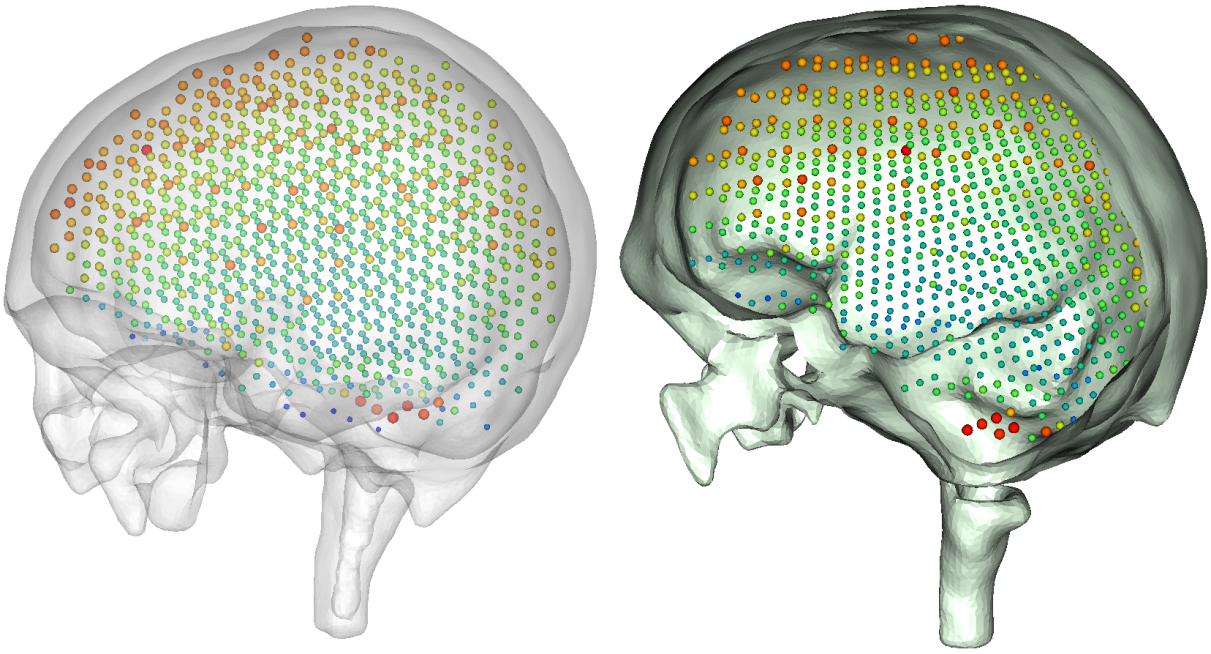


Figure A.9: The sum of the ℓ_2 norms of the three gain-vectors at a given position is depicted. The influence of the hole at the base of the skull (foramen magnum) on the magnitudes of the deep-lying sources is noticeable (this feature occurs with realistic sensor configurations as well)

