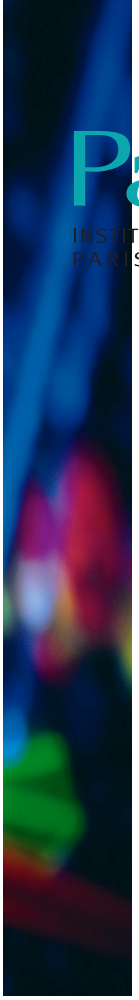
Contributions to sparse source localization for MEG/EEG brain imaging

Yousra Bekhti



ParisTech
INSTITUT DES SCIENCES ET TECHNOLOGIES
PARIS INSTITUTE OF TECHNOLOGY

EDITE - ED130 2018-ENST-0017

PhD THESIS
Télécom ParisTech

A dissertation submitted in partial fulfillment
of the requirements for the degree of

DOCTOR OF SCIENCE
Specialized in Signal and Image Processing

Yousra Bekhti

March 2018

Contributions to sparse source localization for MEG/EEG brain imaging

THÈSE

Jury :

M. Laurent Albera, MdC HdR Univ. Rennes, France	Reviewer
M. Erkki Somersalo, Prof. Case Western Reserve Univ., USA	Reviewer
M. Charles Soussen, Prof. CentraleSupélec, France	Examiner
Mrs. Maureen Clerc, Research Director, Inria Sophia-Antipolis, France	Examiner
M. Thomas Rodet, Prof. ENS Paris-Saclay, France	Examiner
M. Roland Badeau, MdC HdR Télécom ParisTech, Univ. Paris-Saclay, France	Director
M. Alexandre Gramfort, Researcher Inria, Univ. Paris-Saclay, France	Co-Director

TELECOM ParisTech
École de l'Institut Mines-Télécom - Membre de ParisTech
46 rue Barrault 75013 Paris - (+33) 1 45 81 77 77 - www.telecom-paristech.fr

To my parents.

Abstract

Magnetoencephalography (MEG) and electroencephalography (EEG) are non-invasive techniques for investigating human brain activity. They allow the measurement of ongoing brain activation on a millisecond-by-millisecond basis, which makes them very attractive to study the brain dynamics. Since the neuronal activity is measured at a sensor-level distributed over the head, the main question is how can a brain region be identified as the one producing the measured activity with reasonable accuracy? This is the so-called bio-electromagnetic inverse problem which is ill-posed, meaning there is not a unique solution to the problem. The main goal of this thesis is the development of novel methods able to localize in space and time the origin of the observed head surface signals.

To do so, very challenging mathematical and computational problems need to be tackled. First of all, since the solution to the ill-posed inverse problem is not unique, constraints need to be set in order to identify an appropriate solution among the multiple possible candidates. The constraints are chosen depending on the assumptions or a priori knowledge based on the characteristics of the source distributions. Common priors are based on the Frobenius norm and lead to a family of methods generally referred to as Minimum Norm Estimate (MNE). While these methods have some benefits like simple implementation and robustness to noise, they do not take into account the natural assumption that only a few brain regions are typically active during a specific cognitive task. Interestingly, several source reconstruction techniques have then been proposed, which are based on the assumption to promote focal or *sparse* solutions. These techniques, which are partly used in clinical routine, are suitable, *e.g.* for analyzing evoked responses or epileptic spike activity.

This thesis focuses first on the development of source solvers in the time-frequency domain to promote non-stationary focal source activation. It introduces a novel method for improving the source estimation relying on a multi-scale dictionary, *i.e.* multiple dictionaries with different scales concatenated to fit short transients and slow waves at the same time. We do not address the problem of learning the dictionary as doing so would make the cost function non-convex, which would deteriorate the speed of convergence, and also make the solver dependent on the initialization.

The second part of this thesis investigates the challenge of estimating hyperparameters involved in the regularization of the inverse problem. In the MEG/EEG community, the compromise between the data fit and the regularization controlled

by a hyperparameter is often tuned by hand, which is tedious and time consuming, or it is simply hard coded. This thesis introduces a new way of estimating this hyperparameter automatically when having a synthesis prior.

Since source estimates obtained with convex MEG/EEG sparse source imaging are biased in amplitude and often suboptimal in terms of sparsity, iterative reweighted mixed-norm solvers have been proposed in the literature. These solvers make use of non-convex concave penalties in the time or the time-frequency domain. The framework of hierarchical Bayesian modeling (HBM) is a seemingly unrelated approach to encode sparsity. Yet, the next contribution presented in this thesis shows that for certain hierarchical models, a simple alternating scheme to compute fully Bayesian Maximum-a-posteriori (MAP) estimates leads to the exact same sequence of updates as a standard iterative reweighted strategy (a.k.a. the Adaptive Least Absolute Shrinkage and Selection Operator (Lasso)).

Using simulation and various MEG/EEG datasets, this thesis provides empirical evidence that the novel methods presented here offer promising models and algorithms to improve the estimation of MEG/EEG source activations. A validation of these methods and a comparison with the widely used solvers is also presented using some phantom datasets (*i.e.* actual data recorded with known groundtruth).

keywords— Neuroimaging, magneto/electroencephalography (MEG/EEG), inverse problem, convex/non-convex optimization, sparse regression, multi-scale dictionaries, Gabor transform, hierarchical Bayesian models

Acknowledgements

First and foremost I want to express my sincere gratitude to my advisor, Alexandre Gramfort, PhD., for his valuable guidance, support, encouragements and availability during these 3 short years of my PhD. I also want to thank him for introducing me to such a great multidisciplinary world, it was a great pleasure working on sophisticated problem modeling applied to Neuroscience. Your joy and enthusiasm was motivational for me and thanks for all the funding you made available for going all around the world for the sake of Science. I would also like to thank Roland Badeau, PhD., for accepting me in his team, and for our valuable discussions trying to reduce the gap between signal processing communities in both Audio and Neuroscience research.

Thanks to Daniel Strohmeier, PhD., for his valuable advices at the beginning and all along my PhD, and for his time answering all my questions. I also want to thank Felix Lucka and Joseph Salmon for the awesome project we have done together, it was a great pleasure working with you, where I learned a lot. Also thanks to Virginie van Wassenhove for having me as an external collaborator at Neurospin - CEA Saclay where we have done a great work together. I am also grateful for everyone in the MNE development team for their feedbacks of my work and reviewing my code to be integrated.

Thanks to all my colleagues at the "Image, Data, Signal" department at Télécom ParisTech for the awesome working environment. Thanks to my closest colleagues, Mainak Jas, Tom Dupré la Tour, Stanislas Chambon for our various discussions about my work and your valuable feedback. It has been a pleasure working and sharing the B412 with you. I also want to thank Thierry Guillemot for introducing the Monday cake, everyone had the chance to discover their cooking skills, and we loved having your lemon pie.

Thanks to my music team "Les Airs Andalous" for their support all along my PhD, handling between work and concerts. You are my second family in Paris far away from my home. Thank you.

Finally, my family:
My parents: Fatima Zohra Benmansour and Mokhtar Bekhti. They gave me my name, they gave me my life, and the motivation of pursuing a PhD. I am so grateful for all your support and for always being here when needed. My siblings: SidiMohamed, Nihel, Abderrahmane and Amina, thank you for the awesome moments living together far from Tlemcen. I miss you already. I am not forgetting the latest Bekhti: Hadi, my nephew who is still 2-year old but still giving me all

the inspiration when looking at him. To finish, a big thanks to my beloved, Sami Meziane for his priceless encouragement and patience, you are the best.

My final thoughts go to Raghav RV, it was nice having you with us, may you rest in peace.

Yousra Bekhti
Paris, January 2018

Contents

Symbols and Notation

$\mathbf{I}$	Identity matrix
$\mathbb{R}$	Set of real-valued numbers
$\mathbb{R}_+$	Set of non-negative real-valued numbers
$\mathbb{C}$	Set of complex-valued numbers
$\mathbb{R}^N$	Set of real-valued vectors of length N
$\mathbb{R}^{M\times N}$	Set of real-valued $M \times N$ matrices
$\mathbb{C}^{M\times N}$	Set of complex-valued $M \times N$ matrices
$\mathcal{N}(\mu, \Sigma)$	Multivariate normal distribution with mean μ and covariance matrix Σ

a, B, λ	scalars
$\mathbf{a}$, $\mathbf{b}$, $\mathbf{c}$	column vectors
$\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$	matrices

$\mathbf{a}[i]$, $\mathbf{a}_i$	i^{th} element of $\mathbf{a}$
$\mathbf{A}[i,:]$	i^{th} row of $\mathbf{A}$
$\mathbf{A}[:,j]$	j^{th} column of $\mathbf{A}$
$\mathbf{A}[i,j]$	matrix element in the i^{th} row and j^{th} column of $\mathbf{A}$

$\mathbf{A}^\top$	matrix transpose of $\mathbf{A}$
$\mathbf{A}^\mathcal{H}$	Hermitian transpose of $\mathbf{A}$
$\mathbf{A}^{-1}$	inverse matrix of $\mathbf{A}$

$\|\mathbf{a}\|_p$	ℓ_p-norm of $\mathbf{a}$
$\|\mathbf{A}\|_p$	ℓ_p-norm of $\mathbf{A}$
$\|\mathbf{A}\|_{p,q}$	$\ell_{p,q}$-mixed-norm of $\mathbf{A}$
$\|\mathbf{A}\|_{Fro}$	Frobenius norm of $\mathbf{A}$
$\|\mathbf{A}\|$	Spectral norm of $\mathbf{A}$

$\langle \mathbf{a}, \mathbf{b} \rangle$	Inner product of $\mathbf{a}$ and $\mathbf{b}$
$\otimes$	Kronecker product
$\mathbf{AB}$	Matrix product of $\mathbf{A}$ and $\mathbf{B}$
$Tr(\mathbf{A})$	Trace of matrix $\mathbf{A}$
$(x)_+$	$\max(x, 0)$ for $x \in \mathbb{R}$

N	Number of sensors
T	Number of time samples
S	Number of sources
O	Number of orientations (1 or 3)

$\mathbf{M}$	MEG/EEG measurement matrix
$\mathbf{G}$	Gain / leadfield / forward matrix
$\mathbf{X}$	Source activity
$\mathbf{X}[i, :]$	i^{th} row of source activity $\mathbf{X}$
$\mathbf{X}[i]$, $\mathbf{X}_i$	$O \times T$ sub-matrix of $\mathbf{X}$
$\mathbf{E}$	Measurement noise
$\mathbf{Z}$	Time-frequency coefficients of the source activity
ϕ	Tight Gabor frame
$\mathcal{L}$	Lipschitz constant

Chapter 1

Introduction

1.1 Context of the thesis

Understanding the full complexity of the brain has been a challenging research project for decades, yet there are many mysteries that remain unsolved. Being able to model how the brain represents, analyzes, processes, and transforms information of millions of different tasks in a record time is primordial for both cognitive and clinical studies. These tasks can go from language, perception, memory, attention, emotion, to reasoning and creativity. Studying the behavior of the brain at each task and extracting information to define its involved network will result in a better understanding of its functions. This has been widely used in other fields such as Artificial Intelligence where scientists and engineers try to implement aspects they learned from the human brain in computers. Unlike the cognitive science questions, in the clinical diagnostic, understanding how a pathology is affecting the brain helps to find a cure or a way to improve patients' life. For example, being able to detect autism in early age of childhood helps the parents to provide a specific education and a better future.

To make this brain scanning possible, several cutting-edge technologies are used depending on the question one is asking. These techniques differ from their degree of invasiveness, and their spatial and temporal resolutions as it can be seen in Figure 1.1.1. For the different tasks I mentioned above, one very important aspect is time. The brain is able to process most of the tasks in a fraction of a second, for example to recognize an emotion, to perceive a familiar face, etc. In this thesis, to study this high temporal resolution of the brain, I was interested in two direct brain imaging techniques MEG and EEG.

MEG and EEG are functional neuro-imaging techniques for mapping the brain activity. They respectively record the magnetic fields and electric currents produced by electrical activity naturally occurring in the brain within the neurons. They use an array of sensors positioned over the scalp that are extremely sensitive to minuscule changes in the magnetic field (measured by MEG) produced by small changes in the electrical activity (measured by EEG) within the brain. It is, therefore, a direct measurement of neural activity. MEG/EEG as a technique for investigating the neural function in the brain is not new but was originally pioneered in the late 1960s. However, it is only since the early 1990s, with the introduction of high density detector grids covering the whole head, that the full potential of MEG has begun to be realized. The biggest advantage of MEG and EEG, compared to fMRI which is much more established in the neuroscience research, is the time resolution. In fMRI, the neuronal activation is indirectly measured via local changes in the level of blood oxygenation, and a long time window is typically compressed in one measured brain volume. The other techniques mentioned in Figure 1.1.1 are also indirect functional brain imaging techniques.

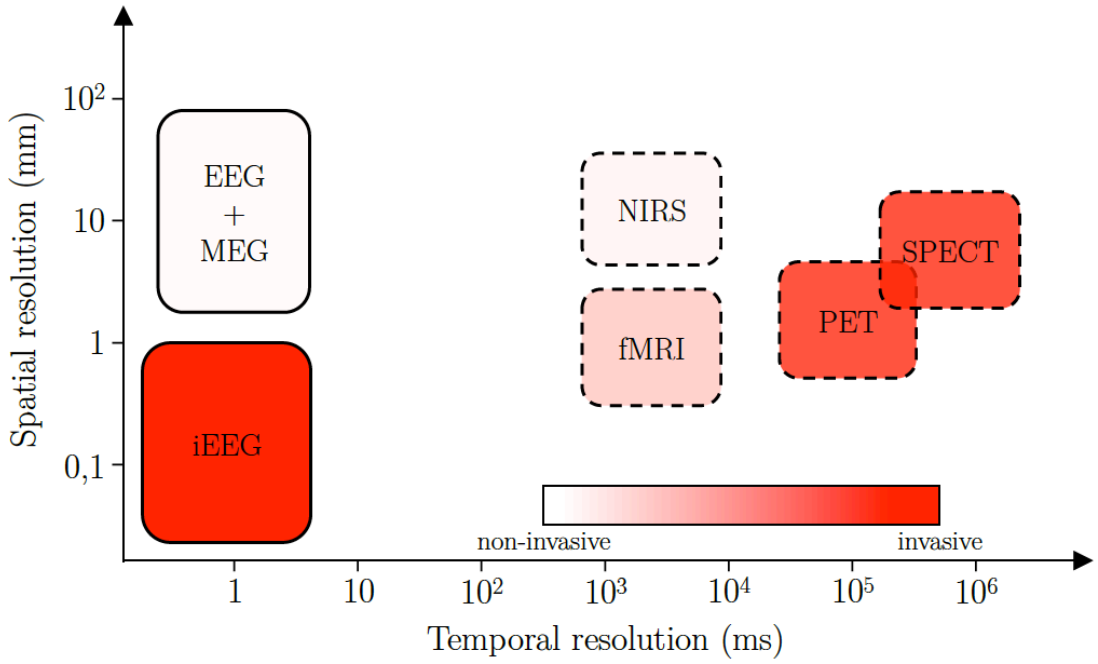


FIGURE 1.1.1: Overview of spatial and temporal resolutions of different functional neuroimaging methods. Direct approaches (EEG, iEEG, MEG) are indicated by solid boxes and indirect approaches (fMRI, NIRS, PET, and SPECT) by dashed boxes. The colors of the boxes indicate the degree of invasiveness.

Using very sensitive magnetometers/electrodes (sensors), MEG and EEG deliver insight into the brain activity with high temporal and good spatial resolution. They allow the measurements of the ongoing activity which describe the active brain sources' state at each millisecond. This problem of computing the result of the measurements is called *forward problem*. The bioelectromagnetic forward problem describes the relationship between a given neural activity in the brain and the observable MEG and EEG signals. Its solution models mathematically the neural activity, the volume conductor, and the measurement setup. It allows us to link the scalp potentials and external fields given an internal current distribution by a stable and unique solution, which is thus a well-posed problem.

Its counterpart, the bioelectromagnetic *inverse problem*, consists in using the actual measurements to infer the parameters (locations, amplitude, orientations) giving the distribution of the neural generators. It is an ill-posed problem in the sense of Hadamard [Had02] due to its non-uniqueness and high sensitivity to noise, which makes its solution unstable. The inverse problem is the so called $n \ll p$ problem in machine learning, where you have much more unknowns p to be estimated than the number of observations or variables n. This problem has infinite solutions, mainly due to the small number of sensors (obervations n) present

in the MEG and EEG. Actually, even if the MEG and EEG were measured simultaneously at infinitely many points over the head, the information would still be insufficient to uniquely compute the brain source distribution that generated the measured brain signals. This is due to the fact that there are different combinations of sources able to cause *exactly* the same potential fields on the head. Thus, to infer the neural activity generating the data at the sensor level, different source reconstruction techniques can be applied, which typically employ a priori knowledge on the state of the brain activity in order to reduce the set of solutions to a unique one.

In the past twenty years, several lines of research have emerged to tackle the problem. The most common approach widely used in clinical application is the Equivalent Current Dipole (ECD), which assumes the underlying neural sources to be focal, as for epileptic study. The limitation of the dipole fitting technique is its non-linearity, which makes the reconstruction challenging. Also, there are difficulties to accurately estimate the correct number of dipoles in advance.

Unlike the dipole fitting method, distributed source models divide the source space into a grid containing a large number of possible dipoles (ECDs). The reconstruction of the source space is done simultaneously over all ECDs, which is less challenging when having correlated sources. However due to the large number of dipoles, the corresponding regression problem is undertermined, and then requires regularization. The regularization will define the type of a priori one needs to put on the solution, such as structural information, spatio-temporal characteristics of the source estimate. The disadvantage of these models compared to dipole fitting is that the solution is smeared, meaning it is not focal, so harder to interpret.