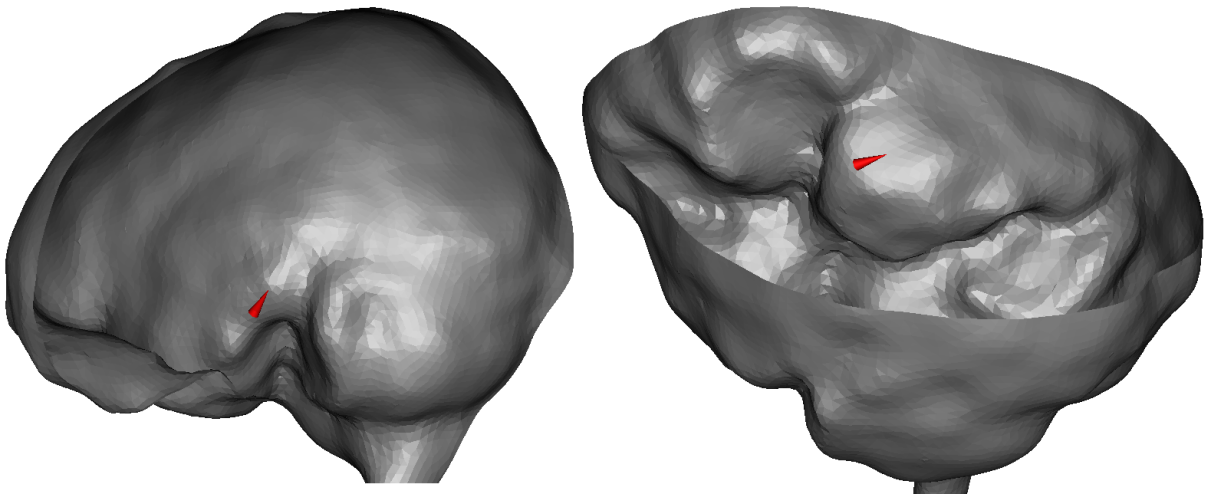


Figure A.10: Dipole used for multimodality illustration

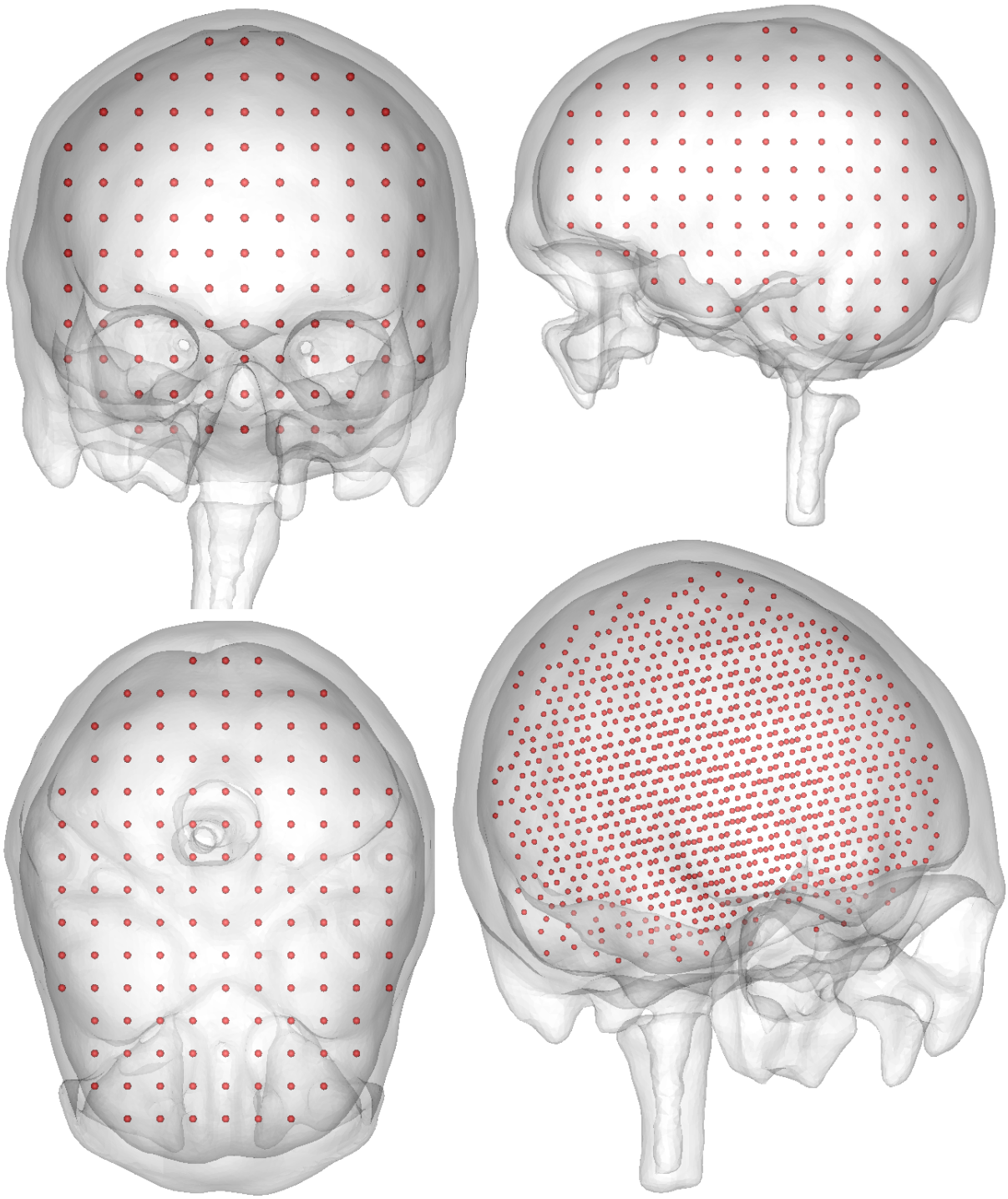


Figure A.11: The locations of the 1000 source space nodes that were constructed in the following way: The nodes of the gray matter surface are clipped below a certain z-value to exclude the brain stem volume, and the convex hull of the remaining nodes is constructed and slightly contracted. Of all FEM nodes only the ones within the resulting surface are labeled as active. After that, all of these nodes that do not fulfill a condition related to the approach used for forward computation (the Venant approach) are delabeled again. A regular grid is laid through the whole volume, and every grid node whose nearest FEM node is labeled is accepted. If the number of accepted grid nodes matches the desired number of source nodes, the locations are fixed, and the lead-field is computed. If not, an automatic procedure adapts grid size and offset of the grid until the desired grid is found.

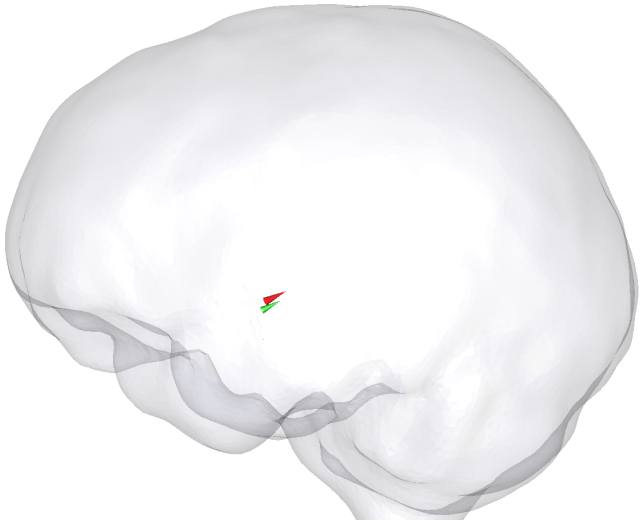


Figure A.12: AS_CM approximation (red-yellow cones) for a single dipole (green cone).

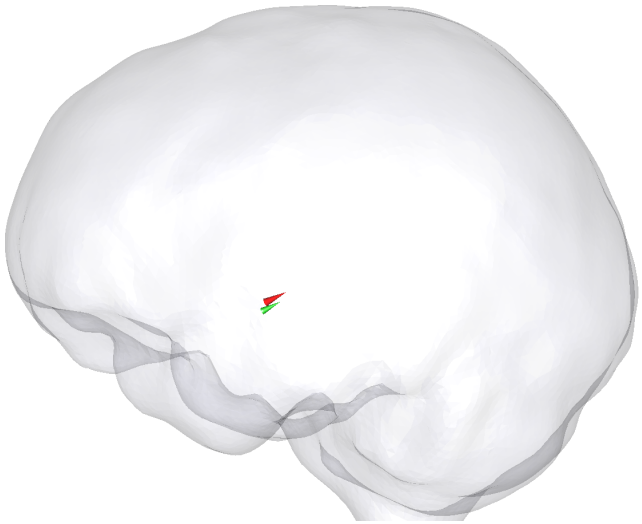


Figure A.13: cmAO_MAP approximation (red-yellow cones) for a single dipole (green cone).

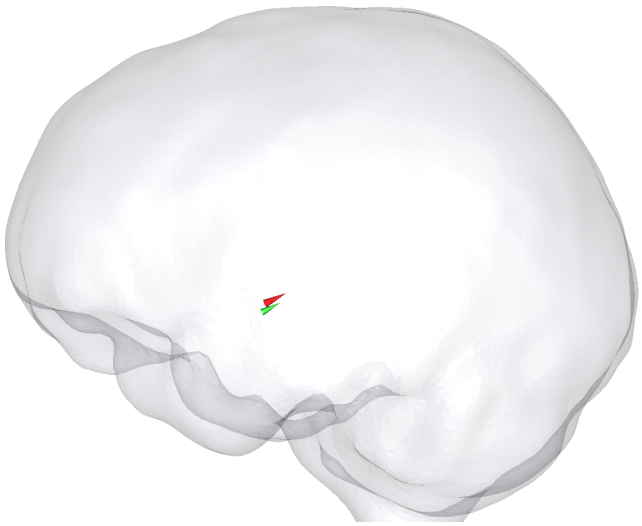


Figure A.14: McmAO_MAP approximation (red-yellow cones) for a single dipole (green cone).