Nevertheless, the sparsity in the solution can be promoted with a specific type of regularization in the regression model. Sparse models are actually the main interest of this thesis. They have been proposed in other fields of research and are widely used. In the signal processing literature, various signals can be defined as the linear combination of basis vectors, called *atoms*. The technique of representing the signal using few atoms is also known as compressed sensing. These atoms are defined in a fix overcomplete dictionary; the underlying motivation is that even though the observed signal lies in a high-dimensional space, the actual signal is organized in a lower-dimensional space. This property has been used in the audio domain, specifically in the analysis of speech, sounds, and music, *e.g.* in order to classify a sound sample. The idea of sparse decomposition is also behind the JPEG2000 compression, which aims to keep only a few atoms best approximating the image. In the image processing literature, sparse models were used for denoising and image reconstruction (Magnetic Resonance Imaging (MRI),...). They are also linked to the dictionary learning literature, where one tries to learn a redundant and overcomplete dictionary, which makes it possible to reconstruct a

signal/image using a sparse setting.

1.2 Objective and scope

In this thesis, I have been interested in sparse models to reconstruct the source estimate for MEG and EEG applications. Obtaining acceptable solutions, easy to interpret, does not depend only on the a priori knownledge we impose, but several questions might be asked:

- How do we best set the regularization to promote sparsity, in such a way to obtain interpretable source estimates?
- How do we set the hyperparameters of the regression problem?
- How do we quantify the uncertainty of these models?
- How do we objectively compare the different state of the art solvers?

These points define the scope of this thesis. It tries to first tackle the problem of non-stationary sources, *i.e.*, how to estimate a source that has a neuroscientific explanation as being active during a short time window only, when studying a longer window. This involves the formulation of the problem in the time-frequency domain, which needs to explicit the dictionary of the decomposition. Secondly, this thesis tries to find a way to automatically estimate the hyperparameter of the regression model to make comparison between solvers easier. The next step was to rewrite the problem as done by other communities in a Bayesian formulation. This paved the way to bridge the gap between the variational and the Bayesian worlds by writing down their equivalence under a specific parametrization of the same problem. The advantage of the Bayesian formulation is the ability to investigate the posterior distribution, making a study of the solution's uncertainty possible. This involves the presentation of Markov Chain Monte-Carlo (MCMC) algorithms to sample from the posterior distribution. A third and last project of this thesis was to put together all the actual knowledge on source localization in MEG/EEG and have a complete study of their reconstruction on a phantom dataset.

1.3 Contributions of the author

This thesis presents novel approaches for source reconstruction in MEG/EEG. It can be divided into four main projects:

- The implementation of a widely known algorithm for the MEG/EEG inverse problem, called Recursively Applied and Projected (RAP) MUSIC [ML99a]. The aim was to have a comparison with a state of the art sparse solver based

on a non-convex regularization which promotes more sparsity by getting rid of all spurious sources. This work has been published in the **IEEE Transactions on Medical Imaging (TMI)** [Str+16].

- The improvement of a previous work by Daniel Strohmeier on source reconstruction in the time frequency domain, which resulted in the introduction of the TF-MxNE (Time-Frequency mixed-norm) algorithm. The contribution tackles the problem of choosing the dictionary used to decompose the data when working in the time-frequency domain. This consists of enabling the possibility to use combined dictionaries to make the algorithm able to find both transient and longer waveforms present in the brain signal. This work has been published in the IEEE workshop on **Pattern Recognition in NeuroImaging (PRNI)** [Bek+16].

- Different lines of research to solve the MEG/EEG inverse problem gave different formulations. The most frequently used formulation in this thesis is a regularized regression model as done in most machine learning problems. With these types of models, one needs to find a good compromise between the term that tries to fit the data called the data fit, and the term regularizing the problem which takes into account any assumption one has onto the problem. This compromise is controlled by an external parameter usually called hyperparameter. For a practical example, when using sparse regularization, if this hyperparameter is fixed to a small value, *i.e.* the regularization term is not as important as the data fit, then the resulting solution will not be sparse enough and vice versa. Thus, the second contribution of this thesis was then to find an automated way to estimate this hypermarameter under some conditions of the model. This work has been published in the **European Signal Processing Conference (EUSIPCO)** [BBG17].

- The biggest drawback of the sparse solvers is the fact that they give one solution without any estimation of variance or any kind of confidence interval. Some other application areas make use of Bayesian inference, mainly because it allows the estimation of uncertainty and its quantification is paramount. Therefore, the third contribution of this thesis is to rewrite to problem as done in a Bayesian world, and tries to bridge this formulation with what has been presented so far. This project shows that under some conditions, the Bayesian formulation and the variational one are equivalent. Then, it shows how we can take advantage of the posterior distribution to extract uncertainty maps. This work has been submitted and is under review in the **Inverse Problems journal** [Bek+17].

- The final project of this thesis is to test and validate our solvers and several other ones that are the mostly used at this day for neuroscience applications. This is done on a phantom dataset which is a simulated dataset with a realistic

environment similar to a real humain brain. This work should be submitted soon to a journal paper.

- An extra project on brain decoding is presented at the end of this thesis. This work presents a novel approach based on a ridge regression with a specific metric that takes ordered target into account. The approach is novel in terms of application to MEG data. This work has been submitted and is under review in the journal **Plos One** [Bek+17].

- The implementation of some of the contributions is already on the MNE-Python package [Gra+14; Gra+13b], the others should also be integrated soon. Another contribution with a coworker's project is published in both **Pattern Recognition in NeuroImaging** [Jas+16] and **Neuroimage** [Jas+17a].

1.4 Structure of the thesis

- **Chapter 2: Background and work related to the MEG/EEG inverse problem**
This chapter defines the basics and the background needed for what will be presented in the rest of this thesis. It starts by giving the origin of the MEG and EEG recordings, *i.e.*, what do the techniques really measure? It gives then more insight on the forward operator and how it is computed efficiently. At this stage, I present a full state of the art of inverse problems defining the three main approaches: beamforming or scanning techniques, image-based methods with distributed models, and sparse source models. Afterwards, I present some basics of time-frequency decomposition, and compare several dictionaries by giving their advantages. I finish this chapter by an optimization section, defining different ways to regularize the ill-posed problem and then how do we solve them. It also gives a comparison between several solvers.

- **Chapter 3: Source localization with multi-scale dictionaries**
This chapter is dedicated to our first contribution, *i.e.* solving the inverse problem in the time-frequency domain using a multi-scale dictionary. Source localization in the time-frequency domain has already been investigated using a Gabor dictionary in a convex [Gra+13a] and a non-convex way [SGH15]. However, the choice of an optimal dictionary remains unsolved. Due to a mixture of signals, i.e. short transient signals (right after the stimulus onset) and slower brain waves, the choice of a single dictionary simultaneously explaining both signals types in a sparse way is difficult. This chapter introduces a method to improve the source estimation relying on a multi-scale dictionary, i.e. multiple dictionaries with different scales concatenated to fit

short transients and slow waves at the same time. The benefits of this approach are shown in terms of reduced leakage (time courses mixture), temporal smoothness and detection of both signals types.

- **Chapter 4: Bridges between Bayesian models and sparsity inducing norms**
  This chapter gives the basic concepts of the Bayesian formulation of the MEG/EEG inverse problem. It also aims to explain the different jargon to link the variational and the Bayesian definitions. This ends up in defining an equivalence between the two communities under some conditions, while taking advantage of the Bayesian formulation which enables us to study the multiple modes of the posterior distribution. The modes of the posterior will define several possible solutions to the inverse problem, allowing then the obtention of uncertainty maps of the source estimates.

- **Chapter 5: Benchmarking on phantom datasets**
  This chapter is a validation chapter on a phantom dataset. Phantom data is a dataset obtained by measuring the MEG/EEG activity with a humain skull phantom head. All real aspects of a head are simulated to generate the same conductivity which is expected with a real skull. The dataset shown in this chapter has four simulated dipoles at different depth. With the knowledge of the groundtruth, this chapter investigates the efficiency of each solver in terms of source localization, orientation and amplitude.

- **Chapter 6: Decoding visual motion from MEG**
  This chapter illustrates an extra project outside of the inverse problem topic. It is based on an application of machine learning to neuroscience. The aim was to develop an efficient approach to decode brain activity recorded with MEG while participants discriminated the coherence of two intermingled clouds of dots.

1.5 Publications of the author

Peer-reviewed journal papers:

- **Y. bekhti**, and A. Gramfort, "Validation of dipole localization using phantom data in MEG source imaging," in preparation.
- **Y. Bekhti**, F. Lucka, J. Salmon, and A. Gramfort, "A hierarchical Bayesian perspective on majorization-minimization for non-convex sparse regression: application to M/EEG source imaging," ArXiv preprint, (submitted).
- **Y. Bekhti**, A. Gramfort, N. Zilber, and V. van Wassenhove, "Decoding the categorization of visual motion with magnetoencephalography," Plos One, (submitted).
- D. Strohmeier, **Y. Bekhti**, J. Haueisen, and A. Gramfort, "The iterative reweighted Mixed-Norm Estimate for spatio-temporal MEG/EEG source reconstruction," IEEE Transactions on Medical Imaging, vol. 35, no. 10, pp. 2218-2228, 2016.
- M. Jas, D.A. Engemann, **Y. Bekhti**, F. Raimondo, and A. Gramfort, "Autoreject: Automated artifact rejection for MEG and EEG data," NeuroImage, vol. 159, pp. 417-429, 2016.

Peer-reviewed conference papers:

- **Y. Bekhti**, R. Badeau, and A. Gramfort, "Hyperparameter estimation in maximum a posteriori regression using group sparsity with an application to brain imaging," Signal Processing Conference (EUSIPCO), pp. 246-250, 2017.
- **Y. Bekhti**, D. Strohmeier, M. Jas, R. Badeau, and A. Gramfort, "M/EEG source localization with multi-scale time-frequency dictionaries," International workshop on Pattern Recognition in NeuroImaging (PRNI), pp. 1-4, 2016.
- M. Jas, D.A. Engemann, F. Raimondo, **Y. Bekhti**, and A Gramfort, "Automated rejection and repair of bad trials in MEG/EEG", International workshop on Pattern Recognition in NeuroImaging (PRNI), pp. 1-4, 2016.

Chapter 2

Background and work related to the M/EEG inverse problem

2.1 Signal sources of MEG and EEG recordings

At the cellular level of the brain, its nervous system is defined by the presence of a special type of neural cells. Despite the apparent simplicity in the structure of the neural cell, the biophysics of the neural current flow relies on a complex network of billions of cells, neurons and glial cells [BML01; Hod64]. Neurons are nerve cells that transmit nerve signals to and from the brain. They are about 100 billions neurons. The neuron consists of a cell body (or soma) with branching dentrites (signal receivers). They send these signals in the form of electrochemical waves traveling along thin fibers called axons, which cause chemicals called neurotransmitters to be released at junctions called synapses. A cell that receives a synaptic signal from a neuron may be excited, inhibited, or otherwise modulated. At a synapse, the cell that sends signals is called presynaptic, and the cell that receives signals is called postsynaptic.

Every neuron maintains a voltage gradient across its membrane, due to metabolically driven differences in ions of sodium, potassium, chloride and calcium within the cell, each of which has a different charge. If the voltage changes significantly, an electro-chemical pulse called an action potential (or nerve impulse) is generated. This electrical activity can be measured and displayed as a waveform called brain wave or brain rhythm. This pulse travels rapidly along the cell's axon, and is transferred across a synapse to a neighbouring neuron, which receives it through its feathery dendrites. Each individual neuron can form thousands of links with other neurons in this way, giving a typical brain well over 100 trillion synapses (up to 1,000 trillion, by some estimates).

[BML01] explains that roughly, when a neuron is excited by other —and possibly remotely located— neurons via an afferent volley of action potentials, Excitatory PostSynaptic Potentials (EPSP)s are generated at its apical dendritic tree. The apical dendritic membrane becomes transiently depolarized and consequently extracellularly electronegative with respect to the cell soma and the basal dendrites. This potential difference causes a current to flow through the volume conductor from the nonexcited membrane of the soma and basal dendrites to the apical dendritic tree sustaining the EPSPs [Glo85]. Some of the current takes the shortest route between the source and the sink by traveling within the dendritic trunk. Conservation of electric charges imposes that the current loop be closed with extracellular currents flowing even through the most distant part of the volume conductor. Intracellular currents are commonly called primary currents, while extracellular currents are known as secondary, return, or volume currents.

Both primary and secondary currents contribute to magnetic fields outside the head and to electric scalp potentials, but spatially structured arrangements of cells

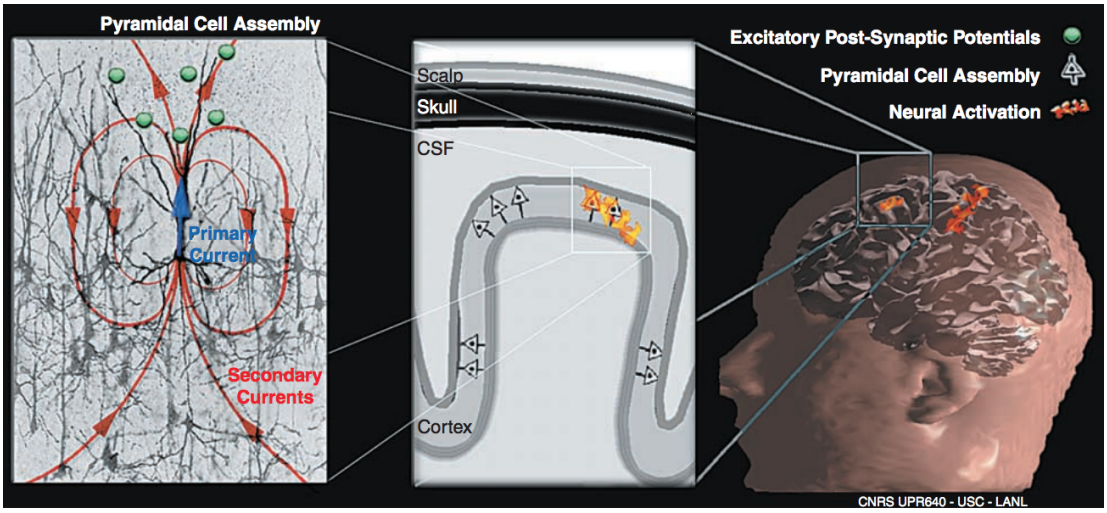


FIGURE 2.1.1: Networks of cortical neural cell assemblies are the main generators of MEG/EEG signals. Left: Excitatory postsynaptic potentials (EPSPs) are generated at the apical dendritic tree of a cortical pyramidal cell and trigger the generation of a current that flows through the volume conductor from the non-excited membrane of the soma and basal dendrites to the apical dendritic tree sustaining the EPSPs. Center: Large cortical pyramidal nerve cells are organized in macro-assemblies with their dendrites normally oriented to the local cortical surface. This spatial arrangement and the simultaneous activation of a large population of these cells contribute to the spatio-temporal superposition of the elemental activity of every cell, resulting in a current flow that generates detectable EEG and MEG signals. Right: Functional networks made of these cortical cell assemblies and distributed at possibly mutliple brain locations are thus the putative main generators of MEG and EEG signals. The origin of this image is [BML01].

are of crucial importance to the superposition of neural currents such that they produce measurable fields. Tens of thousands of synchronously activated large pyramidal cortical neurons are thus believed to be the main MEG and EEG generators because of the coherent distribution of their large dendritic trunks locally oriented in parallel, and pointing perpendicularly to the cortical surface [NS00]. The currents associated with the EPSPs generated among their dendrites are believed to be at the source of most of the signals detected in MEG and EEG because they typically last longer than the rapidly firing action potentials traveling along the axons of excited neurons [NS06].

MEG and EEG are non-invasive functional imaging techniques for analyzing the neuronal activity on a macroscopic scale. In contrast to indirect neuroimaging modalities, MEG and EEG signals derive from the net effect of ionic currents flowing in the dendrites of neurons during synaptic transmission. In accordance with Maxwell's equations, any electrical current will produce a magnetic field, and it is this field that is measured. The measurement principle of MEG and EEG is illustrated in Figure 2.1.2.

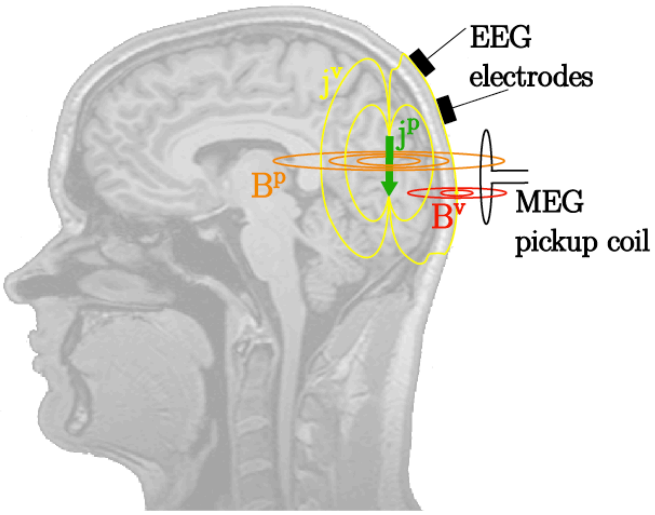


FIGURE 2.1.2: Simplified model of the measuring principle of MEG and EEG. The EEG measures the difference of the electric potential between the EEG electrode and a reference due to volume currents generated by primary currents in the brain. The MEG captures the magnetic field generated by both primary and volume currents.

The neuronal activity captured by MEG is not, as perhaps expected, generated by the (too brief) axonal action potentials of pyramidal cells, but rather by the net contributions of excitatory and inhibitory dendritic postsynaptic potentials. This current flow through the apical dendrites (represented as a 'dipole') generates a magnetic field that projects radially; thus, MEG excels at detecting dipoles arranged in a tangential orientation to the skull. Fortunately, the extensively folded sulci of the human cortex promote that orientation for the majority of cortical microcolumns. However, MEG is less sensitive to deeper (including subcortical) sources, as the magnetic field change decreases rapidly with distance.

In 1969, the journey to understand the electrical potentials of the brain took an interesting and fruitful detour when David Cohen, a physicist working at MIT, became the first to confidently measure the incredibly tiny magnetic fields produced by the heart's electrical signals. To do this, he constructed a shielded room, blocking interference from the overwhelming magnetic fields generated by earth itself and by other electrical devices in the vicinity, effectively closing the door on a cacophony of voices to carefully listen to a slight whisper. His shielding technique became central to the advent of MEG, which measures the yet even quieter magnetic fields generated by the brain's electrical activity.

This approach to record the brain's magnetic fields, rather than the electrical

potentials themselves, was advanced even further by James Zimmerman and others working at the Ford Motor Company, where they developed the SQUID, a Superconducting QUantum Interference Device. A SQUID is an extremely sensitive magnetometer, operating on the principles of quantum physics, which is able to detect precisely those very tiny magnetic fields produced by the brain. To appreciate the contributions of magnetic shielding and SQUIDs to magnetoencephalography, consider that the earth's magnetic field, the one acting on your compass needle, is at least 200 million times the strength of the fields generated by your brain trying to read that very same compass.

On the other hand, the EEG measures the electric potential difference between the EEG electrode and a reference on the scalp associated with primary currents in the brain. These electric potential differences, which are in the range of a few microvolts, are recorded using amplifiers with high open-loop gain, common-mode rejection ratio, and input impedance. The first human EEG recording was done by Hans Berger in 1924.

MEG and EEG can be recorded simultaneously and reveal complementary properties of the electrical fields. Although the signals of EEG and MEG are generated by the same sources (electrical currents in the brain), they are both sensitive to different aspects of these sources. This could be compared to viewing the shadows of the same object from two different angles; combining the two recordings usually leads to better source estimation [Mal12; Sha+07; Ayd+15].

2.2 The forward model

The bioelectromagnetic forward problem describes the relationship between a given neural activity in the brain and the observable MEG and EEG signals. We assume the electric current (denoted by $\vec{j}_t(\vec{r})$) at any position (denoted by $\vec{r}$ in the head is known at arbitrary time t. The magnetic field or the scalp voltage detected by one sensor can be modeled as an integration or a linearly weighted combination of the currents at all positions, using Maxwell's equations under a reasonable head model that describes the shape, the electrical conductivity and the permeability of various tissues [Häm+93; MLL99].

Maxwell's equations

We consider the head as a finite three-dimensional volume conductor, non magnetic. The quasi-static approximation of Maxwell's equations are a set of partial differential equations forming the foundation of classical electromagnetism. We denote by $\mathbf{E}$ the electric field, $\mathbf{B}$ the magnetic field, $\mathbf{J}$ the current density, and ρ the charge density.

$$\begin{cases} \nabla.\mathbf{E} = \frac{\rho}{\epsilon} \\ \nabla \times \mathbf{E} = \frac{-\partial \mathbf{B}}{\partial t} \\ \nabla \cdot \mathbf{B} = 0 \\ \nabla \times \mathbf{B} = \mu_0(\mathbf{J} + \epsilon \frac{\partial \mathbf{E}}{\partial t}) \end{cases}$$

For the biological signals of interest in MEG/EEG, the time-derivatives of the associated electric and magnetic fields are sufficiently small to be ignored in Maxwell's equations. Recent discussions and details of this quasi-static approximation can be found in[Häm+93; Tri83; HH92].

The propagation of the electric potentials and magnetic field measured by EEG and MEG suffers from no temporal delay, meaning that the recording is instantaneous. Let us note $\mathbf{M} \in \mathbb{R}^{N \times T}$ the measurement matrix of MEG/EEG, $\mathbf{G} \in \mathbb{R}^{N \times SO}$ the design matrix (leadfield or gain matrix [HI94a]) with S source locations in the brain and O number of orientations (1 or 3). One has:

$$\mathbf{M} = \mathbf{G}\mathbf{X} + \mathbf{E} \tag{2.1}$$

From now on, $\mathbf{E}$ denotes an additive white Gaussian noise. Equation (2.2) is linear not by assumption but by guarantee from Maxwell's equations.

If the source orientation is set a priori, *e.g.*, by using the cortical constraint assuming sources to be oriented perpendicularly to the cortical surface [DS93] (O = 1), a single dipole with unit norm per source location is used to compute the gain matrix $\mathbf{G} \in \mathbb{R}^{S \times T}$. To allow for arbitrary dipole orientations, the dipole moment per location is represented by a linear combination of O perpendicular unit dipoles. An orthogonal dipole triplet is commonly applied (O = 3). Due to the low sensitivity of MEG to radial sources, the radial component per source location is sometimes neglected (O = 2). The gain matrix $\mathbf{G}$ is generated by solving the MEG/EEG forward problem for each dipole separately and by appending the results column-wise. Hence, each column of the gain matrix provides information on the topography in the sensor space generated by the activity of a specific unit dipole, while each row reflects the sensitivity of a specific sensor to all unit dipoles in the model.

Spherical head models

A very common approximation in the forward modeling consists in assuming that the head is a set of nested concentric spheres, each corresponding to a layer with homogeneous and isotropic conductivity (Figure 2.2.1). Typically, the head is represented by three to five regions, *e.g.*, scalp, skull, cerebrospinal fluid, gray matter, and white matter, and that the conductivity is constant and isotropic within these regions. The gradient of the conductivity is therefore zero except at the surfaces

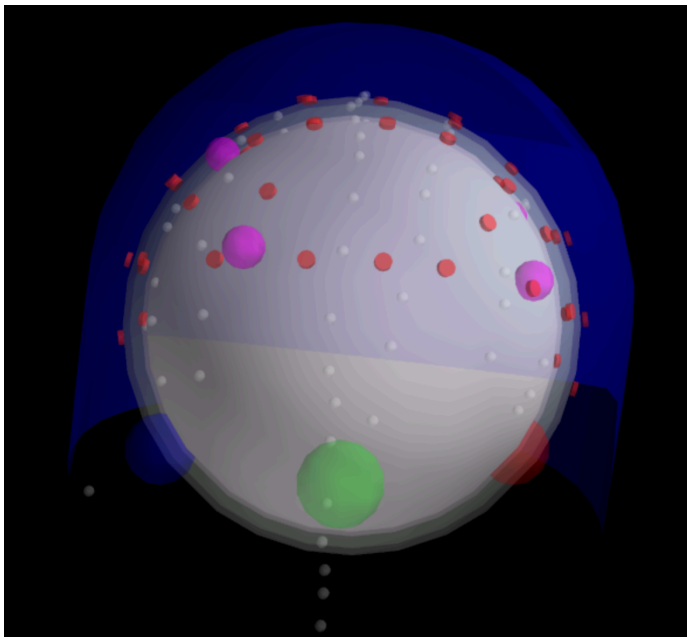


FIGURE 2.2.1: The alignment of a spherical model with three layers and the sensors. The spheres are shown in grey, the sensor space in blue, and the dots to align in order to put the spheres modeling the head and the sensors in a common coordinate system.

between regions. Computable analytic solutions exist for both MEG and EEG forward problems.

A very practical formulation of the EEG and MEG field kernels has been presented by Mosher [MLL99], that only requires vectors expressed in their Cartesian form. For MEG, since the magnetic permeability does not change across layers (and does not change much from the vacuum) and no current exists outside the head (where the sensors are located), the full magnetic field outside a set of concentric spheres can be calculated without explicit consideration of the volume currents. Therefore, the MEG spherical model does not require specifying (or assuming) the number of and the radius ratios between the spherical layers.

For EEG, the number and radii of the spherical layers are to be specified. Nonetheless, previous empirical work on closed-form approximations by Berg and Scherg [BS94] and related theoretical studies by Zhang [Zha95] have gathered a valid and convenient method for approximating an EEG field kernel from a multilayer spherical model as the weighted sum of three kernels from a single-layer spherical model applied to a modified source configuration. The optimal values of the "Berg parameters" (Eccentricity and Magnitude) in this approximation depend on the layer radii and conductivities.

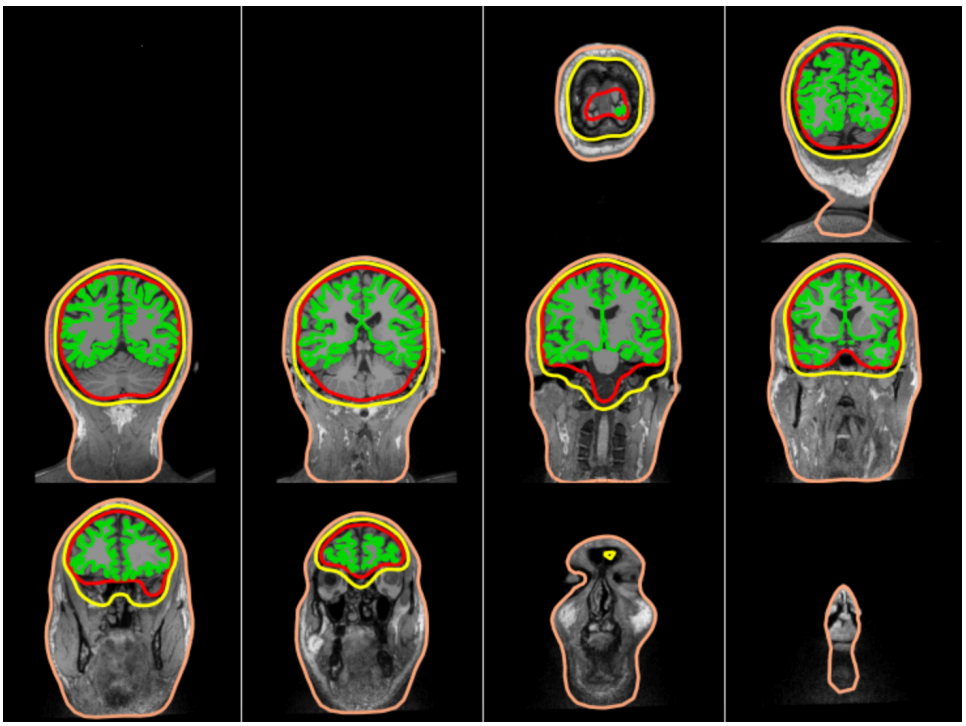


FIGURE 2.2.2: The head model: the BEM surfaces containing the three layers (inner skull, outer skull, and skin).

Realistic head models

A more realistic head model requires that the real geometry and conductivity of the head layers be taken into account as much as possible (Figure 2.2.2). For real (non-spherical) geometry and conductivity fields, numerical solutions for Maxwell's equations are to be computed.

Assuming a piecewise constant distribution of the conductivity field, approximate, yet efficient and accurate, numerical solutions can be obtained for a realistically shaped head model using the so-called boundary element methods (BEM). For BEM solutions, an MRI-based simplified description of the geometry is needed for computing the lead fields, and this can be provided in terms of (strictly) nested and closed surfaces corresponding to the boundaries separating the main tissue compartments (also called layers). In practice, BEM solutions will only require to set a few conductivity parameters (one per tissue), and to specify a few triangular meshes, as can be obtained, *e.g.*, with 3D volume segmentation tools from anatomical MRI data, each representing a separate interface between layers.

2.3 The MEG/EEG Inverse problem

An important question in the MEG/EEG community since the neuronal activity is measured at a sensor-level distributed over the head, is "how to recover the brain region(s) involved in producing the measured activity?". This is the so-called bio-electromagnetic inverse problem which is ill-posed in contrast to the forward problem. The uniqueness of the inverse problem solution is due to the fact that MEG/EEG signals can be produced by an infinite number of source configurations. Thus, to identify a stable and a unique solution among all of these infinite configurations, constraints need to be set. The constraints are chosen depending on the assumptions or a priori knowledge based on the characteristics of the source distributions, *e.g.*, spatial and/or temporal characteristics of the neural activity. The source reconstruction techniques can be in general categorized as parametric (Section 2.3.1), scanning (Section 2.3.2), and probabilistic methods (Section 2.3.3).

2.3.1 Parametric models: *dipole fitting*

Parametric methods model the problem as a small number of sources defined by their location, orientation and the strengths of the current sources that generate the MEG/EEG measurements.
The most common parametric method is the dipole fitting approaches [SVC85; MLL92; Sch90]. It assumes that the measured data have been produced by a small number of active brain regions that can each be modeled using an equivalent current dipole (ECD). These algorithms minimize a data-fit cost function such as the Frobenius norm of the residual, and they estimate five non-linear parameters per dipole: the 3D (x, y, z) position, and the two angles to define the dipole orientation. The main limitation of these methods is that they cannot be used when complex cognitive tasks are performed. This is due to the fact that the optimization problem to be solved is non-convex and multimodal, which implies that it gets easily trapped in local minima as soon as one tries to localize more than two dipoles. Furthermore, the number of dipoles to be estimated is not known and then needs to be set in advance.

2.3.2 Scanning methods: *beamforming & MUSIC*

Scanning methods, *a.k.a.* beamforming, use a discrete grid to search for optimal dipole positions throughout the source space [Hil+05; MBL99; SVC85]. An estimator of the contribution of each source location to the data can be derived either via spatial filtering or signal classification settings. The simplest spatial filter is *a matched filter* which uses the normalized columns of the gain matrix for spatial filtering, but the most common one is the linearly constrained minimum variance (LCMV) beamformer [VV+97].

LCMV performs a spatial filtering on data to discriminate between signals originating from a location of interest and those coming from elsewhere, and limits the influence of the noise. In practice, it implies that the measurement matrix is multiplied by a weighting matrix. The weighting matrix should let pass signals coming from the location of interest, while attenuating signals from elsewhere [Was08]. LCMV determines the weighting matrix by minimizing the output power of a filter under a constraint that its gain (forward operator) is unity at the location of interest. An attractive feature of beamforming is that it does not require any assumption on the number of the underlying sources. However it makes the strong assumption that the activations of the different sources are uncorrelated, which is not necessarily the case. An alternative to LCMV which integrates some information related to the experimental paradigms is called Synthetic Aperture Magnetometry (SAM) [VR01]. Beamforming can also be applied in the frequency domain using the Dynamic Imaging of Coherent Sources (DICS) [Gro+01].

Alternatives to beamformers are methods based on signal classification using subspace decompositions. The MUltiple SIgnal Classification (MUSIC) is a widely known signal processing technique that was first applied to EEG data by *Mosher* [MLL92]. The primary assumptions for this method are that the dipolar time series are mutually linearly independent [Was08]. MUSIC is based on a singular value decomposition (SVD) of the measurement data, which results in orthogonal basis vectors and singular values. Any true source localization will have a lead field (forward) vector which lies in the signal subspace computed with the SVD. MUSIC scans the brain space for source locations that satisfy this condition. The lead field vector at every candidate dipole location is systematically projected onto the signal subspace. The dipole source locations with the largest projections on the signal subspace are the active sources [MLL92; ML99b]. However, MUSIC suffers from some problems. Firstly, when the noise present in the data is correlated across channels, MUSIC can produce larger errors in the dipole localization than would have been observed with uncorrelated noise of the same power. Another problem is the detection of multiple MUSIC peaks in a 3D space of the head, each of which may correspond to a different ECD. A related problem is to determine which peaks are truly indicative of a dipolar source rather than a local minimum in the error function [ML99b].

The latter problem is solved in an extended version of MUSIC, Recursively Applied and Projected (RAP)-MUSIC, by recursive estimation of multiple sources [ML97; ML99b]. In other words, it consists in applying MUSIC successively after removing the contribution of the previously identified sources. Such as matching pursuit algorithms are used for sparse signal decomposition over dictionary of atoms [MZ93], the RAP-MUSIC method adopts a greedy strategy to select the relevant dipoles in a dictionary of sources. The implementation of RAP-MUSIC and its comparison with another solver was the first contribution of this thesis [Str+16].

2.3.3 Probabilistic modeling: *distributed sources & Bayesian approaches*

Distributed source localization estimates the amplitudes of a dense set of dipoles distributed at fixed locations within the head surface or volume. These methods are based on the reconstruction of the brain electric activity in each point of a 3D grid of solution points, the number of points being much larger that the number of electrodes on the scalp. Each solution point is considered as a possible location of a current source, thus there is no a priori assumption on the number of dipoles in the brain [Was08]. When orientations are fixed and only the amplitudes of the sources are estimated, the forward problem results in a regression formulation.

The MEG/EEG inverse problem with distributed source models leads to a regularized regression problem. The widely known method *Minimum Norm Estimate* (MNE) minimizes the ℓ_2-norm [HI94b]. MNE has been very attractive due to the fact that the inverse solution is given by a simple matrix multiplication as ℓ_2-based methods have closed-form solutions:

$$\mathbf{X}^\star = \underset{\mathbf{X}\in\mathbb{R}^{SO\times T}}{\arg\min} \|\mathbf{M}-\mathbf{G}\mathbf{X}\|_{Fro}^2 + \lambda\|\mathbf{X}\|_2^2 \quad \text{with} \quad \lambda > 0 \tag{2.2}$$

However, the choice of the λ parameter as the trade-off between data fit and regularization is sometimes tricky, because it depends on the data since λ is related to the noise level present in the measurement. λ is most of the time chosen with a cross-validation strategy to find the optimal value in the machine learning community. While in the MEG/EEG world, it is mostly tuned by hand.

When applying a minimum-norm estimate as in Equation (2.2), all the sources are penalized equivalently. This approach introduces bias over sources which are far from the sensors. Indeed sources which are close to the sensors have a higher forward field. Those sources are called superficial sources, and this bias is often known as the depth bias [PM99]. This led to the introduction of the Weighted Minimum Norm (WMN) estimate [Lin+06], which downweights the dipoles in the head that are closer to the surface.

$$\mathbf{X}^\star = \underset{\mathbf{X}\in\mathbb{R}^{SO\times T}}{\arg\min} \|\mathbf{M}-\mathbf{G}\mathbf{X}\|_{Fro}^2 + \lambda\|\mathbf{W}\mathbf{X}\|_2^2 \quad \text{with} \quad \lambda > 0 \tag{2.3}$$

$\mathbf{W}$ is a weighting matrix, which represents a priori knowledge on the source covariance. Assigning a higher variance to a deep sources is a standard approach to reduce the depth bias in MEG/EEG source imaging [Lin+06; Gra+13a; Hau+08; Hau+11; Hua+14; PS+07]. A common practice is to normalize the columns of $\mathbf{G}_s \in \mathbb{R}^{S\times O}$ of the gain matrix corresponding to the s^{th} source location using its Frobenius norm $\|\mathbf{G}_s\|_{Fro}^\delta$ or spectral norm $\|\mathbf{G}_s\|^\delta$ [Gra+14; Gra+13a; Koh+06]. The hyperparameter δ is used to prevent a bias towards very deep sources (often fixed to $\delta = 0.9$).