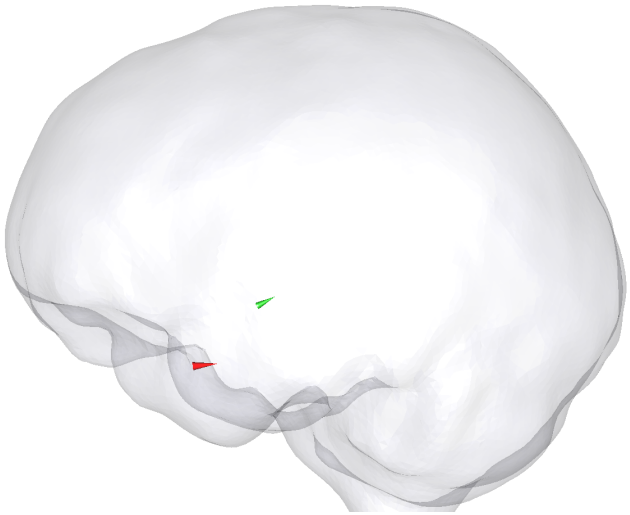


Figure A.15: uAO_MAP approximation (red-yellow cones) with inverse gamma hyperprior for a single dipole (green cone).

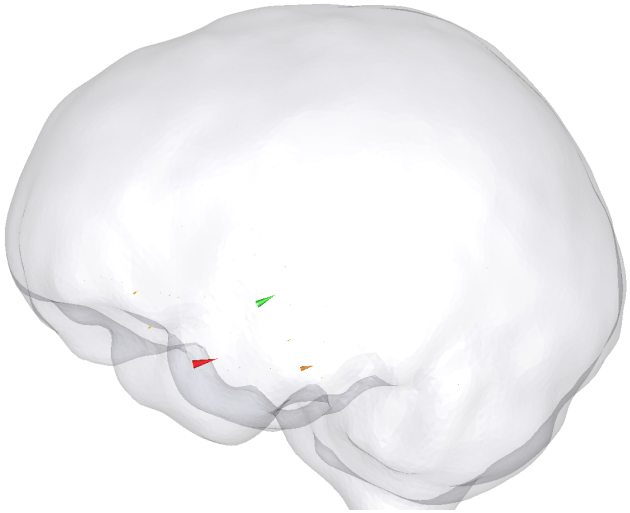


Figure A.16: uAO_MAP approximation (red-yellow cones) with gamma hyperprior for a single dipole (green cone).

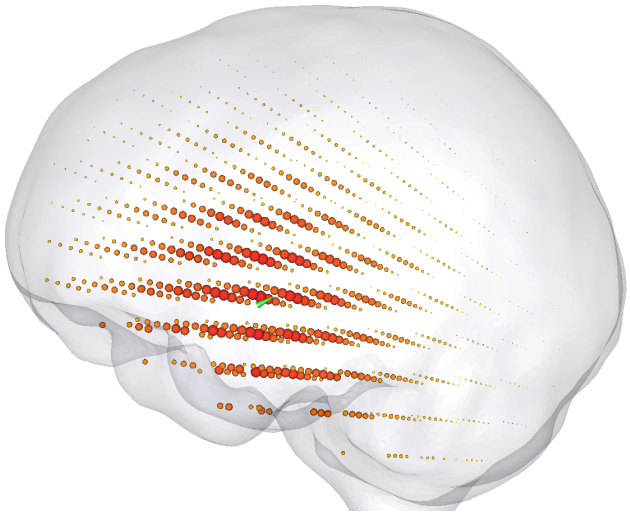


Figure A.17: sLORETA result (red-yellow spheres) for a single dipole (green cone).

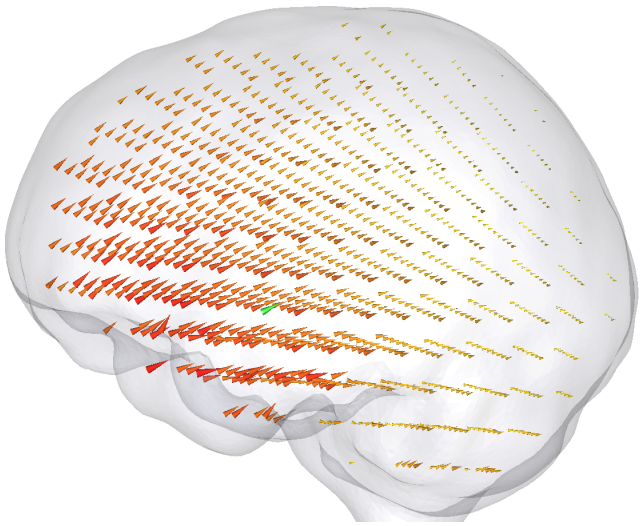


Figure A.18: MNE result (red-yellow cones) for a single dipole (green cone).

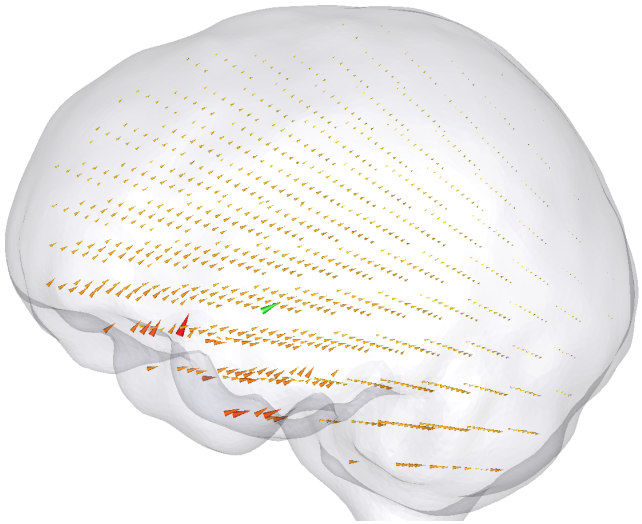


Figure A.19: WMNE result (red-yellow cones) with ℓ_2 weighting for a single dipole (green cone).