However, the current predicted inside the head with WMN as in Equation (2.3) is very blurred. Therefore, an alternative method has been developed called FOCUSS (FOCal Underdetermined System Solution) [GGR95]. This method changes the weights at each iteration to overcome the problem. The limitation of this method is that it does not take any biophysiological information into account, and might get stuck in local minima.

Another method for the MEG/EEG inverse problem is Low Resolution brain Electromagnetic Tomography (LORETA) [PMML94]. This method applies $\mathbf{W} = \mathbf{L}$, where $\mathbf{L}$ is the Laplacian operator. This choice of the weighting matrix imposes neighboring sources to be correlated and thus tries to find the smoothest possible solution, however it generally provides very blurred (over-smoothed) solutions.

The dSPM [Dal+00] and sLORETA [PM+02] are variants of weighted MNE which apply noise-normalized methods based on an estimate of the variance of the estimated current density. Those methods aim to represent on the cortex not the activity itself, but a *dimensionless* statistical quantity depicting the significance of each source activity. The reason why dSPM has been widely used in the MEG community is the fact that using this statistical quantity reduces the bias towards the superficial sources and makes all kinds of thresholding easy on the source estimate.

MNE and its variants solve the MEG/EEG inverse problem for each time point separately. They consider a spatial smoothness prior on the inverse problem, but they do not take the time dimension of the MEG/EEG data into account. Moreover they are dense models, which do not fit the assumption that only a few focal brain regions are involved in a specific cognitive task. MNE or dSPM for example will both have nonzero sources for every time instant. For this aim, several methods favoring sparse focal source configurations have been proposed based on a relaxation of the ℓ_0-norm. A popular approximation is ℓ_p-norms with $0 < p \leq 1$:

$$\|\mathbf{X}\|_p = \Big(\sum_{s=1}^{S} \sum_{t=1}^{T} |\mathbf{X}[s,t]|^p \Big)^{\frac{1}{p}} \tag{2.4}$$

Assuming spatially whitened MEG/EEG data, a sparse source estimate can be obtained by solving the regularized problem:

$$\mathbf{X}^\star = \underset{\mathbf{X} \in \mathbb{R}^{SO \times T}}{\arg\min} \frac{1}{2} \|\mathbf{M} - \mathbf{G}\mathbf{X}\|_2^2 + \lambda \|\mathbf{X}\| p^p \quad \text{with} \quad \lambda > 0. \tag{2.5}$$

The optimization problem is non-differentiable, and has no a closed-form solution. Hence iterative approaches need to be applied to solve the problem in Equation (2.5). Fixing $p = 1$ in Equation (2.5), the problem is known as Lasso [Tib96a] in statistics, Basis Pursuit Denoising (BPDN) [CDS98] in signal processing, and Selective Minimum Norm (SMN) [MO95] in MEG/EEG. However applying the ℓ_1-norm for the free or the loose orientation ($O = 3$) promotes sparsity even within

the orientation at each source location. To overcome this issue, Minimum Current Estimate (MCE) has been proposed by [UHS99] solving SMN by fixing the orientation priori by computing first a MNE solution.

Several other approaches can be cited that investigate the same idea of having focal source estimates. Sparse Bayesian Learning (SBL) [Wip06], Spatio-Temporal TOmographic NonNegative Independent Component Analysis (STTONNICA) [VS+09], mixed-norms [OHG09b], Champagne [Owe+12], hierarchical Bayesian inference [Luc+12b], or the Mixed-Norm estimates (MxNE) [GKH12b]. Those methods are called spatio-temporal because they work in a predefined time window, however they completely ignore the temporal correlation. This can be verified by shifting the columns of the source estimate: it will have no effect on the source estimate itself.

To introduce a "true" spatio-temporal constraint in the model, [ZR11b; ZR11a] incorporated the temporal correlation to improve the source estimates, the vector-based spatio-temporal minimum ℓ_1-norm solver (VESTAL) [Hua+06] applies a temporal projection to reduce the sensitivity to noise after using the ℓ_1-norm. The fast-VESTAL [Hua+14] is a sort of postprocessing to the VESTAL method. The Fast-VESTAL technique consists of two steps. First, ℓ_1-minimum-norm MEG source images were obtained for the dominant spatial modes of sensor-waveform covariance matrix. Next, accurate source time-courses with millisecond temporal resolution were obtained using an inverse operator constructed from the spatial source images of the first step. However this postprocessing step implicitly assumes that the source estimates are stationary. To overcome these issues, Ou et al. [OHG09a] proposed an approach to reconstruct multiple time instants simultaneously by applying the $\ell_{2,1}$-mixed-norm to impose group sparsity as a spatio-temporal regularization [GKH12b; OHG09a].

The Time-Frequency Mixed-Norm Estimate (TF-MxNE) solver [Gra+13a] reused the $\ell_{2,1}$-mixed-norm (MxNE) in the time-frequency domain by adding a second regularization over time ($\ell_{2,1} + \ell_1$). It multiplies the gain matrix by a dictionary of spatial basis functions. They obtain a modified gain matrix, which can be used to estimate spatially extented sources with temporally smooth waveforms. This approach was also investigated by [CC+15], calling the method Spatio-Temporal Unifying Tomography (STOUT).

Although these spatio-temporal methods improve the MEG/EEG source reconstruction, they are based on convex penalties. This allows fast algorithms with guaranteed global convergence. However, the resulting source estimates are biased in amplitude and often suboptimal in terms of support recovery, *i.e.*, active sources [CWB08]. As shown *e.g.* in the field of compressed sensing, promoting sparsity by applying non-convex penalties, such as logarithmic or ℓ_p-quasinorm penalties with $0 < p < 1$, improves support reconstruction in terms of feature selection, amplitude

bias, and stability [CWB08; Cha07; SCY08]. Several approaches for solving the resulting non-convex optimization problem have been proposed, including iterative reweighting ℓ_1 optimization [CWB08]. [SHG14] used an iterative reweighted approach to solve the composite non-convex penalty in the time-frequency domain.

2.3.4 Conclusion

This part has presented an overview on the state of the art of the MEG and EEG inverse solvers. Although multiple solvers have been provided, this list is definitely not exhaustive. We have kept at the end the distributed solvers which are the main interest of this thesis. Here, we will model the problem as a regularized regression with sparse priors. In the next part of the chapter, we discuss all aspects of linear regression, different penalization terms, especially the ones promoting sparsity, and the algorithms for solving those optimization problems.

2.4 Time-Frequency representation

In signal processing, time-frequency analysis encompasses those techniques that study a signal in both time and frequency domains simultaneously, using various Time-Frequency Representations (TFR). During the last decades, the signal processing community has provided many new techniques for expanding signals into "elementary" waveforms, such as wavelet bases, Modified Discrete Cosine Transform (MDCT), short time Fourier transform (STFT), Gabor wavelets (frames), etc. More often, the key issue is to obtain a sparse representation of the signal, when it is better defined in the frequency domain than the time domain. For example, a sine wave is sparsely represented in the Fourier domain, not in the time domain.

A signal representation is sparse when most information is concentrated in a small amount of data or coefficients. Several applications such as denoising make use of the sparse TFR because the noise is not sparse: source separation, signal modeling, etc. A key ingredient is to decompose a signal into a linear combination of "elementary" waveforms ϕ_i:

$$x(t) = \sum_i \alpha_i \phi_i(t) \tag{2.6}$$

with α_i the coefficients, and ϕ_i the waveforms. See [Mal08; HBB92; Wic94] for detailed examples of signal representations.

2.4.1 Modified Discrete Cosine Transform: MDCT

The mathematically simplest tool for signal decomposition is based on orthonormal bases. The waveform system $\mathcal{W} = \{\phi_i, i \in \Lambda\}$ is an orthonormal basis of the signal space (assumed to be a Hilbert space with an inner-product) $\mathcal{H}$ is:

- The atoms are mutually orthogonal and normalized: $\langle \phi_i, \phi_j \rangle = 0$ and $\|\phi_i\| = 1$
- They form a complete set of $\mathcal{H}$: if the signal $x \in \mathcal{H}$ is such that $\langle x, \phi_i \rangle = 0$ for all $i \in \Lambda$, then $x = 0$.

Then, any signal can be written in a unique way as in Equation (2.6) with $\alpha_i = \langle x, \phi_i \rangle$.

MDCT basis vectors are shift variant and its coefficients are real valued. They cannot be easily interpreted in terms of magnitude and phase.

2.4.2 Gabor dictionaries

Given a signal observed over a time interval, its conventional Fourier transform computes the frequency content but loses the time information. To analyze the evolution of the spectrum over time and hence the non-stationarity of the signal, Gabor introduced windowed Fourier atoms which correspond to a Short Time Fourier Transform (STFT) with a Gaussian window. In practice, for numerical computation, a challenge is to properly discretize the continuous STFT. The discrete STFT with a Gaussian window is also known as the discrete Gabor Transform [Gab46].

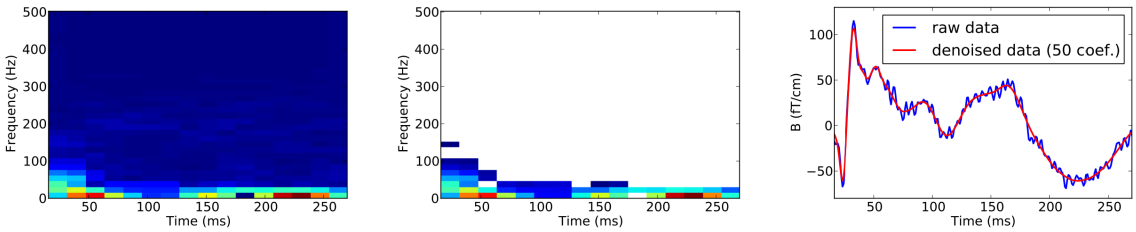
The setting we consider is the finite-dimensional one. Let $\mathbf{g} \in \mathbb{R}^T$ be a "mother" analysis window. Let $f_0 \in \mathbb{N}$ and $k_0 \in \mathbb{N}$ be the frequency and time sampling rates in the time-frequency plane generated by the STFT, respectively. The family of the translations and modulations of the mother window generates a family of Gabor atoms $(\phi_{mf})_{mf}$ forming the dictionary $\mathbf{\Phi} \in \mathbb{C}^{T \times C}$, where C denotes the number of atoms. The atoms can be written as:

$$\phi_{mf}[n] = \mathbf{g}[n - mk_0] e^{\frac{i2\pi f_0 f n}{T}}, m \in \{0, \dots, \frac{T}{k_0} - 1\}, f \in \{0, \dots, \frac{T}{f_0} - 1\}. \tag{2.7}$$

If the product $f_0 k_0$ is small enough, *i.e.*, the time-frequency plane is sufficiently sampled, the family $(\phi_{mf})_{mf}$ is a frame of $\mathbb{R}^T$, *i.e.*, one can recover any signal $\mathbf{x} \in \mathbb{R}^T$ from its Gabor coefficients $(\langle \mathbf{x}, \phi_{mf} \rangle) = \mathbf{\Phi}^{\mathcal{H}} \mathbf{x}$. More precisely, there exists two constants $A, B > 0$ such that:

$$A\|x\|_2^2 \leq \sum_{m,f} |\langle \mathbf{x}, \phi_{mf} \rangle|^2 \leq B\|x\|_2^2 \ . \tag{2.8}$$

When $A = B$, the frame is *tight*. When the vectors ϕ_{mf} are normalized, the frame is an orthogonal basis if and only if $A = B = 1$. The Balian-Low theorem says



(a) STFT (b) STFT (50 coef.) (c) MEG data

FIGURE 2.4.1: a) STFT of a single channel MEG signal sampled at 1000Hz showing the sparse nature of the transformation (window size 64 time points and time shift $k_0 = 16$ samples). b) STFT restricted to the 50 largest coefficients. c) Original data and reconstructed data using only the 50 largest coefficients.

that it is impossible to construct a Gabor frame which is a basis. Consequently, a Gabor transform is redundant or overcomplete and there exists an infinite number of ways to reconstruct $\mathbf{x}$ from a given family of Gabor atoms. In the following, the considered $\mathbf{\Phi}$ dictionaries are tight frames.

The canonical reconstruction of $\mathbf{x}$ from its Gabor coefficients requires a canonical dual window, denoted by $\tilde{\mathbf{g}}$. Following Equation (2.7) to define $(\tilde{\phi}_{mf})_{mf}$ we have:

$$\mathbf{x} = \sum_{m,f} \langle \mathbf{x}, \phi_{mf} \rangle \tilde{\phi}_{mf} = \mathbf{\Phi}^{\mathcal{H}} \mathbf{x} \tilde{\mathbf{\Phi}} = \tilde{\mathbf{\Phi}}^{\mathcal{H}} \mathbf{x} \mathbf{\Phi}, \tag{2.9}$$

where $\tilde{\mathbf{\Phi}}$ is the Gabor dictionary formed with the dual windows. If the frame is tight, then we have $\tilde{\mathbf{g}} = \mathbf{g}$, and more particularly we have $\mathbf{\Phi}\mathbf{\Phi}^{\mathcal{H}} = \|\mathbf{\Phi}\mathbf{\Phi}^{\mathcal{H}}\|\mathbf{I}$. The representation being redundant, for any $\mathbf{x} \in \mathbb{R}^T$ one can find a set of coefficients z_{mf} such that $\mathbf{x} = \sum_{m,f} z_{mf} \phi_{mf}$, while the z_{mf} verify some suitable properties dictated by the application. For example, it is particularly interesting for MEG/EEG to find a sparse representation of the signal. Indeed, a spectrogram, sometimes simply called TF transform of the data in the MEG literature, generally exhibits a few peaks localized in the time-frequency domain. In other words, MEG/EEG signals can be expressed as linear combinations of a few oscillatory atoms. In order to demonstrate this, Fig. 2.4.1 shows the STFT of a single signal from a MEG channel from a somatosensory experiment, the same STFT restricted to the 50 largest coefficients (approximately only 10% of the coefficients), and the signal reconstructed with only these coefficients compared to the original signal. We observe that the original signal can be very well approximated by only a few coefficients, *i.e.*, a few Gabor atoms.

In practice, the Gabor coefficients are computed using the Fast Fourier Transform (FFT) and not by a multiplication by a $\mathbf{\Phi}$ matrix as suggested above. Such operations can be efficiently implemented as in the LTFAT toolbox[1] [STB12]. Another practical concern to keep in mind is the trade-off between the size of the

[1] http://ltfat.sourceforge.net

window g and the time shift k_0. A long window will have a good frequency resolution and a limited time resolution. The time precision can be improved with a small time shift, leading however to a larger computational cost, both in time and memory. Finally as any computation done with an FFT, the STFT implementations assume circular boundary conditions for the signal. To take this into account and avoid edge artifacts, the signal has to be windowed.

For more details about Gabor dictionaries, please refer to [Dau+92].

2.4.3 Conclusion

This last part briefly introduced some important properties of time-frequency representations. It presented MDCT and Gabor dictionaries. It mainly explained the advantage of using the tight Gabor frames concept. The STFT or Gabor frames are time invariant, contrary to MDCT. Chapter 3 will model the inverse problem in the Time-Frequency (TF) domain, which is based on the construction of Gabor frames.

2.5 Cost functions and optimization

Due to the fact that the MEG/EEG sensors are linear combinations of the electromagnetic fields produced by all current sources, the linear forward operator, called *gain matrix*, or the *mixing matrix*, predicts the MEG/EEG measurements. Here we introduce the linear regression on which the formulation of the MEG/EEG inverse problem is based in this thesis.

2.5.1 Linear model and regression

In statistics, linear regression consists of modeling the linear relationship between an observation $\mathbf{y}$ and some explanatory variables $\mathbf{A} = [A_1, \ldots, A_n]^\top$. This relationship involves the vector of coefficients $\mathbf{x} \in \mathbb{R}^m$ such that:

$$y = Ax \tag{2.10}$$

This linear model is verified for a set of observations coming from the same event, *i.e.* the linear combination defined by $\mathbf{x}$ is the same for all (variable, observation) couples originating from the same event: $(\mathbf{A}_i, y_i) \in \mathbb{R}^m \times \mathbb{R}$ with $i \in [1, \ldots, n]$. We can note $[y_1, \ldots, y_n]^\top = \mathbf{y} \in \mathbb{R}^n$ and $[\mathbf{A}_1, \ldots, \mathbf{A}_n] \in \mathbb{R}^{m \times n}$. If all couples $\{(\mathbf{A}_i, y_i)\}_{i \in [1,\ldots,n]}$ verify exactly a linear model, then there exists a vector $\mathbf{x} \in \mathbb{R}^m$ such that $\mathbf{Ax} = \mathbf{y}$.

In a MEG/EEG application, the equivalent of $\mathbf{A}$ is the forward operator $\mathbf{G}$ which describes the linear relationship between the MEG/EEG measurements $\mathbf{M} \in \mathbb{R}^{N \times T}$ (N number of sensors, T number of time instants) and the source activation

$\mathbf{X} \in \mathbb{R}^{S \times T}$ (S is the number of source locations). The linear model then reads: $\mathbf{M} = \mathbf{G}\mathbf{X}$ where $\mathbf{G} \in \mathbb{R}^{N \times S}$ is the gain or the lead-field matrix (forward operator), a known instantaneous mixing matrix, which links source and sensor signals.

In practice, the linear model is never exactly verified due to external noise. The aim of linear regression is to additionally assume that an unobserved random variable, *i.e.* error term, is added to the linear relationship between the M/EEG measurements $\mathbf{M}$ and the source activation $\mathbf{X}$. The regression model can then be written similarly to Equation (2.2):

$$\mathbf{M} = \mathbf{G}\mathbf{X} + \mathbf{E} \tag{2.11}$$

with $\mathbf{E}$ is the measurement noise, which is assumed to be additive, white, and Gaussian, $\mathbf{E}_{:,j} \sim \mathcal{N}(0, \mathbf{I})$ for all j. This assumption is acceptable on the basis of a proper spatial whitening of the data using an estimate of the noise covariance [EG15].

Several methods exist to approximate the solution of the regression model. The most widely used is the Ordinary Least Square (OLS) approach [Leg05], which minimizes the sum of the squares of the errors or residuals as follows:

$$\mathbf{X}^\star \in \underset{\mathbf{X} \in \mathbb{R}^{S \times T}}{\arg\min} \frac{1}{2} \|\mathbf{M} - \mathbf{G}\mathbf{X}\|_{Fro}^2 \tag{2.12}$$

There exist other types of approaches like the Least Absolute Deviations (LAD), also known as least absolute errors. Instead of minimizing the squares of the errors, LAD tries to minimize the sum of the absolute values of the errors/residuals. Its advantage over OLS is that it is more robust to outliers in the data. However the LAD is not stable, *i.e.*, a small modification of the observation $\mathbf{M}$ may result in a huge variation of the estimation of $\mathbf{X}$. Moreover it can have multiple solutions, because unlike OLS, it does not have an analytical expression but needs to be computed iteratively. This explains why the OLS approach has been the standard one, along with the fact that it has a closed-form solution.

2.5.2 Regularization

Regularization in general can be applied for different reasons. We have seen why the least squares is the standard approach in linear regression. However this approach has some drawbacks: *overfitting* [2] and the fact that the closed-form solution is computed using $\mathbf{G}^\top\mathbf{G}$, which might not be invertible, giving rise to infinitely many solutions. This requires to set regularization.

[2] Overfitting: when the model fits the training data too well and has bad generalization

In general, the penalization term will be marked $\mathcal{P}(\mathbf{X})$ as in Equation (2.13) and it can take any dense or sparse form:

$$\mathbf{X}^\star = \underset{X \in \mathbb{R}^{S \times T}}{\arg\min} \frac{1}{2}\|\mathbf{M} - \mathbf{G}\mathbf{X}\|_{Fro}^2 + \lambda \mathcal{P}(\mathbf{X}) \tag{2.13}$$

In the rest of this section, we list the different regularization terms $\mathcal{P}(\mathbf{X})$ as those using: dense norms, convex sparse norms, non-convex sparse norms, and structured norms.

Non-sparsity promoting norms

Thikonov regularization [TA77] is the most commonly used penalty, also known as rigde regression [HK70]. It is part of dense norms as the estimated $\mathbf{X}^\star$ is dense, even if most of its values are almost zero. It reads:

$$\mathbf{X}^\star = \underset{X \in \mathbb{R}^{S \times T}}{\arg\min} \frac{1}{2}\|\mathbf{M} - \mathbf{G}\mathbf{X}\|_{Fro}^2 + \lambda \|\mathbf{X}\|_2^2 \tag{2.14}$$

The first term of the minimization is called the data fit, and the second term penalizes the solution by keeping the values of $\mathbf{X}$ small. We always keep the same data fit term $\frac{1}{2}\|\mathbf{M} - \mathbf{G}\mathbf{X}\|_{Fro}^2$ and change the second penalization term depending on our priori knowledge. The penalization is controlled by the λ parameter. The higher it is, the more penalized the regression is. An explicit solution of Equation (2.14) is given by $\mathbf{X}^\star = (\mathbf{G}^\top\mathbf{G} + \lambda\mathbf{I})^{-1}\mathbf{G}^\top\mathbf{M}$.

In this thesis, we are interested in sparsity promoting regularizations, which are natural for the MEG/EEG inverse problem. Indeed, it is reasonable to assume that only a few focal regions in the brain are active during a certain cognitive task.

Sparse norms: *Convex norms*

Let $\mathbf{y} \in \mathbb{R}^n$ a vector. The support of y is defined by the set $\mathcal{S}(y) = \{i = [1, \dots, n] \text{ s.t. } y[i] \neq 0\}$. A vector is sparse if its support is small, *i.e.* the cardinal $\#\mathcal{S}(y)$ is small compared to n. The cardinal of $\mathcal{S}(y)$ corresponds to the ℓ_0 pseudo-norm. The optimization problem implies then to identify $\mathcal{S}(y)$.

Coming back to our application, we assume that the signals $\mathbf{M}$ obtained with MEG/EEG are linear combinations of a small number of sources in the brain. This implies that only few sources in $\mathbf{X}$ are active, *i.e.*, $\mathbf{X}$ is sparse. However the minimization of Equation (2.13) with the ℓ_0-norm ($\mathcal{P}(\mathbf{X}) = \|\mathbf{X}\|_0$) is unfortunately an NP-hard combinatorial problem.

Due to the above undesired properties, we need to consider a convex relaxation of the ℓ_0-norm. The use of the least squares with the ℓ_1-norm (i.e. $\mathcal{P}(\mathbf{X}) = \|\mathbf{X}\|_1$ in Equation (2.13)) is the natural approximation, since it is the closest convex

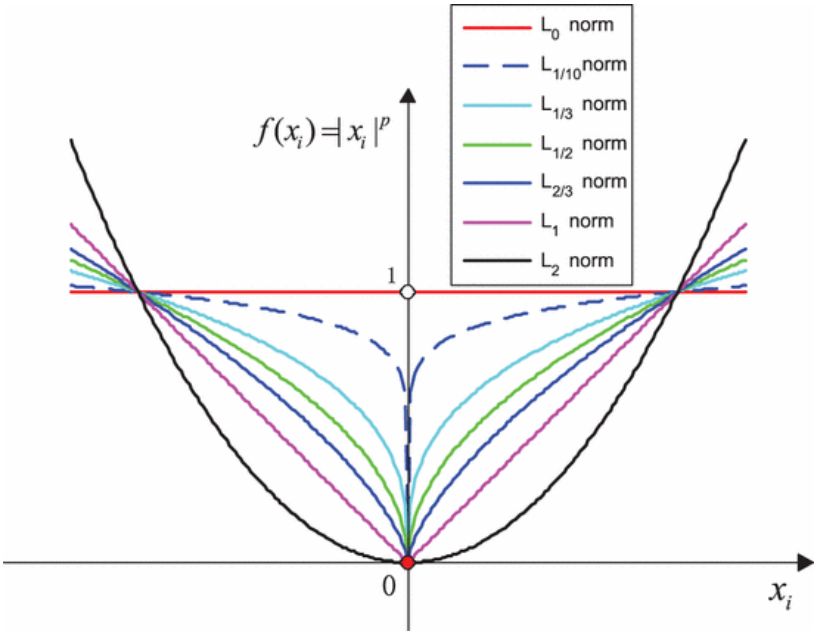


FIGURE 2.5.1: Geometric interpretation of the different norms in 1D space.

norm to the ℓ_0-norm. The ℓ_1-norm is known as *Lasso* in statistics [Tib96b], and as Basis Pursuit Denoising [CDS01] in signal processing literature.

Lasso

$$\mathbf{X}^\star = \underset{\mathbf{X}\in\mathbb{R}^{S\times T}}{\arg\min} \frac{1}{2}\|\mathbf{M} - \mathbf{G}\mathbf{X}\|_{Fro}^2 + \lambda\|\mathbf{X}\|_1 \tag{2.15}$$

The use of a convex approximation of the ℓ_0-norm is convenient, as the method always converges to a globally optimal solution. There exist also very efficient algorithms.

Sparser norms: ***Non-Convex norms***

These convex approaches allow for fast algorithms with guaranteed global convergence. However, the resulting source estimates are biased in amplitude and often suboptimal in terms of support recovery [CWB08]. This is particularly due to the high spatial correlation of the MEG/EEG forward model. As shown, **e.g.** in the compressed sensing literature, promoting sparsity by applying non-convex penalties, such as logarithmic or ℓ_p-quasinorm penalties with $0 < p < 1$, can improve support identification, as well as reduce amplitude bias [CWB08; Cha07; SCY08]. Figure 2.5.1 shows the geometric interpretation of different norms in the 1-dimensional space. The smaller p, the closer is this approximation to the exact definition of sparsity.

We investigated the $\ell_{0.5}$-quasinorm as part of the regularization, written as:

$\ell_{0.5} - quasinorm$

$$\mathbf{X}^\star = \underset{\mathbf{X}\in\mathbb{R}^{S\times T}}{\arg\min} \frac{1}{2}\|\mathbf{M} - \mathbf{G}\mathbf{X}\|_{Fro}^2 + \lambda\|\mathbf{X}\|_{0.5} \tag{2.16}$$

As the $\ell_{0.5}$-quasinorm is non-convex, it cannot be solved in the same way and with the same guarantees as the ℓ_1-norm. One algorithm for solving Equation (2.13) using the $\ell_{0.5}$ quasi-norm consists in applying an iterative reweighted approach, where each iteration boils down to a convex problem with a weighted ℓ_1 regularization.

Weighted Lasso

$$\mathbf{X}^\star = \underset{\mathbf{X}\in\mathbb{R}^{S\times T}}{\arg\min} \frac{1}{2}\|\mathbf{M} - \mathbf{G}\mathbf{X}\|_{Fro}^2 + \lambda\|\mathbf{X}\|_{\mathbf{W};1} \tag{2.17}$$

with

$$\|\mathbf{X}\|_{\mathbf{W};1} = \sum_i \sum_j W[i,j]|X[i,j]|$$

where $\mathbf{W}$ is the weight applied to matrix $\mathbf{X}$ aiming to regularize more the low coefficients, resulting in a higher sparsity. The update of the weight $\mathbf{W}$ will be presented in Chapter 3.

Structured Norms: ***non stationary sources in TF domain***

In some applications, like for MEG/EEG, one is not only interested in sparsity, as a-priori knowledge is available on the structure of the support of $\mathbf{X}$. To go beyond the sparsity with the ℓ_p-norms where $0 < p < 1$, [YL06] introduced the Group Lasso in order to take grouped structures in the data into account. It uses a mixed ℓ_2 and ℓ_1-norm on $\mathbf{X}$. The idea is to keep a small number of groups active (ℓ_1) but once a group is active, then the coefficients of that group will be all nonzero (ℓ_2).

Group Lasso

$$\mathbf{X}^\star = \underset{\mathbf{X}\in\mathbb{R}^{S\times T}}{\arg\min} \frac{1}{2}\|\mathbf{M} - \mathbf{G}\mathbf{X}\|_{Fro}^2 + \lambda\|\mathbf{X}\|_{2,1} \tag{2.18}$$

where

$$\|\mathbf{X}\|_{2,1} = \sum_i \left(\sum_j |X[i,j]|^2\right)^{1/2}$$

While the Group Lasso gives only a sparse set of groups, sometimes we would like to obtain sparsity in groups and within each group. In our application, a group is basically a source, *i.e.* a position in the brain. The Group Lasso is definitely convenient to obtain sparse source estimates, however it is not efficient for sources which are active only during small time windows. Toward this end, one can use Sparse Group Lasso [Sim+13], which is a convex combination of the Lasso and the Group Lasso penalties.

Sparse Group Lasso

$$\mathbf{X}^{\star} = \underset{\mathbf{X}\in\mathbb{R}^{S\times T}}{\arg\min} \frac{1}{2}\|\mathbf{M}-\mathbf{G}\mathbf{X}\|_{Fro}^2 + \lambda_1\|\mathbf{X}\|_{2,1} + \lambda_2\|\mathbf{X}\|_1 \tag{2.19}$$

If $\lambda_1 = 0$ then it would be equivalent to the Lasso penalty, and if $\lambda_2 = 0$, it results in the Group Lasso penalty.

2.5.3 Methods for solving sparse inverse problems

The previous section describes the MEG/EEG inverse problem as a penalized regression model. This section enumerates some methods for solving this inverse problem using sparse priors. The corresponding MEG/EEG inverse solver for the ℓ_1-norm is the MCE solver (Minimum Current Estimate) introduced by Matsuura and Okabe [MO95]. One possible way to solve the ℓ_1 penalty is to use the Iterative Least Squares (IRLS). IRLS consists in iteratively computing weighted LS by setting appropriate weights. This is based on the fact that a weighted ℓ_2-norm: $\|\mathbf{x}\|_{\mathbf{w};2} = \sum_i w[i]^k |x[i]|^2$ is equal to the ℓ_1-norm: $\|x\|_1 = \sum_i |x[i]|$, when $w[i]^k = 1/|x[i]|$, where k denotes the iteration index. This corresponds to WMN in Section 2.3.3. Similar iterative weighted methods are used to solve the (Sparse) Group Lasso corresponding to mixed-norms in both standard and time-frequency domains presented in Section 2.3.3. Other methods based on the proximity operator are used to solve non-differentiable convex optimization problems. The idea is to alternate the minimization over the smooth convex data fit using a small gradient step and the computation of the proximal operator associated with the penalty which is non-smooth.

Indeed, the MEG/EEG inverse problem can be written as:

$$\hat{\mathbf{X}} = \underset{\mathbf{X}\in\mathbb{R}^{SO\times T}}{\arg\min}\ f(\mathbf{X}) = \underset{\mathbf{X}\in\mathbb{R}^{SO\times T}}{\arg\min} \left(g(\mathbf{X}) + \lambda\mathcal{P}(\mathbf{X})\right) \text{ with } \lambda > 0\ . \tag{2.20}$$

Here $g(\mathbf{X}) : \mathbb{C}^{SO\times T} \to \mathbb{R}$ is a convex differentiable function with Lipschitz-continuous gradient. The regularization function $\mathcal{P}(\mathbf{X}) : \mathbb{C}^{SO\times T} \to \mathbb{R}$ is a non-smooth function, typically a combination of norms or quasi-norms, inducing sparsity in the time or time-frequency domain.

Proximal operators

Let $h : \mathbb{R}^n \to \mathbb{R}$ be a convex, non differentiable function. The proximity operator associated with h and $\lambda \in \mathbb{R}_+$ denoted by $\mathbf{prox}_{\lambda h} : \mathbb{R}^n \to \mathbb{R}^n$ is given by:

$$\mathbf{prox}_{\lambda h}(\mathbf{y}) = \underset{\mathbf{x}\in\mathbb{R}^n}{\arg\min} \frac{1}{2}\|\mathbf{y}-\mathbf{x}\|_2^2 + \lambda h(\mathbf{x}) \tag{2.21}$$

Algorithm 1: GROUP LASSO WITH FISTA

Input : $\mathbf{M}, \mathbf{G}, \lambda > 0$
Auxiliary variables: $\mathbf{Y}, \mathbf{X}_0 \in \mathbb{R}^{S \times T}$, $\tau_0 \in \mathbb{R}$
1. Initialization: $\mathbf{X} \in \mathbb{R}^{S \times T}$, $\mathbf{Y} = \mathbf{X}$, $\tau = 1$, and $0 < \mu < \mathcal{L}^{-1} = \|\mathbf{G}^\top \mathbf{G}\|^{-1}$
2. **repeat**
3. $\quad \mathbf{X}_0 = \mathbf{X}$
4. $\quad \tau_0 = \tau$
5. $\quad \mathbf{X} = \mathbf{prox}_{\mu\lambda\|.\|_{2,1}}(\mathbf{Y} - \mu \nabla g(\mathbf{X}))$ with $\nabla g(\mathbf{X}) = -\mathbf{G}^\top(\mathbf{M} - \mathbf{G}\mathbf{X})$
6. $\quad \tau = \frac{1+\sqrt{1+4\tau_0^2}}{2}$
7. $\quad \mathbf{Y} = \mathbf{X} + \frac{\tau_0 - 1}{\tau}(\mathbf{X} - \mathbf{X}_0)$
8. **until** convergence

return $\mathbf{X}$

This corresponds to the inverse problem where $\mathbf{G} = \mathbf{I}$. To be able to solve the problem with non-smooth penalties and $\mathbf{G} \neq \mathbf{I}$, one needs to introduce the iterative *forward-backward* algorithm [Mor65]. Each iteration computes the proximity operator of the penalty as:

$$\mathbf{X}^{(k+1)}[:, j] = \mathbf{prox}_{\mu\lambda\mathcal{P}}(\mathbf{X}^{(k)}[:, j] + \mu \mathbf{G}^\top(\mathbf{M}[:, j] - \mathbf{G}\mathbf{X}^{(k)}[:, j]), \forall j \in [1, \ldots, T] \quad (2.22)$$

μ stands for the step size and has been proved to satisfy $0 < \mu < \|\mathbf{G}^\top \mathbf{G}\|_2^{-1}$. In practice, it is fixed to $\mu = \frac{1}{\mathcal{L}} = \|\mathbf{G}^\top \mathbf{G}\|_2^{-1}$, where $\mathcal{L}$ denotes the Lipschitz constant. k represents the iteration index. For more details refer to [Mor65; CW05; DDDM04].

If the penalty is set to be the $\ell_{2,1}$-norm as in Equation (2.18), the solution is obtained by row-wise soft thresholding. These proximal gradient methods are known as the forward-backward algorithm, thresholded Landweber iterations, or the Iterative Soft Thresholding Algorithm (ISTA) or Fast Iterative Soft Thresholding Algorithm (FISTA) [Bac+12; PB+14]. FISTA or any proximal gradient method can be applied when the objective function is a sum of two terms, a convex smooth term and non-smooth term for which the proximity operator is available. A detailed algorithm of FISTA applied to Group Lasso can be found in Algorithm 1.

However, as seen before, the ℓ_1-norm is not very appropriate for M/EEG applications as it does not take the temporal correlation of the data into account. For the spatio-temporal solvers such as TF-MxNE or irTF-MxNE presented in Chapter 3, one needs to introduce the proximity operator for these composite penalties.