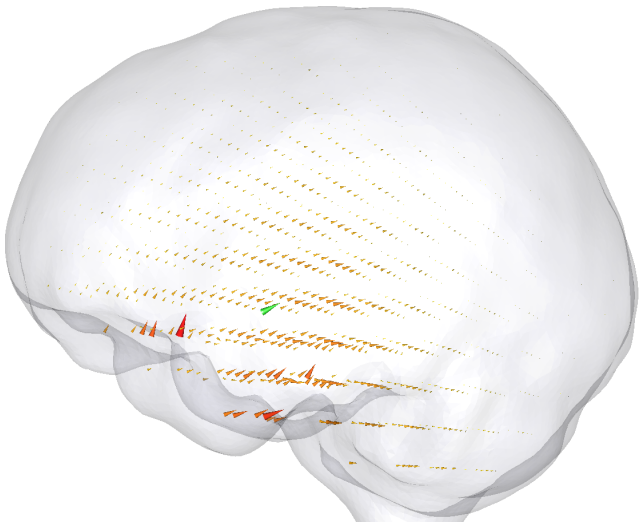
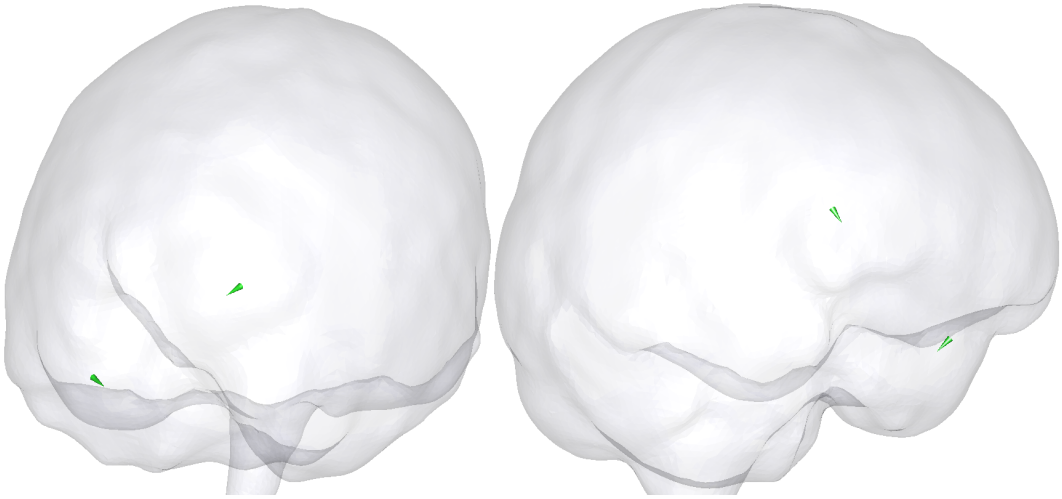
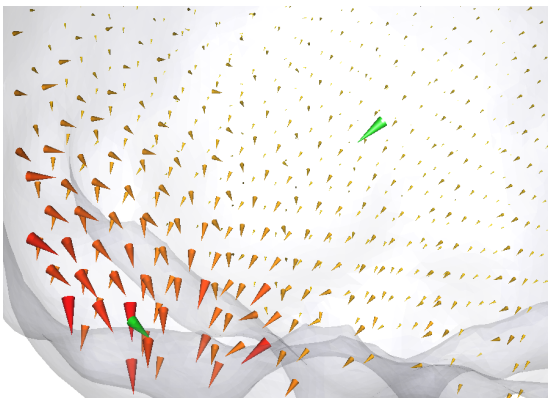


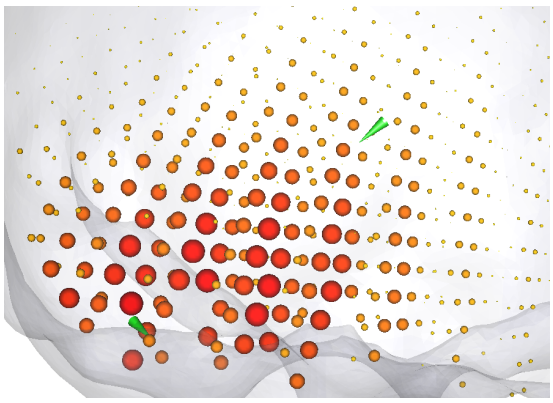
Figure A.20: WMNE result (red-yellow cones) with regularized ℓ_∞ weighting for a single dipole (green cone).



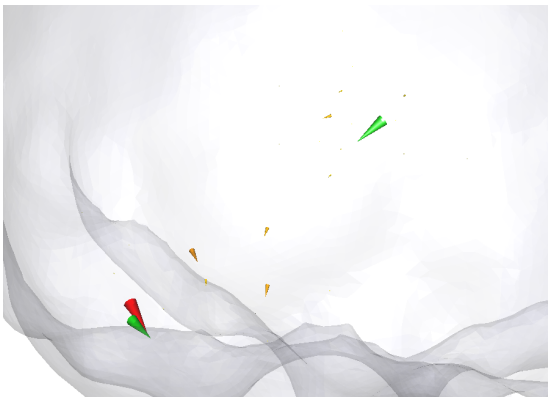
(a) Two sources chosen for visualization.



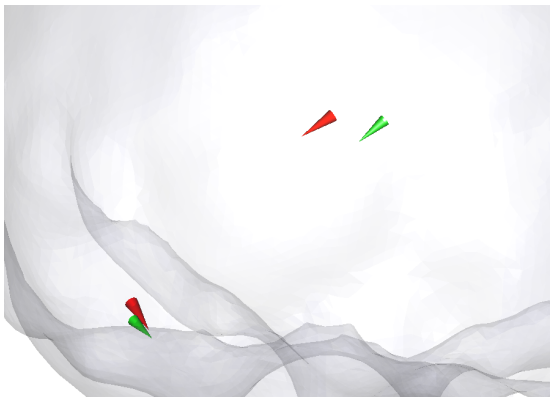
(b) MNE



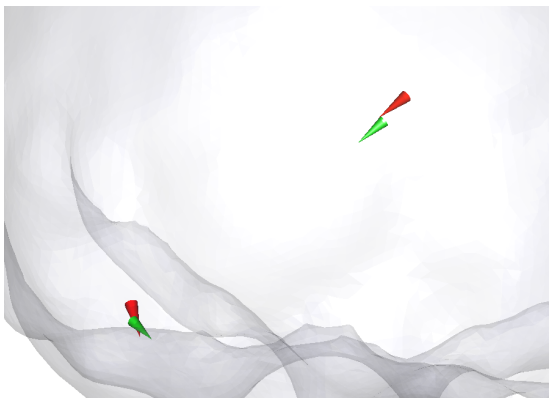
(c) sLORETA



(d) AS_CM, $M = 200\,000$



(e) cmAO_MAP, $M = 200\,000$



(f) McmAO_MAP

Figure A.21: An Example for the masking of deep-lying sources.

Bibliography

Ahlfors, S. P., Ilmoniemi, R. J., and Hämäläinen, M. (1992). Estimates of visually evoked cortical currents. *Electroencephalogr Clin Neurophysiol*, 82(3):225–36.

Ambrosio, L., Gigli, N., and Savaré, G. (2008). *Gradient Flows in Metric Spaces and in the Spaces of Probability Measures*. Birkhauser.

Andersen, P. (2007). *The Hippocampus Book*. Oxford University Press, USA.

Aubert, G. and Kornprobst, P. (2006). *Mathematical Problems in Image Processing*, volume 147 of *Applied Mathematical Sciences*. Springer, 2nd edition.

Bardsley, J., Calvetti, D., and Somersalo, E. (2010). Hierarchical regularization for edge-preserving reconstruction of PET images. *Inverse Problems*, 26:035010.

Ben-Israel, A. and Greville, T. N. E. (2003). *Generalized Inverses : Theory and Applications*. Springer, New York, 2nd edition.

Bernstein, D. (2009). *Matrix Mathematics: Theory, Facts, and Formulas*. Princeton Univ Pr.

Björck, Å. (1996). *Numerical Methods for Least Squares Problems*. SIAM, Philadelphia.

Braess, D. (2007). *Finite Elements: Theory, Fast Solvers, and Applications in Solid Mechanics*. Cambridge University Press, 3rd edition.

Brazier, M. A. B. (1949). A study of the electric field at the surface of the head. *Electroencephalography and Clinical Neurophysiology*, pages 38–52.

Brenner, S. C. and Scott, R. L. (2008). *The Mathematical Theory of Finite Element Methods*. Springer, New York, 3nd edition.

Calvetti, D., Hakula, H., Pursiainen, S., and Somersalo, E. (2009). Conditionally Gaussian hypermodels for cerebral source localization. *SIAM J. Imaging Sci.*, 2(3):879–909.

Calvetti, D. and Somersalo, E. (2007a). A Gaussian hypermodel to recover blocky objects. *Inverse Problems*, 23(2):733–754.

Calvetti, D. and Somersalo, E. (2007b). *Introduction to Bayesian Scientific Computing*, volume 2 of *Surveys and Tutorials in the Applied Mathematical Sciences*. Springer New York.

Calvetti, D. and Somersalo, E. (2008a). Hypermodels in the Bayesian imaging framework. *Inverse Problems*, 24(3):034013 (20pp).

Calvetti, D. and Somersalo, E. (2008b). Recovery of shapes: Hypermodels and Bayesian learning. In *Journal of Physics: Conference Series*, volume 124, page 012014. IOP Publishing.

Chang, B. S. and Lowenstein, D. H. (2003). Epilepsy. *N Engl J Med*, 349(13):1257–66.

Cuffin, B. (1996). EEG localization accuracy improvements using realistically shaped head models. *IEEE Transactions on Biomedical Engineering*, 43(3):1.

Dale, A. M., Liu, A. K., Fischl, B., Buckner, R., 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(1):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. *J. Cogn. Neurosci*, 5:162–176.

Dannhauer, M., Knösche, T. R., Lanfer, B., and Wolters, C. H. (2009). Skull tissue conductivity modeling affects forward and inverse solution: an EEG simulation study across subjects. *NeuroImage*, 47(Supplement 1):S74–S74. Organization for Human Brain Mapping 2009 Annual Meeting.

Dannhauer, M., Lanfer, B., Wolters, C. H., and Knösche, T. R. (2010). Modeling of the human skull in EEG source analysis. *Hum Brain Mapp*.

De Munck, J. (1988). The potential distribution in a layered anisotropic spheroidal volume conductor. *Journal of applied Physics*, 64(2):464–470.

de Munck, J. C., van Dijk, B. W., and Spekreijse, H. (1988). Mathematical dipoles are adequate to describe realistic generators of human brain activity. *IEEE Trans Biomed Eng*, 35(11):960–6.

Dempster, A., Laird, N., Rubin, D., et al. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38.

Drechsler, F., Wolters, C. H., Dierkes, T., Si, H., and Grasedyck, L. (2009). A full subtraction approach for finite element method based source analysis using constrained Delaunay tetrahedralisation. *NeuroImage*, 46(4):1055–1065.

Duvernoy, H. (2005). *The Human Hippocampus: Functional Anatomy, Vascularization, and Serial Sections with MRI*. Springer Verlag.

Engl, H., Hanke-Bourgeois, M., and Neubauer, A. (1996). *Regularization of Inverse Problems*. Mathematics and its applications. Kluwer Acad. Publ.

Friston, K. J., Glaser, D. E., Henson, R. N., Kiebel, S. J., Phillips, C., and Ashburner, J. (2002a). Classical and Bayesian inference in neuroimaging: applications. *Neuroimage*, 16(2):484–512.

Friston, K. J., Harrison, L., Daunizeau, J., Kiebel, S. J., Phillips, C., Trujillo-Barreto, N. J., Henson, R. N., Flandin, G., and Mattout, J. (2008). Multiple sparse priors for the M/EEG inverse problem. *Neuroimage*, 39(3):1104–20.

Friston, K. J., Mattout, J., Trujillo-Barreto, N. J., Ashburner, J., and Penny, W. D. (2007). Variational free energy and the Laplace approximation. *Neuroimage*, 34(1):220–34.

Friston, K. J., Penny, W. D., Phillips, C., Kiebel, S. J., Hinton, G., and Ashburner, J. (2002b). Classical and Bayesian inference in neuroimaging: theory. *Neuroimage*, 16(2):465–83.

Fuchs, M., Wagner, M., Köhler, T., and Wischmann, H.-A. (1999). Linear and nonlinear current density reconstructions. *Journal of clinical Neurophysiology*, 16(3):267.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis*, 1(3):515–533.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2003). *Bayesian Data Analysis*. Texts in Statistical Science. Chapman and Hall/CRC, 2nd edition.

Gelman, A., Van Dyk, D. A., Huang, Z., and Boscardin, J. W. (2007). Transformed and Parameter-expanded Gibbs Samplers for Multilevel Linear and Generalized Linear Models. Technical report, Department of Statistics, Columbia University.

Gencer, N. and Williamson, S. (2002). Differential characterization of neural sources with the bimodal truncated SVD pseudo-inverse for EEG and MEG measurements. *Biomedical Engineering, IEEE Transactions on*, 45(7):827–838.

Gencer, N. G. and Williamson, S. J. (1998). Differential characterization of neural sources with the bimodal truncated SVD pseudo-inverse for EEG and MEG measurements. *IEEE Trans Biomed Eng*, 45(7):827–38.