Proximity operator of $\ell_{2,1} + \ell_1$

Let $\mathbf{Y} \in \mathbb{R}^{S \times T}$; $\mathbf{X} = \mathbf{prox}_{\lambda_1\|\cdot\|_1 + \lambda_2\|\cdot\|_{2,1}}(\mathbf{Y}) \in \mathbb{R}^{S \times T}$ is given for each coordinate (s,t) by:

$$X[s,t] = \frac{Y[s,t]}{|Y[s,t]|}(|Y[s,t]| - \lambda_1)_+ \left(1 - \frac{\lambda_2}{\sqrt{\sum_t (|Y[s,t]| - \lambda_1)_+^2}}\right)_+ \tag{2.23}$$

where for $z \in \mathbb{R}$, $(z)_+ = \max(0, z)$ and by convention $\frac{0}{0} = 0$.

Block Coordinate Descent: BCD

Other methods for solving the MEG/EEG inverse problem with non-smooth penalties exist. We mention here the Block Coordinate Descent (BCD) scheme [Tse10]. BCD is an extension of the well known Coordinate Descent (CD) [LO09; Nes12]. CD is based on the idea of decomposing a large optimization problem into a sequence of one-dimensional optimization problems.

BCD was used to solve the Group Lasso in [Rak11b; QSG13], it is based on the same idea of alternating between a gradient step and the computation of the proximity operator of $\mathcal{P}(\mathbf{X})$ (for instance: $\ell_{2,1} + \ell_1$). BCD is used on block-separable schemes where a block is a set of coordinates and can be defined depending on the data. Here a block maps a location in the brain, *i.e.*, a block is one source. Similarly to the CD method, the order in which the different blocks are processed can be cyclic, random which improves theoretical performance [Tse01; WYL12].

As both BCD and FISTA are based on the same idea of alternating between the gradient and the proximal operator, their difference is that BCD uses at each step a subproblem specific to one block. The subproblem per block has a closed form solution, which involves applying the group soft-thresholding operator, the proximity operator associated to the $\mathcal{P}(\mathbf{X})$, for instance that defined in Equation 2.23 when using $\ell_{2,1} + \ell_1$ ($\mathcal{P}(\mathbf{X}) = \|\mathbf{X}\|_{2,1} + \|\mathbf{X}\|_1$. Accordingly, the closed form solution for the BCD subproblems solving the Group Lasso problem can be derived as:

$$\bar{\mathbf{X}}_s^{(k)} = \mathbf{X}_s^{(k-1)} + \mu_s \mathbf{G}_s^\top (\mathbf{M} - \mathbf{G}\mathbf{X}^{(k-1)})$$

$$\tilde{\mathbf{X}}_s^{(k)} = \tilde{\mathbf{X}}_s^{(k)} \max(1 - \frac{\mu[\mathbf{s}]\lambda}{\max(\|\bar{\mathbf{X}}_s^{(k)}\|_{Fro}, \mu[s]\lambda)}, 0) \tag{2.24}$$

The step length $\mu[s]$ for each BCD subproblem is determined by $\mu[s] = \mathcal{L}_s^{-1}$ with $\mathcal{L}_s = \|\mathbf{G}_s^\top \mathbf{G}_s\|$ being the Lipschitz constant of the data-fit restricted to the s^{th} source location. This step length is typically larger than the step length applicable in any proximal gradient method, which is upper-bounded by the inverse of $\mathcal{L} = \|\mathbf{G}^\top \mathbf{G}\|$.

Optimality conditions and stopping criterion

Stopping criterion: The standard way is to check if the solution at iteration k has not been improved more than a fixed tolerance threshold ϵ, for either the objective function $|f(\mathbf{X}^{(k-1)}) - f(\mathbf{X}^{(k)})| < \epsilon$, or the source estimate itself $\|\mathbf{X}^{(k-1)} - \mathbf{X}^{(k)}\|_\infty < \epsilon$. This is an acceptable strategy, although not the best one. A more rigorous criteria would be based on the *duality gap* [BV04; Bac+12].

Duality gap: It is a way to check the optimality criterion when optimizing a convex cost function f. For a subset of convex problems, the Slater's conditions apply, therefore the gap at the optimum is exactly zero [BV04]. Computing the gap needs to derive first a dual formulation of the original problem, also called the *primal* problem. For a general minimization problem, the minimum of the primal objective function $f_p(\mathbf{X})$ is bounded below by the maximum of the dual objective function $f_d(\mathbf{X})$. Then, the duality gap is defined as the difference between the minimum of the primal cost function f_p and the maximum of the dual cost f_d. For a value of of $\mathbf{X}^{(k)}$ of the primal variable at iteration k, if one can exhibit a dual variable $\mathbf{Y}^{(k)}$, the duality gap $\eta(k)$ is defined as:

$$\eta^{(k)} = f_p(\mathbf{X}^{(k)}) - f_d(\mathbf{Y}^{(k)}) \geq 0 \tag{2.25}$$

At the optimum (corresponding to $\hat{\mathbf{X}}$), if the $\mathbf{Y}^{(k)}$ is well chosen, $\eta^{(k)}$ is 0. By exhibiting a pair $(\mathbf{X}^{(k)}, \mathbf{Y}^{(k)})$, one can guarantee that $\|f_p(\mathbf{X}^{(k)}) - f_p(\hat{\mathbf{X}})\| \leq \|f_p(\mathbf{X}^{(k)}) - f_d(\mathbf{Y}^{(k)})\|$. A good stopping criterion is therefore given by a duality gap $\eta^{(k)} < \epsilon$. The solution meeting this condition is called ϵ-optimal. The challenge in practice is to find an expression for f_d and to be able to associate a good $\mathbf{Y}$ with a given $\mathbf{X}$. Experimental studies showed that for whitened data a duality gap lower than 10^{-6} does not produce distinguishable solutions [GKH12a]. For more details on how to compute the duality gap in this kind of problems, see [Bac+12; GKH12a; Str+16]

Screening rules and active set

The regularization term $\mathcal{P}(\mathbf{X})$ used in this thesis promotes spatial sparsity, which makes most of the blocks of $\hat{\mathbf{X}}$ equal to zero. We can thus reduce the computation time by primarily updating blocks that are likely to be non-zero, while keeping the remaining blocks at zero. For this purpose, data-dependent sweep patterns (such as greedy approaches based on steepest descent [LO09; WYL12]) or active set strategies can be applied [FHT10a; RF08a].

The active set strategy can be used for both Group Lasso and Sparse Group Lasso based on [RF08b; WY14]. The main idea is to start with $\mathbf{X} = 0$, which corresponds to an empty active set $\Gamma = \{\}$. We estimate an initial active set of sources Γ

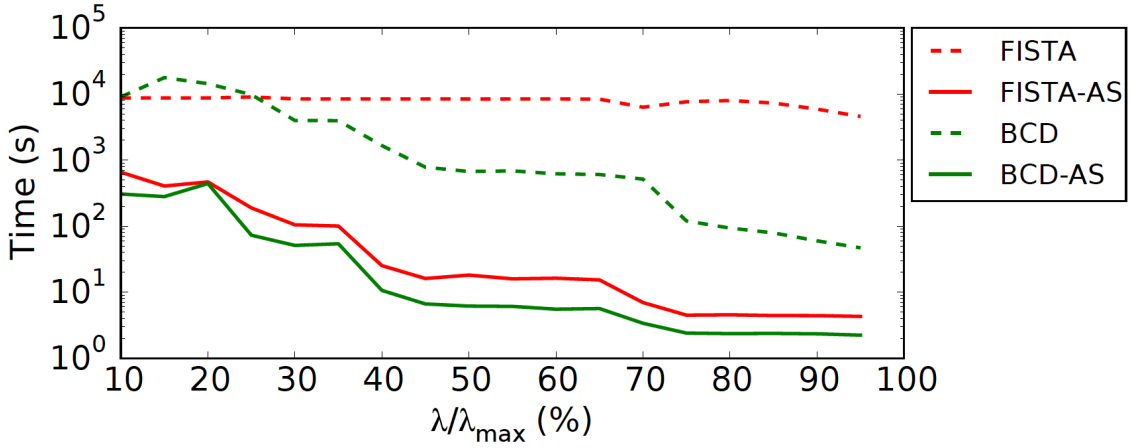


FIGURE 2.5.2: Computation time as a function of λ for group Lasso on real MEG data using BCD and FISTA with (solid) and without (dashed) active set strategy. The size of the data was: 306 sensors, 7498 cortical locations, and free orientation (O=3)

by evaluating the Karush-Kuhn-Tucker (KKT) optimality conditions [RF08b; WY14], which states that $\hat{\mathbf{X}}_s = 0$ under some conditions depending on the regularization term. We select the N sources that violates the KKT conditions the most (*e.g.* $N = 10$). Subsequently, we restrict the source space to the sources in Γ and estimate $\hat{\mathbf{X}}^\Gamma$ with convergence controlled by the duality gap. After convergence of this restricted optimization problem, we check whether $\hat{\mathbf{X}}^\Gamma$ is an ϵ-optimal solution for the original problem (without restricting the source space to Γ). If $\hat{\mathbf{X}}^\Gamma$ is not an ϵ-optimal solution indicated by $\eta \leq \epsilon$, we re-evaluate the KKT optimality conditions and update the active set Γ by adding the N sources that violate again these optimality conditions. The same procedure is then repeated with warm start.

Comparison of the different solvers

In Strohmeier et al. [Str+16], the BCD scheme was used for solving the MEG/EEG inverse problem. For the problem at hand, BCD outperforms FISTA proposed in a former work in [GKH12b]. BCD converges faster due to the reasons discussed in the BCD subsection 2.5.3. Taking bigger step depending on the current block makes the algorithm go faster to the optimal solution.

Combining the BCD and the active set strategy reduces the computation time by a factor of 100 and allows us to compute the group Lasso on real MEG/EEG data in a few seconds. All the experimental results shown in the rest of this thesis will be obtained by using BCD with active set strategy.

2.5.4 Conclusion

This chapter gives all the needed background which has been used to develop and demonstrate the upcoming results. It demonstrated how to model the inverse problem as a regularized regression problem. It defined the multiple priors that have been used in the literature including the sparse approaches that are of interest in this thesis. Then it introduced some of the methods for solving the different convex optimization problems. This is again not an exhaustive list and not all details have been presented here.

This chapter also defined the state of the art of the MEG/EEG inverse problem and how research in this field have been evolving. From the penalized regression formulation to the hierarchical Bayesian formulation, I will show in the next chapters how this thesis tries to bridge the gap between those two communities. Especially, the aim is to take advantage of each part, the computationally fast solvers developed so far by one community and the ability to quantify uncertainties of the solution in the second community.

Chapter 3

Source localization with multi-scale dictionaries

Parts of this chapter have been published in the following:

- **Y. Bekhti**, D. Strohmeier, M. Jas, R. Badeau, and A. Gramfort, "M/EEG source localization with multi-scale time-frequency dictionaries," International workshop on Pattern Recognition in NeuroImaging (PRNI), pp. 1-4, 2016.

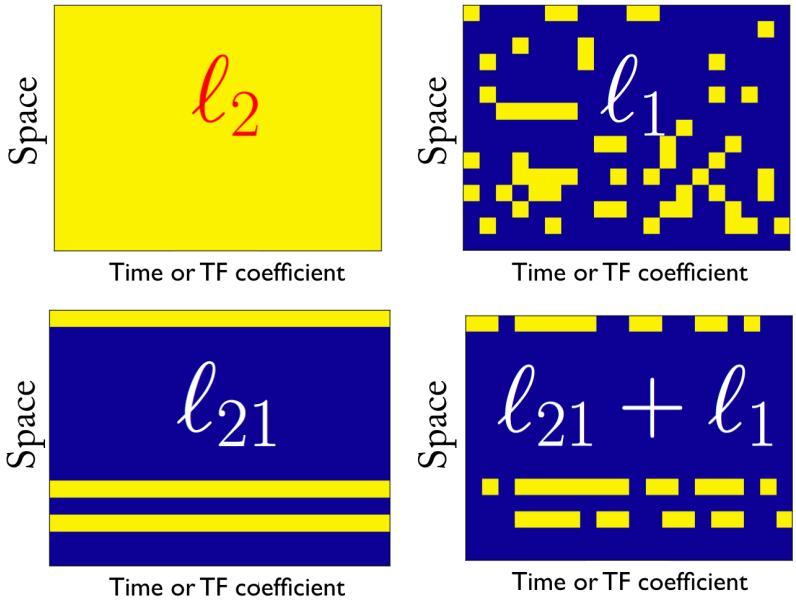


FIGURE 3.1.1: Sparsity patterns promoted by the different regularizations: ℓ_2 all non-zero, ℓ_1 scattered and unstructured non-zero, ℓ_{21} block row structure, and $\ell_{21}+\ell_1$ (TF domain) block row structure with intra-row sparsity. The yellow color indicates non-zero coefficients.

3.1 Introduction

In Chapter 2, we have seen all the background of the inverse problem in the MEG and EEG field. We justified the motivation for having sparse priors as regularization for the regression problem. Sparse priors were presented under different approaches. This chapter considers the variational problem in the time-frequency domain by fixing the penalization term as a Sparse Group Lasso as in Equation (2.19) (page 31), with λ_1 a hyperparmeter over space, and λ_2, a second hyperparameter over time. Figure 3.1.1 justifies this choice. It shows how $\ell_{21} + \ell_1$ allows for modeling non-stationary sources which cannot be estimated with the ℓ_2 or the ℓ_{21} due to the non-sparsity promoting ℓ_2-norm over time, while the ℓ_1 estimate is completely scattered and unstructured.

This chapter describes the source localization in the TF domain. We have showed in Chapter 2-Section 2.4 (page 23) why localizing the source in the TF domain was a "true" spatio-temporal approach taking the time correlation into account. The Time-Frequency Mixed Norm Estimate (TF-MxNE) [Gra+13a], Spatio-Temporal Unifying Tomography (STOUT) [CC+15] and the iterative reweighted

TF-MxNE (irTF-MxNE) [SGH15] improve the reconstruction of transient and non-stationary sources by promoting structured sparsity in the TF domain. Those methods apply a Sparse Group Lasso on the TF coefficients. TF-MxNE and STOUT apply a composite convex penalty, the sum of an $\ell_{2,1}$-mixed-norm and an ℓ_1-norm penalty, on the Gabor transform of the source time courses. On the other hand, irTF-MxNE applies a composite *non-convex* penalty, the sum of an $\ell_{2,0.5}$-quasinorm and an $\ell_{0.5}$-quasinorm penalty on the TF coefficients. The non-convex penalties have been shown to outperform convex approaches both in terms of source recovery and amplitude bias [CWB08; al.10], as explained in Chapter 2. However, the choice of an optimal Gabor dictionary for decomposing the data remains difficult.

This issue of the choice of the dictionary is specially encountered when a mixture of signals is available in the data, *e.g.* a short transient signal right after the onset of a stimuli, and slower brain waves afterward. The choice of a unique dictionary describing both signals in a sparse way is hard. We show in this chapter how to incorporate a multi-scale dictionary in the iterative reweighted optimization algorithm, *i.e.* multiple dictionaries with different scales concatenated to fit short transients and slow waves at the same time, while keeping computational efficiency. The optimization problem is solved in the same way as irTF-MxNE [SGH15], *i.e.* each iteration is a weighted TF-MxNE, which we solve using BCD (Section 2.5.3) and an active set strategy (Section 2.5.3) [FHT10b]. We demonstrate the benefit of the multi-scale dictionary in terms of reconstructed source time courses and temporal unmixing of activations.

3.2 Inverse problem in the Time-Frequency domain

Using a dictionary of TF atoms, such as a tight Gabor frame (*cf.* Section 2.4.2 - page 24), $\mathbf{\Phi} \in \mathbb{C}^{T \times C}$ (T samples, C atoms), the neuronal activation $\mathbf{X} \in \mathbb{R}^{SO \times T}$ (S sources, O orientations) can be modeled as a linear combination of atoms, $\mathbf{X} = \mathbf{Z}\mathbf{\Phi}^{\mathcal{H}}$, where $\mathbf{Z} \in \mathbb{C}^{SO \times C}$ is the TF coefficients matrix. A Gabor frame $\mathbf{\Phi}$ is tight (see Section 2.4) when the Euclidean norm of the input signal and the vector of TF coefficients are proportional ($\|\mathbf{Z}\|_2^2 = A_{\mathbf{\Phi}} \|\mathbf{X}\|_2^2$ where $A_{\mathbf{\Phi}} > 0$). When $A_{\mathbf{\Phi}} = 1$ the frame is said to be normalized. We will use tight frames in the following.

The MEG/EEG measurements matrix$\mathbf{M} \in \mathbb{R}^{N \times T}$ (N sensors) follows the forward model equivalent to Equation (2.2) (page 15):

$$\mathbf{M} = \mathbf{G}\mathbf{X} + \mathbf{E} = \mathbf{G}\mathbf{Z}\mathbf{\Phi}^{\mathcal{H}} + \mathbf{E} \tag{3.1}$$

where $\mathbf{G} \in \mathbb{R}^{N \times SO}$ stands for the forward operator. $\mathbf{E} \in \mathbb{R}^{N \times T}$ is the measurement noise, which can be assumed to be additive white noise: $\mathbf{E}[:, j] \sim \mathcal{N}(0, \mathbf{I})$ for all j after spatial whitening [Eng+15]. Estimating the coefficients $\mathbf{Z}$ given the measurement $\mathbf{M}$ is an ill-posed problem and constraints have to be imposed on $\mathbf{Z}$ to obtain

a unique source estimate, as described for $\mathbf{X}$ in the last chapter. For analyzing evoked responses, we assume that the neuronal activation is spatially sparse and temporally smooth. This corresponds to a row sparsity [Gra+13a], which we promote by applying a composite non-convex regularization $\mathcal{P}(\mathbf{Z})$ (see Figure 3.1.1). The associated regularized regression problem equivalent to Equation (2.13) is:

$$\mathbf{Z}^\star = \arg\min_{\mathbf{Z}} \frac{1}{2}\|\mathbf{M} - \mathbf{G}\mathbf{Z}\mathbf{\Phi}^{\mathcal{H}}\|_{Fro}^2 + \mathcal{P}(\mathbf{Z}) \tag{3.2}$$

with

$$\mathcal{P}(\mathbf{Z}) = \lambda_{space}\|\mathbf{Z}\|_{2,0.5} + \lambda_{time}\|\mathbf{Z}\|_{0.5} \tag{3.3}$$

where $\lambda_{space} > 0$, $\lambda_{time} > 0$. A large regularization parameter λ_{space} will lead to a spatially very sparse solution or even an empty solution, while a large λ_{time} will promote sources with smooth time series and might loose sharp aspects of the neural activity.

3.3 Fast iterative reweighted TF-MxNE with tight frames

Given a dictionary $\mathbf{\Phi}$, the optimization problem in Equation (3.2) can be solved by iteratively minimizing convex surrogate problems [SGH15]. The regularization term at each iteration k is a weighted convex mixed norm that can be written as:

$$\mathcal{P}(\mathbf{Z}) = \lambda_{space}\|\mathbf{Z}\|_{\mathbf{W}_1^{(k)};2,1} + \lambda_{time}\|\mathbf{Z}\|_{\mathbf{W}_2^{(k)};1} \tag{3.4}$$

with $\forall s, c$,

$$\begin{aligned}\mathbf{W}_1^{(k)}[s,c] &= \left(2\sqrt{\|\hat{\mathbf{Z}}^{(k-1)}[s,:]\|_2 + \epsilon^{(k-1)}}\right)^{-2} \\ \mathbf{W}_2^{(k)}[s,c] &= \left(2\sqrt{|\hat{\mathbf{Z}}^{(k-1)}[s,c]| + \epsilon^{(k-1)}}\right)^{-1}\end{aligned},$$

where $\mathbf{W}_1$ and $\mathbf{W}_2$ are the weights applied to the TF coefficients, and $\hat{\mathbf{Z}}^{(k-1)}$ are the estimated coefficients at iteration $(k-1)$. $\epsilon^{(k-1)} \in \mathbb{R}^+$ is used to prevent infinite weights. To have an intuition about the update rule for the weights, one can prove that updating the weights with w_i=$|x_i|^{p-2}$ leads to a solution of the ℓ_p-norm penalized problem.

For solving Equation (3.2), we use BCD [Tse10]. The algorithm boils down to sequentially computing a gradient step and the proximity operator of the $\ell_{2,1} + \ell_1$ norm for each block s of coefficients (see Section 2.5.3-page 33). Here a block maps to a location in the brain. One update of a block of coefficients is given by a first

gradient step:

$$\mathbf{R} = \mathbf{M} - \mathbf{G}\hat{\mathbf{X}} \tag{3.5}$$

$$\bar{\mathbf{X}}[s,:] = \hat{\mathbf{X}}[s,:] + \mu[s]\mathbf{G}[:,s]^\top \mathbf{R} \tag{3.6}$$

$$\bar{\mathbf{Z}}[s,:] = \bar{\mathbf{X}}[s,:]\mathbf{\Phi} \tag{3.7}$$

followed by the computation of the proximity operator of the weighted $\ell_{2,1} + \ell_1$ described in Chapter 2-Equation (2.23):

$$\tilde{\mathbf{Z}}[s,c] = \bar{\mathbf{Z}}[s,c]\left(1 - \frac{\mu[s]\lambda_{time}\mathbf{W}_2^{(k)}[s,c]}{|\bar{\mathbf{Z}}[s,c]|}\right)_+ \tag{3.8}$$

$$\hat{\mathbf{Z}}[s,c] = \tilde{\mathbf{Z}}[s,c]\left(1 - \frac{\mu[s]\lambda_{space}\sqrt{\mathbf{W}_1^{(k)}[s,c]}}{\|\tilde{\mathbf{Z}}[s,:]\|_2}\right)_+ \tag{3.9}$$

When $\mathbf{\Phi}$ is a tight frame, $\mu[s]$ is the step length for each BCD subproblem and it is given by $\mu[s] = \sqrt{\mathbf{A}_{\mathbf{\Phi}}}(\|\mathbf{G}[:,s]^T\mathbf{G}[:,s]\|)^{-1}$. This step length, *i.e.* the inverse of the Lipschitz constant restricted to source s, is typically larger than the step length applicable in iterative proximal gradient methods, which is upper bounded by $\|\mathbf{G}^T\mathbf{G}\|^{-1}$. This implies a bigger step to speed up the convergence. Finally:

$$\hat{\mathbf{X}}[s,:] = \hat{\mathbf{Z}}[s,:]\mathbf{\Phi}^{\mathcal{H}}. \tag{3.10}$$

Equations (3.8) and (3.9) are respectively solutions of the proximity operator for the weighted ℓ_1-norm and for the weighted $\ell_{2,1}$-norm. As the ℓ_1 proximity operator shrinks coefficients towards zero, if a block of coefficients were set to zero by the $\ell_{2,1}$ proximity operator, it would also be set to zero after the application of the ℓ_1 proximity operator. As a consequence, it is possible to know just by applying the $\ell_{2,1}$ proximity operator to $\bar{\mathbf{X}}[s,:]$ if the set of coefficients $\tilde{\mathbf{Z}}[s,:]$ will be set to zero. Note that this is just a sufficient condition and that we may have to compute all steps to know if the block is set to zero. This is summarized in the following lemma.

Lemma 1 *Let $\mathbf{\Phi}$ be a frame with constant $A_{\mathbf{\Phi}}$; if $\|\bar{\mathbf{X}}[s,:]\|_2 \leq \mu[s]\lambda_{space}\sqrt{\mathbf{W}_1^{(k)}[s,c]}/\sqrt{A_{\mathbf{\Phi}}}$, then $\hat{\mathbf{Z}}[s,c] = 0, \forall c$.*

Computing the TF decomposition at each iteration can be costly. The consequence of the lemma is that for a lot of source locations one can avoid computing their TF decomposition during the optimization, just by computing the ℓ_2-norm of the time courses after the gradient step. In order to speed up the computation even more, we combine the BCD scheme with an active set strategy (Section 2.5.3 - page 34) [FHT10b], which primarily updates sources that are likely to be active, while keeping the remaining sources inactive.

Orientation constraints

Let $\mathbf{Z}_s \in \mathbb{C}^{3\times C}$ be the block of the $\mathbf{Z} \in \mathbb{C}^{3S\times C}$ ($O = 3$) corresponding to the s^{th} source, $\mathbf{Z}[1,:]$ be the activity of the dipole oriented normal to the cortical surface, and $\mathbf{Z}[2,:]$ and $\mathbf{Z}[3,:]$ be the two other orientations tangent to the surface. We modify the ℓ_1-norm and the $\ell_{2,1}$-norm for the free orientation constraints ($O = 3$) as follows:

$$\|\mathbf{Z}\|_1 = \sum_{s,c} \sqrt{|\mathbf{Z}_s[1,c]|^2 + \frac{1}{\kappa^2}|\mathbf{Z}_s[2,c]|^2 + \frac{1}{\kappa^2}\mathbf{Z}_s[3,c]|^2}$$

$$\|\mathbf{Z}\|_{2,1} = \sum_{s} \sqrt{\sum_{c} |\mathbf{Z}_s[1,c]|^2 + \frac{1}{\kappa^2}|\mathbf{Z}_s[2,c]|^2 + \frac{1}{\kappa^2}\mathbf{Z}_s[3,c]|^2}$$

where s indexes the source location and c the TF coefficient. When $\kappa = 1$, no orientation contraint is applied and the modified penalties amount to grouping the orientation in a common ℓ_2-norm. In practice κ is fixed to 0.2.

3.4 Inverse problem with multi-scale tight Gabor frames

As shown in Section 2.4.2 (page 24), a tight Gabor frame is computed by setting two parameters: the length of the window (window size), and an overlap parameter (time shift). The window size defines the time/frequency resolution. If its length is short, it would be more focused on time than frequency, and vice versa. If it is long, it will be more focused on frequency than in time. The time resolution also depends on the time shift parameter. The time shift parameter defines the time step from one window to another (*cf.* Section 2.4). This affects the redundancy of the dictionary. A dense sampling of the TF space, however, increases the computational complexity on both time and memory.

Each source waveform is a sparse linear combination of atoms from this dictionary. Fixing those parameters is then critical for having an optimal dictionary. Learning the dictionary might be a solution to avoid fixing the parameters, or the need to have an overcomplete dictionary covering a broad range of scales. However, learning both $\mathbf{Z}$ and $\mathbf{\Phi}$ simultaneously is a non-convex optimization problem, for which one needs to alternate between a convex optimization for the two variables [MM+14].

Let us define a multi-scale TF dictionary, where we concatenate Q tight Gabor frames $\mathbf{\Phi}_q$, $1 \leq q \leq Q$, with different resolutions. One can realize that this union of tight frames $\mathbf{\Phi} = [\mathbf{\Phi}_1, \ldots, \mathbf{\Phi}_Q]$ is also a tight frame with $A_{\mathbf{\Phi}} = \sum_q A_{\mathbf{\Phi}_q}$. The strategy presented in the previous section 3.2 is therefore still relevant for a multi-scale dictionary, where the activation $\mathbf{Z}$ is a concatenation of $\mathbf{Z}_1, \mathbf{Z}_2, ..., \mathbf{Z}_Q$. Algorithm 2

Algorithm 2: MULTI-SCALE TF-MXNE WITH ACTIVE SET STRATEGY

input : $\mathbf{M}, \mathbf{G}, \mathbf{\Phi} = [\mathbf{\Phi}_1, \ldots, \mathbf{\Phi}_Q], \lambda_{space} > 0, \lambda_{time}$, and $\epsilon > 0$
init : $\mathbf{Z} \in \mathbb{R}^{SO \times C}$, $\Gamma = \{\}$, $\eta = f_p(\mathbf{Z}) - f_d(\mathbf{Y})$, μ with $\mu[s] = \mathcal{L}_s^{-1} = \|\mathbf{G}_s^\top \mathbf{G}_s\|^{-1}$
while $\eta \geq \epsilon$ **do**
  $\Gamma^\star = \{s \mid \|\mathbf{prox}_{\lambda_{time}\|.\|_1}(\mathbf{G}_s^\top(\mathbf{M} - \mathbf{G}\mathbf{Z}\mathbf{\Phi}^{\mathcal{H}})\mathbf{\Phi})\|_{Fro} > \lambda_{space}\}$
  $\Gamma = \Gamma \cup \Gamma^\star$ (Update of the active set)
  $\mathbf{Z}_\Gamma^\star \leftarrow$ output of Algorithm 3 with μ and Γ
  $\mathbf{Z}[\Gamma, :] = \mathbf{Z}_\Gamma^\star$
  $\eta = f_p(\mathbf{Z}) - f_d(\mathbf{Y})$

Algorithm 3: MULTI-SCALE TF-MXNE WITH BCD

input : $\mathbf{M}, \mathbf{G}, \mathbf{\Phi}, \mu, \lambda_{space} > 0, \lambda_{time} > 0, \epsilon > 0$, and Γ
init : $\eta = f_p(\mathbf{X}) - f_d(\mathbf{Y})$
while $\eta \geq \epsilon$ **do**
  for $s \in \Gamma$ **do**
    $\mathbf{Z}_s = \mathbf{prox}_{\mu[s](\lambda_{space}\|.\|_{2,1} + \lambda_{time}\|.\|_1)}(\mathbf{Z}_s + \mu[s]\mathbf{G}_s^\top(\mathbf{M} - \mathbf{G}\mathbf{Z}\mathbf{\Phi}^{\mathcal{H}})\mathbf{\Phi})$
  $\eta = f_p(\mathbf{X}) - f_d(\mathbf{Y})$

describes how to solve the inverse problem in the TF domain with a multi-scale dictionary.

3.5 Experiments with different dictionaries

We first evaluate the accuracy of irTF-MxNE with and without multi-scale on realistic simulations. We then apply our new solver on MEG somatosensory data.

3.5.1 Simulation

We generated a realistic simulation dataset based on a fixed-orientation source model with 7549 cortical locations and 102 magnetometers. Two of these locations were selected to be active in the primary and secondary somatosensory cortex (S1 and S2). The corresponding time courses are shown in Figure 3.5.1-a in blue (S1) and green (S2). We have both a transient source around 40 ms and slow waves afterwards around 70, 100 and 150 ms. The irTF-MxNE solver improves the source recovery [SGH15]. Therefore, we do not compare the solvers presented here over the active set size or an F_1 measure [1], as both solvers are already able to recover all the sources. We evaluate our approach by computing the explained variance between simulated source courses and the source estimation from each solver as follows:

$$\theta = 1 - \frac{\|GX_{sim} - GX_{est}\|_{Fro}^2}{\|GX_{sim}\|_{Fro}^2} \tag{3.11}$$

[1] The F_1 score is the harmonic mean of precision and recall: $F_1 = 2.\frac{precision.recall}{precision+recall}$

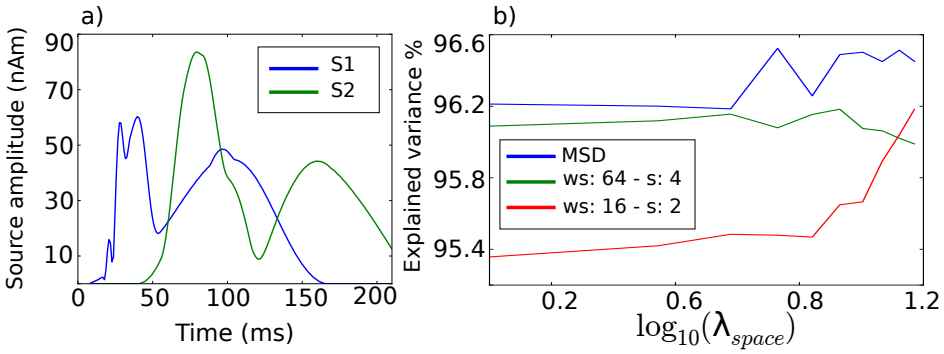


FIGURE 3.5.1: (a) Simulated source time courses in S1 (blue) and S2 (green). (b) The explained variance for irTF-MxNE using two different dictionaries: long window size (ws) 64 with time shift (s) 4 (green), and small window size 16 with time shift 2 (red). The combination of the two dictionaries is shown in blue. This shows how the multi-scale dictionary (MSD) improves the explained variance.

Fig. 3.5.1-b shows the explained variance for the irTF-MxNE with different dictionaries over a logarithmic grid of λ_{space}. The first Gabor dictionary is constructed with a 64-sample-long window and a 4 samples time shift (green), the second Gabor dictionary is constructed with a 16 sample-long window and 2 samples time shift (red) and the third one is the combination of the two dictionaries (blue). We observe that the irTF-MxNE solver using the combination of two dictionaries outperforms the solver with each dictionary separately in terms of explained variance measure over all parameter range. Higher values of $\log(\lambda_{space}) > 1.2$ impose high penalization on the active set size, resulting in a too sparse source estimate, where the solution does not explain the measurement anymore. The results show a source reconstruction improvement, which leads to a larger explained variance.

3.5.2 Experimental results with MEG somatosensory data

In order to demonstrate the advantage of irTF-MxNE with a multi-scale dictionary over the basic irTF-MxNE, we tested different parameters for different solvers on a MEG dataset: somatosensory study of the MIND dataset (see details in [Wei+07]). The evoked response is shown in Figure 3.5.2. One can already notice this mixture of brain waves in the evoked. Sharper waves right after the onset are due to a nice alignment of the trials whose information is not lost after averaging. This is mainly known as a response of the primary somatosensory area (S1) which answers quickly after a painless electrical stimulation of the median nerve. A longer wave which comes later around 70ms is clearly seen from the evoked too. This is what makes this data a challenging dataset and a very good one for testing the multi-scale solver.

Source estimation was first performed using several solvers: irTF-MxNE, ir-MxNE [SHG14] and dSPM [Dal+00]. Regarding irTF-MxNE, two dictionaries were tested (both STFT dictionaries). A dictionary with a 64 sample-long window and a

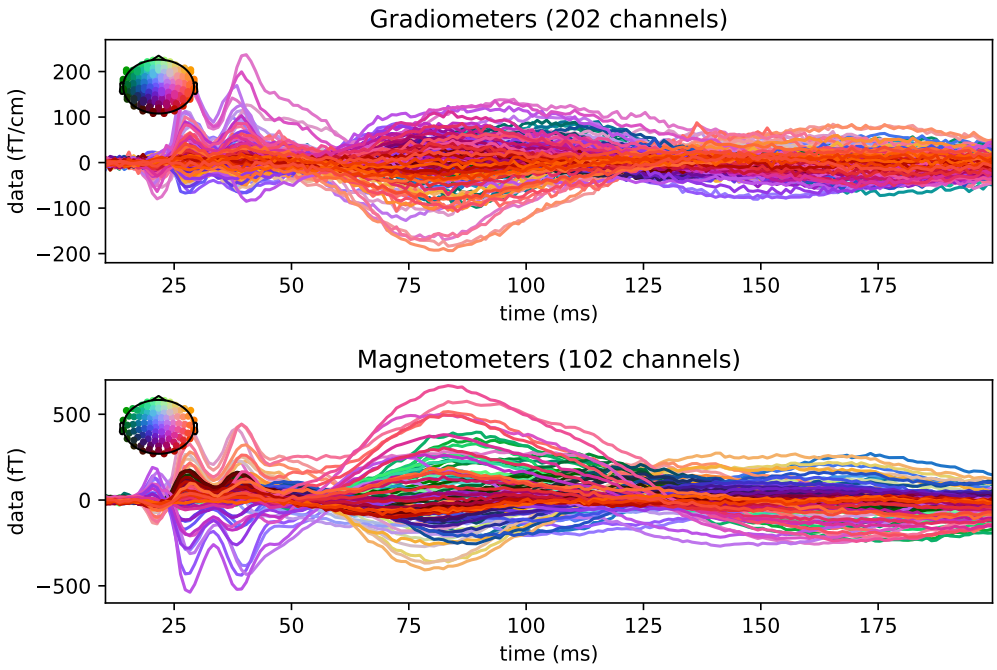


FIGURE 3.5.2: Somatosensory evoked response after preprocessing and averaging (gradiometers and magnetometers data). The top left circle gives the position of the sensors over the head which are color-coded

4 samples time shift, which leads to smooth source courses; and a dictionary with a 16 sample-long window and a 2 samples time shift, which helps capture short transient sources. After inspection of the residual in Figure 3.5.3, results show that at least four sources are necessary to capture all evoked components.