Gorodnitsky, I. F., George, J. S., and Rao, B. D. (1995). Neuromagnetic source imaging with FOCUSS: a recursive weighted minimum norm algorithm. *Electroencephalogr Clin Neurophysiol*, 95(4):231–51.

Gorodnitsky, I. F. and Rao, B. D. (1997). Sparse signal reconstruction from limited data using FOCUSS: A re-weighted minimum norm algorithm. *IEEE Trans. Signal Processing*, pages 600–616.

Grave de Peralta, R., Hauk, O., and Gonzalez, S. L. (2009). The neuroelectromagnetic inverse problem and the zero dipole localization error. *Comput Intell Neurosci*, page 659247.

Greenblatt, R. E., Ossadtchi, A., and Pflieger, M. E. (2005). Local Linear Estimators for the Bioelectromagnetic Inverse Problem. *IEEE Transactions on Signal Processing*, 53(9):3403–3412.

Hackbusch, W. (1997). *Integralgleichungen. Theorie und Numerik.* Teubner, Stuttgart, 2nd edition.

Hadamard, J. (1923). *Lectures on Cauchy's Problem in Linear Partial Differential Equations.* New Haven.

Hallez, H. (2008). *Incorporation of Anisotropic Conductivities in EEG Source Analysis.* PhD thesis, Faculteit Ingenieurswetenschappen, Universiteit Gent, Belgium.

Hämäläinen, M., Haario, H., and Lehtinen, M. (1987). Inference about sources of neuromagnetic fields using Bayesian parameter estimation. *Preprint, TKK-F-A620.*

Hämäläinen, M., 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. *Rev. Mod. Phys.*, 65(2):413–497.

Hämäläinen, M. and Ilmoniemi, R. J. (1984). Interpreting measured magnetic fields of the brain: minimum norm estimates of current distributions. *Helsinki University of Technology, Technical Report TKK-F-A559.*

Hämäläinen, M. and Ilmoniemi, R. J. (1994). Interpreting magnetic fields of the brain: minimum norm estimates. *Med Biol Eng Comput*, 32(1):35–42.

Hämäläinen, M. and Sarvas, J. (1989). Realistic conductivity geometry model of the human head for interpretation of neuromagnetic data. *IEEE Transactions on Biomedical Engineering*, 36(2):165–171.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* Mathematics and Statistics. Springer New York, 2nd edition.

Helin, T. (2010a). *Discretization and Bayesian Modeling in Inverse Problems and Imaging*. PhD thesis, Aalto University School of Science and Technology.

Helin, T. (2010b). On infinite-dimensional hierarchical probability models in statistical inverse problems. *Inverse Problems and Imaging*, 3:567–597.

Helin, T. and Lassas, M. (2009). Hierarchical Models in Statistical Inverse Problems and the Mumford–Shah Functional. Technical Report arXiv:0908.3396. Comments: 31 pages, 2 figures.

Henson, R. N., Flandin, G., Friston, K. J., and Mattout, J. (2010). A Parametric Empirical Bayesian framework for fMRI-constrained MEG/EEG source reconstruction. *Hum Brain Mapp*.

Henson, R. N., Mattout, J., Phillips, C., and Friston, K. J. (2009a). Selecting forward models for MEG source-reconstruction using model-evidence. *Neuroimage*, 46(1):168–76.

Henson, R. N., Mouchlianitis, E., and Friston, K. J. (2009b). MEG and EEG data fusion: simultaneous localisation of face-evoked responses. *Neuroimage*, 47(2):581–9.

Hillebrand, A., Singh, K. D., Holliday, I. E., Furlong, P. L., and Barnes, G. R. (2005). A new approach to neuroimaging with magnetoencephalography. *Hum Brain Mapp*, 25(2):199–211.

Ioannides, A. A., Bolton, J. P. R., and Clarke, C. J. S. (1990). Continuous probabilistic solutions to the biomagnetic inverse problem. *Inverse Problems*, 6(4):523.

Jackson, J. D. (1998). *Classical Electrodynamics*. Wiley, 3rd edition.

Jaynes, E. and Bretthorst, G. (2003). *Probability Theory: The Logic of Science*. Cambridge Univ Pr.

Jun, S. C., George, J. S., Kim, W., Paré-Blagoev, J., Plis, S. M., Ranken, D. M., and Schmidt, D. M. (2008). Bayesian brain source imaging based on combined MEG/EEG and fMRI using MCMC. *Neuroimage*, 40(4):1581–94.

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

Kaipio, J. P. and Somersalo, E. (2007). Statistical inverse problems: Discretization, model reduction and inverse crimes. *Journal of Computational and Applied Mathematics*, 198(2):493–504. Applied Computational Inverse Problems.

Kantorovich, L. (1942). On the translocation of masses, CR (Doklady) Acad. *Sci. URSS (NS)*, 37:199–201.

Kantorovich, L. and Gavurin, M. (1949). The application of mathematical methods in problems of freight flow analysis. *Collection of Problems Concerned with Increasing the Effectiveness of Transports, Publication of the Akademii Nauk SSSR, Moscow-Leningrad*, pages 110–138.

Klenke, A. (2008). *Probability Theory: A Comprehensive Course*. Springer Verlag.

Kybic, J., Clerc, M., Abboud, T., Faugeras, O., Keriven, R., and Papadopoulo, T. (2005). A common formalism for the integral formulations of the forward EEG problem. *IEEE Trans Med Imaging*, 24(1):12–28.

Lassas, M., Saksman, E., and Siltanen, S. (2009). Discretization-invariant Bayesian Inversion and Besov Space Priors. Technical Report arXiv:0901.4220.

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

Lin, F.-H., Witzel, T., Ahlfors, S. P., Stufflebeam, S. M., Belliveau, J. W., and Hämäläinen, M. (2006). Assessing and improving the spatial accuracy in MEG source localization by depth-weighted minimum-norm estimates. *Neuroimage*, 31(1):160–71.

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. *Proc Natl Acad Sci U S A*, 95(15):8945–50.

Liu, A. K., Dale, A. M., and Belliveau, J. W. (2002). Monte Carlo simulation studies of EEG and MEG localization accuracy. *Hum Brain Mapp*, 16(1):47–62.

Liu, Z., Ding, L., and He, B. (2006a). Integration of EEG/MEG with MRI and fMRI. *IEEE Engineering in Medicine and Biology Magazine*.

Liu, Z., Kecman, F., and He, B. (2006b). Effects of fMRI-EEG mismatches in cortical current density estimation integrating fMRI and EEG: a simulation study. *Clinical neurophysiology: official journal of the International Federation of Clinical Neurophysiology*, 117(7):1610.

MacKay, D. (1991). *Bayesian Methods for Adaptive Models*. PhD thesis, California Institute of Technology.

MacKay, D. (2003). *Information Theory, Inference, and Learning Algorithms*. Cambridge Univ Pr.

Malmivuo, J. and Plonsey, R. (1995). *Bioelectromagnetism : Principles and Applications of Bioelectric and Biomagnetic Fields*. Oxford University Press, USA.

Matsuura, K. and Okabe, Y. (1995). Selective minimum-norm solution of the biomagnetic inverse problem. *IEEE Trans Biomed Eng*, 42(6):608–15.

Mattout, J., Pélégrini-Issac, M., Garnero, L., and Benali, H. (2005). Multivariate source prelocalization (MSP): use of functionally informed basis functions for better conditioning the MEG inverse problem. *Neuroimage*, 26(2):356–73.

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(3):753–67.

Molins, A., Stufflebeam, S. M., Brown, E. N., and Hämäläinen, M. (2008). Quantification of the benefit from integrating MEG and EEG data in minimum l2-norm estimation. *Neuroimage*, 42(3):1069–77.

Mosher, J. C., Lewis, P. S., and Leahy, R. M. (1992). Multiple dipole modeling and localization from spatio-temporal MEG data. *IEEE Trans Biomed Eng*, 39(6):541–57.

Munck, J. and Peters, M. (1993). A fast method to compute the potential in the multisphere model. *IEEE Transactions on Biomedical Engineering*, 40(11):1166–1174.

Nagarajan, S. S., Portniaguine, O., Hwang, D., Johnson, C., and Sekihara, K. (2006). Controlled Support MEG imaging. *Neuroimage*, 33(3):878–85.

Neal, R. (2003). Slice sampling. *Annals of Statistics*, 31(3):705–741.

Neal, R. M. (1994). *Bayesian Learning for Neural Networks*. PhD thesis, University of Totonto.

Nicholson, C. and Llinas, R. (1971). Field potentials in the alligator cerebellum and theory of their relationship to Purkinje cell dendritic spikes. *J Neurophysiol*, 34(4):509–31.

Nummenmaa, A., Auranen, T., Hämäläinen, M., 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):669–85.

Nummenmaa, A., Auranen, T., Hämäläinen, M., Jääskeläinen, I. P., Sams, M., Vehtari, A., and Lampinen, J. (2007b). Automatic relevance determination based hierarchical Bayesian MEG inversion in practice. *Neuroimage*, 37(3):876–89.

Nunez, P. L. and Srinivasan, R. (2005). *Electric Fields of the Brain: The Neurophysics of EEG*. Oxford University Press, USA, 2nd edition.

Okada, Y. (1993). Empirical bases for constraints in current-imaging algorithms. *Brain Topogr*, 5(4):373–7.

Ollikainen, J., Vauhkonen, M., Karjalainen, P., and Kaipio, J. P. (1999). Effects of local skull inhomogeneities on EEG source estimation. *Medical engineering & physics*, 21(3):143–154.

Palmer, J., Wipf, D., Kreutz-Delgado, K., and Rao, B. (2006). Variational EM algorithms for non-Gaussian latent variable models. *Advances in Neural Information Processing Systems*, 18:1059.

Parker, R. L. (1977). Understanding inverse theory. *Annual Review of Earth and Planetary Sciences*, 5:35–64.

Pascual-Marqui, R. (1999a). Reply to comments made by R. Grave de Peralta Menendez and SL Gozalez Andino. *International Journal of Bioelectromagnetism*, 1(2):75–86.

Pascual-Marqui, R. D. (1999b). Review of methods for solving the EEG inverse problem. *International Journal of Bioelectromagnetism*, 1(1):75–86.

Pascual-Marqui, R. D. (2002). Standardized low-resolution brain electromagnetic tomography (sLORETA): technical details. *Methods Find Exp Clin Pharmacol*, 24 Suppl D(Suppl D):5–12.

Pascual-Marqui, R. D., Michel, C. M., and Lehmann, D. (1994). Low resolution electromagnetic tomography: a new method for localizing electrical activity in the brain. *Int J Psychophysiol*, 18(1):49–65.

Perona, P. and Malik, J. (1990). Scale-space and edge detection using anisotropic diffusion. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 12(7):629–639.

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(4):997–1011.

Phillips, C., Rugg, M. D., and Friston, K. J. (2002a). Anatomically informed basis functions for EEG source localization: combining functional and anatomical constraints. *NeuroImage*, 16(3 Pt 1):678–695.

Phillips, C., Rugg, M. D., and Friston, K. J. (2002b). Systematic regularization of linear inverse solutions of the EEG source localization problem. *NeuroImage*, 17(1):287–301.

Plonsey, R. and Heppner, D. (1967). Considerations on quasi-stationarity in electro-physiological systems. *Bull.math.Biophys.*, (29):657–664.

Pursiainen, S. (2008). *Computational Methods in Electromagnetic Biomedical Inverse Problems*. PhD thesis, Helsinki University of Technology.

Roth, B. J., Balish, M., Gorbach, A., and Sato, S. (1993). How well does a three-sphere model predict positions of dipoles in a realistically shaped head? *Electroencephalography and clinical Neurophysiology*, 87(4):175–184.

Rudin, L. I., Osher, S., and Fatemi, E. (1992). Nonlinear total variation based noise removal algorithms. *Physica D Nonlinear Phenomena*, 60:259–268.

Rullmann, M., Anwander, A., Dannhauer, M., Warfield, S. K., Duffy, F. H., and Wolters, C. H. (2009). EEG source analysis of epileptiform activity using a 1 mm anisotropic hexahedra finite element head model. *Neuroimage*, 44(2):399–410.

Sadleir, R. and Argibay, A. (2007). Modeling skull electrical properties. *Annals of Biomedical engineering*, 35(10):1699–1712.

Sarvas, J. (1987). Basic mathematical and electromagnetic concepts of the biomagnetic inverse problem. *Physics in Medicine and Biology*, 32(1):11.

Sato, M., Yoshioka, T., Kajihara, S., Toyama, K., Goda, N., Doya, K., and Kawato, M. (2004). Hierarchical Bayesian estimation for MEG inverse problem. *NeuroImage*, 23(3):806–26.

Schmitt, U. and Louis, A. K. (2002). Efficient algorithms for the regularization of dynamic inverse problems: I. Theory. *Inverse Problems*, 18(3):645.

Schmitt, U., Louis, A. K., Wolters, C. H., and Vauhkonen, M. (2002). Efficient algorithms for the regularization of dynamic inverse problems: II. Applications. *Inverse Problems*, 18(3):659.

Sekihara, K. and Nagarajan, S. S. (2008). *Adaptive Spatial Filters for Electromagnetic Brain Imaging (Series in Biomedical Engineering)*. Springer, 1st edition.

Sekihara, K., Nagarajan, S. S., Poeppel, D., Marantz, A., and Miyashita, Y. (2001). Reconstructing spatio-temporal activities of neural sources using an MEG vector beamformer technique. *IEEE Transactions on Biomedical Engineering*, 48(7):760–771.

Sekihara, K., Sahani, M., and Nagarajan, S. S. (2005). Localization bias and spatial resolution of adaptive and non-adaptive spatial filters for MEG source reconstruction. *Neuroimage*, 25(4):1056–67.

Somersalo, E., Cheney, M., and Isaacson, D. (1992). Existence and uniqueness for electrode models for electric current computed tomography. *SIAM Journal on Applied Mathematics*, 52(4):1023–1040.

Stefan, H., Hildebrandt, M., Kerling, F., Kasper, B. S., Hammen, T., Dörfler, A., Weigel, D., Buchfelder, M., Blümcke, I., and Pauli, E. (2009). Clinical prediction of postoperative seizure control: structural, functional findings and disease histories. *J Neurol Neurosurg Psychiatry*, 80(2):196–200.

Steinsträter, O., Sillekens, S., Junghofer, M., Burger, M., and Wolters, C. H. (2010). Sensitivity of beamformer source analysis to deficiencies in forward modeling. Human Brain Mapping. *Human Brain Mapping*, n/a.

Tanzer, O., Järvenpää, S., Nenonen, J., and Somersalo, E. (2005). Representation of bioelectric current sources using Whitney elements in the finite element method. *Phys Med Biol*, 50(13):3023–39.

Taubin, G. (1995). A signal processing approach to fair surface design. In *Proceedings of the 22nd annual conference on Computer graphics and interactive techniques*, pages 351–358. ACM.

Tikhonov, A. N. and Arsenin, V. Y. (1977). *Solutions of Ill-Posed Problems*. Winston& Sons Washington.

Trujillo-Barreto, N. J., Aubert-Vázquez, E., and Penny, W. D. (2008). Bayesian M/EEG source reconstruction with spatio-temporal priors. *Neuroimage*, 39(1):318–35.

Trujillo-Barreto, N. J., Aubert-Vázquez, E., and Valdés-Sosa, P. A. (2004). Bayesian model averaging in EEG/MEG imaging. *NeuroImage*, 21(4):1300–1319.

Tuch, D. S., Wedeen, V. J., Dale, A. M., George, J. S., and Belliveau, J. W. (2001). Conductivity tensor mapping of the human brain using diffusion tensor MRI. *Proceedings of the National Academy of Sciences of the United States of America*, 98(20):11697.

Uutela, K., Hämäläinen, M., and Somersalo, E. (1999). Visualization of magnetoencephalographic data using minimum current estimates. *Neuroimage*, 10(2):173–80.

van der Linde, A. (2001). Model Complexity and Model Priors. Workshop on "Nonlinear Estimation and Classification".

von Helmholtz, H. (1853). Ueber einige Gesetze der Vertheilung elektrischer Ströme in körperlichen Leitern, mit Anwendung auf die thierisch-elektrischen Versuche. *Annalen der Physik und Chemie, Band 165*.

Wahba, G. (1990). *Spline Models for Observational Data*. Society for Industrial Mathematics.

Wang, J. Z., Williamson, S. J., and Kaufman, L. (1992). Magnetic source images determined by a lead-field analysis: the unique minimum-norm least-squares estimation. *IEEE Trans Biomed Eng*, 39(7):665–75.

Wipf, D. (2006). *Bayesian Methods for Finding Sparse Representations*. PhD thesis, UC San Diego.

Wipf, D. and Nagarajan, S. S. (2009). A unified Bayesian framework for MEG/EEG source imaging. *Neuroimage*, 44(3):947–66.

Wipf, D. and Nagarajan, S. S. (2010). Iterative Reweighted ℓ_1 and ℓ_2 Methods for Finding Sparse Solutions. *Journal of Selected Topics in Signal Processing (Special Issue on Compressive Sensing)*, 4(2).

Wipf, D., Ramírez, R., Palmer, J., Makeig, S., and Rao, B. D. (2007). Analysis of Empirical Bayesian Methods for Neuroelectromagnetic Source Localization. In Schölkopf, B., Platt, J., and Hoffman, T., editors, *Advances in Neural Information Processing Systems 19*, pages 1505–1512. MIT Press, Cambridge, MA.

Wolters, C. and de Munck, J. C. (2007). Volume conduction. *Scholarpedia*, 2(3):1738.

Wolters, C. H., Anwander, A., Tricoche, X., Lew, S., and Johnson, C. (2005). Influence of Local and Remote White Matter Conductivity Anisotropy for a Thalamic Source on EEG/MEG Field and Return Current Computation. *Int.Journal of Bioelectromagnetism*, 7(1):203–206.

Wolters, C. H., Anwander, A., Tricoche, X., Weinstein, D. M., Koch, M., and MacLeod, R. S. (2006). Influence of tissue conductivity anisotropy on EEG/MEG field and return current computation in a realistic head model: A simulation and visualization study using high-resolution finite element modeling. *NeuroImage*, 30(3):813–826.

Wolters, C. H., Köstler, H., Möller, C., Härtlein, J., Grasedyck, L., and Hackbusch, W. (2007). Numerical mathematics of the subtraction method for the modeling of a current dipole in EEG source reconstruction using finite element head models. *SIAM J. on Scientific Computing*. in press.

Erklärung der Eigenständigkeit

Hiermit versichere ich, Felix Lucka, dass ich die vorliegende Arbeit selbstständig verfasst und keine anderen als die angegebenen Hilfsmittel verwendet habe.
Gedanklich, inhaltlich oder wörtlich Übernommenes habe ich durch Angabe von Herkunft und Text oder Anmerkung belegt bzw. kenntlich gemacht. Dies gilt in gleicher Weise für Bilder, Tabellen und Skizzen, die nicht von mir selbst erstellt wurden.
Diese Arbeit wurde in gleicher oder ähnlicher Form noch keiner Prüfungsbehörde vorgelegt.

Münster, 10. März, 2011