We have therefore fixed the parameters of the irTF-MxNE solvers as to obtain only four sources while explaining as much variance as possible. After that, we experimented with a set of different parameters and we show two of them, $\lambda_{time} = 1.5$ and $\lambda_{time} = 2.5$, to demonstrate their impact on the smoothness of the different time sources obtained. The parameters were chosen in such a way to reduce the residual *i.e.* to maximize the explained data by having at least four sources. Figures 3.5.5 (a-b) represent the four time courses obtained with irTF-MxNE using the short window dictionary for the selected values of λ_{time}.

We show that for high values of λ_{time} (b), the solver is not able to capture the short transient component around 30 ms. While for a small value (a), the unmixing is not reliable since the light blue and the green source estimates catch the activity from the red source. Additionally, the time courses are not smooth. On the other hand, Figures 3.5.5 (c-d) represent the four time courses obtained with irTF-MxNE using the long window dictionary for the selected λ_{time}. The figure confirms that both parameters are not able to capture the transient effect after the stimulus, although the time courses are smooth. These four sub-figures reveal that one needs a short window to capture the transient effect of the brain signal (see

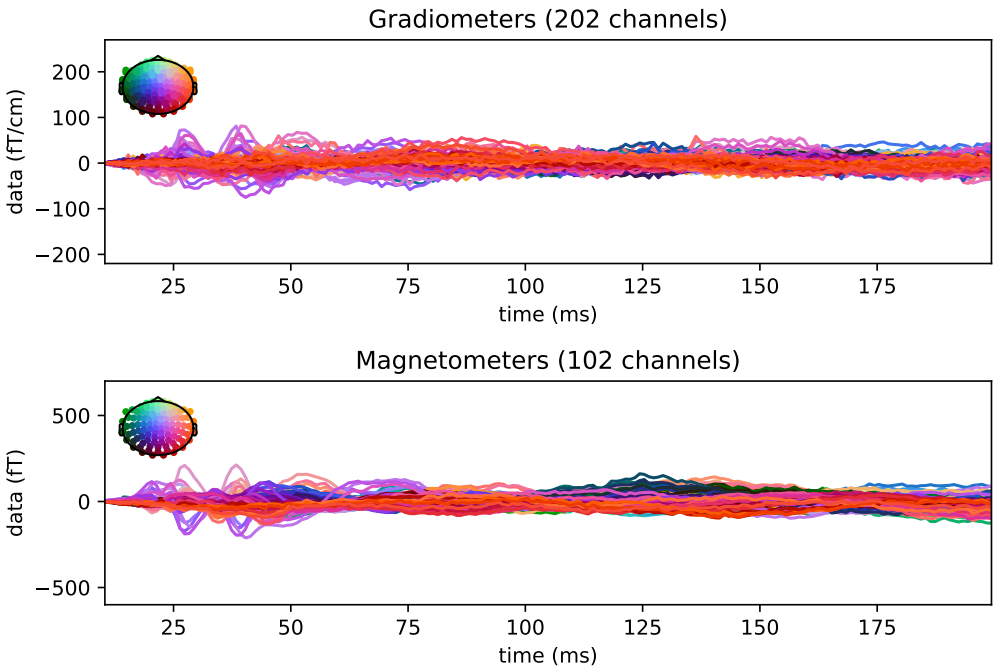


FIGURE 3.5.3: Residual of the somatosensory data after applying the multi-scale irTF-MxNE. The top left circle gives the position of the sensors over the head which are color-coded

Figure 3.5.4), while it needs to have a long window to capture the long waves and to have smooth source estimates. This result demonstrates how a combination of the two dictionaries is critical to acquire source estimates with high precision, but the hyperparameters need to be tuned as well, as shown in this Figure 3.5.5 that their values drastically change the results.

Moreover, Figure 3.5.5 (e) displays the amplitudes obtained with MxNE for five sources. As for MxNE, one is not able to obtain the four relevant sources unmixed (see for more demonstrative figures [GKH12b]). We notice that the light blue source in Figures 3.5.5 (a) to (d) appears as two separate sources in (e): light blue and purple. If we increase the λ parameter, we increase the amplitude bias due to the l_1 norm of the solver. If we set it too high ($\lambda = 50$) we obtain four sources, but the blue source which is relevant to the study would be removed and the duplicated purple source is kept. The last panel Figure 3.5.5 (f) displays the source estimates for dSPM values corresponding to the four locations of the sources obtained with irTF-MxNE. These sub-figures show that none of MxNE or dSPM solvers is able to obtain smooth sources without any leakage between the time courses.

Source estimation was then achieved using irTF-MxNE with the combination of the two dictionaries. Figure 3.5.6 shows source reconstruction using the multi-scale irTF-MxNE for the regularization parameters $\lambda_{space} = 28.5$ and $\lambda_{time} = 1.5$. Each source's location is marked by a sphere in Figure 3.5.6-left, and its amplitude over time is color-coded in the right panel. The results show a suitable succession of the sources. The transient source (red) is the only source explaining the

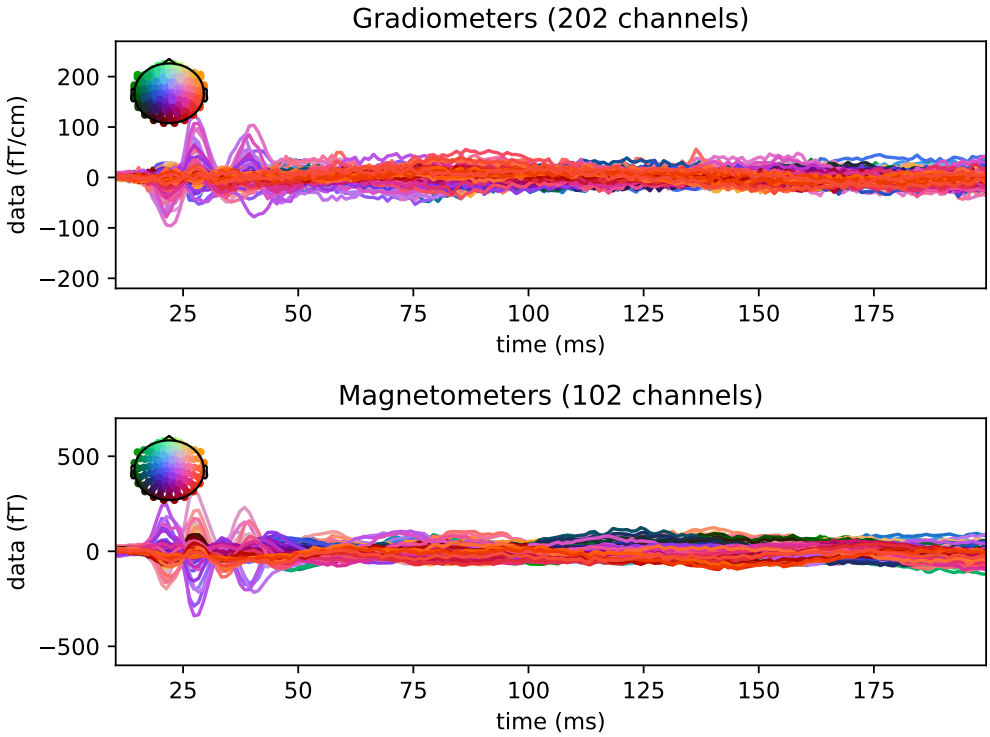


FIGURE 3.5.4: Residual of the somatosensory data after applying irTF-MxNE with a long window dictionary (window size = 64, time shift = 4). The transient part of the brain signal is left in the residual as it cannot be modeled by the long dictionary.

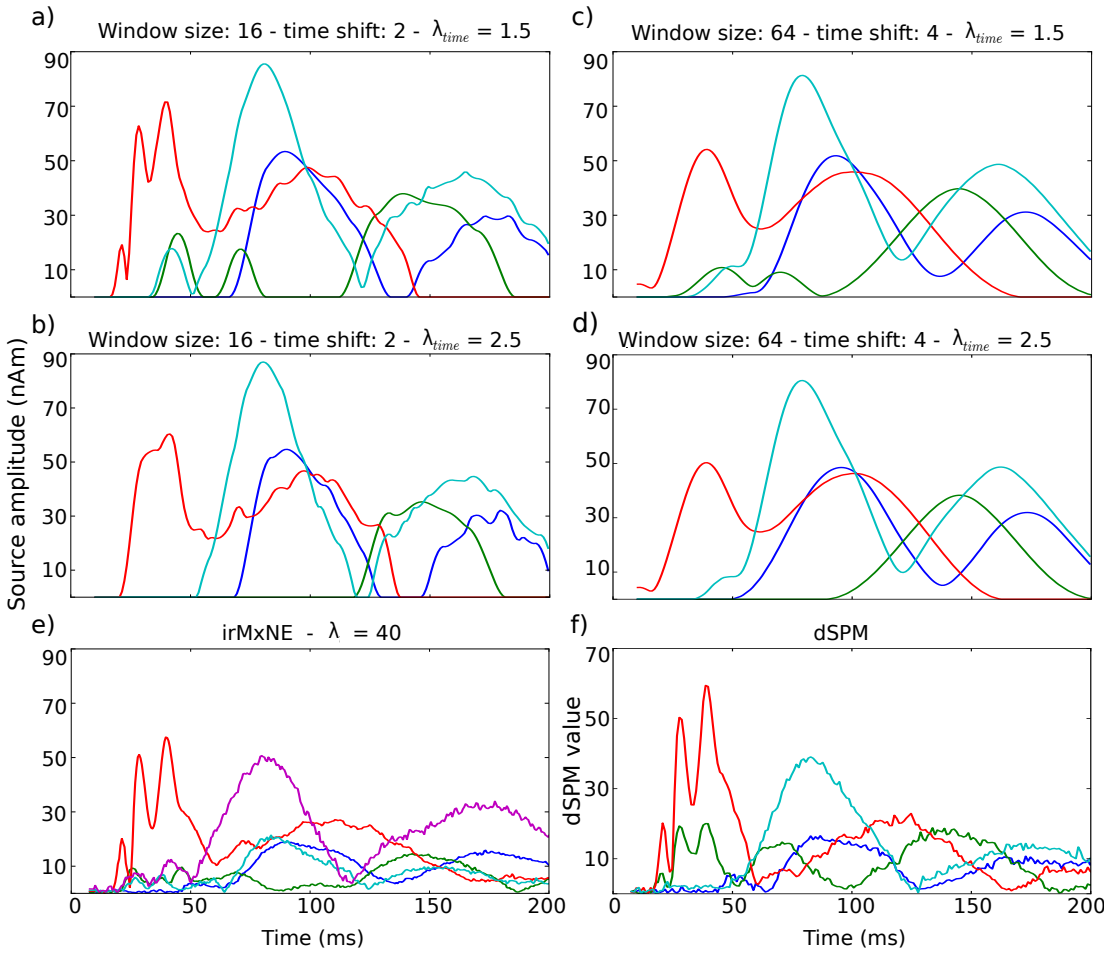


FIGURE 3.5.5: Source reconstruction using somatosensory data with different solvers. (a) - (b) irTF-MxNE on a small window dictionary with $\lambda_{time} = 1.5$ and $\lambda_{time} = 2.5$ respectively. (c) - (d) irTF-MxNE on a long window dictionary with $\lambda_{time} = 1.5$ and $\lambda_{time} = 2.5$ respectively. From (a) to (d) $\lambda_{space} = 28.5$ (e) MxNE for $\lambda = 40$ and (f) dSPM activation for the four activated sources.

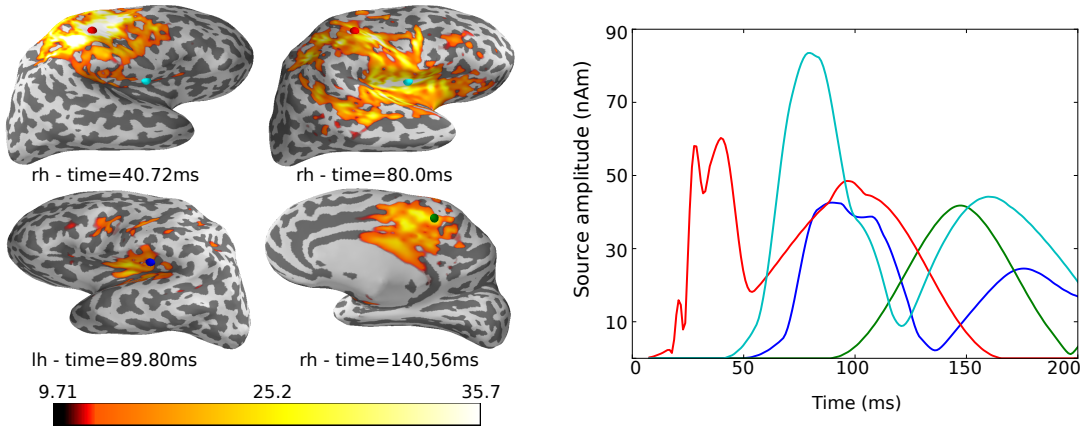
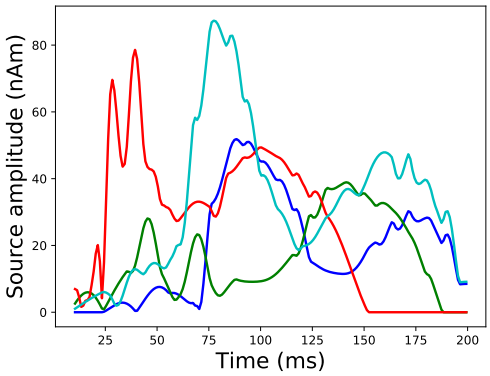


FIGURE 3.5.6: Source reconstruction using somatosensory data with a multi-scale irTF-MxNE. The solver estimates four sources for $\lambda_{space} = 28.5$ and $\lambda_{time} = 1.3$. The source locations marked with spheres in right (rh) and left (lh) hemisphere, and their corresponding activation are color-coded. The colorbar is over dSPM values which has no units as they are statistical values.

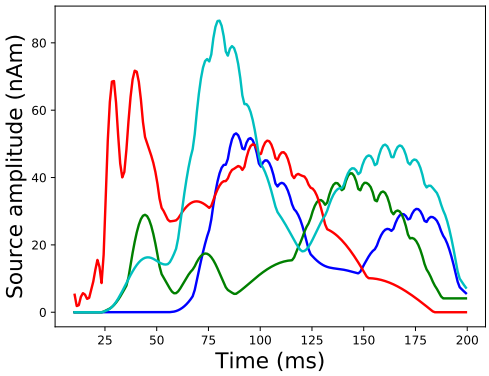
event related field until 48 ms. This red source corresponds to the contralateral primary somatosensory cortex (cS1) located in the postcentral gyrus of the parietal lobe (right hemisphere (rh)). The red sphere on the lateral view coincides with the smeared dSPM activation around 40 ms. The second source (light blue) corresponds to the secondary somatosensory cortex (cS2), and also occurs with dSPM activation around 80 ms. About 100 ms after stimulus, additional cortical sources are activated, such as ipsilateral secondary somatosensory cortex (iS2) (blue-lh), and contralateral medial wall (green-rh).

The multi-scale version of the MEG/EEG inverse problem in the TF domain does not only allow the capture of mixture of brain signals. An interesting point is the non-stationary aspect of the sources, which can be activated only for a short time window within a longer one. This multi-scale solver then allows us to analyze and reconstruct signals with variable characteristics over time. So far, all the results presented have used a Gabor transform by fixing its window length and the time shift. The Gabor transform is a special case of STFT (the discrete case), and the question is what if this dictionary is not the best choice for decomposing the data. The choice of the STFT was driven mostly by its flexibility of the choice of the dictionary being redundant or overcomplete. Moreover, its efficient implementation using FFT makes the STFT/iSTFT computation possible even with very redundant dictionaries.

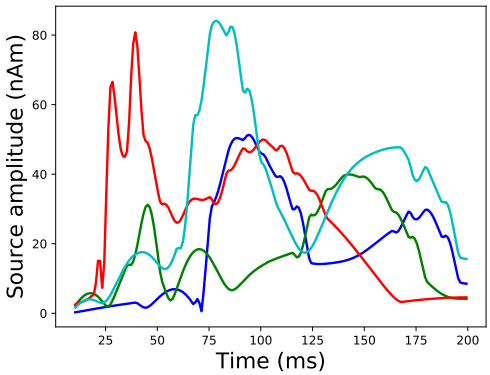
We investigated a second choice: MDCT (*cf.* Section 2.4.1 - page 24). The problem found with the MDCT is the fact that it is critically sampled. The sliding time windows are overlapping, so that the second half of one block coincides with the first half of the next block, *i.e.* the time shift is equal to half the window's



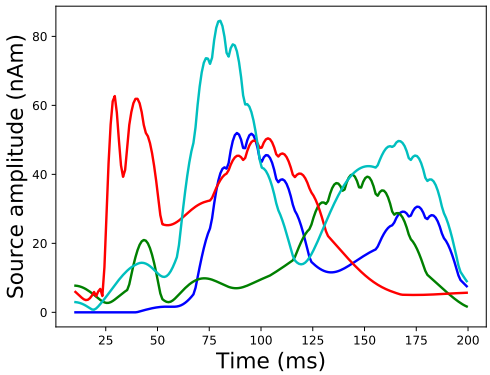
(A) MDCT: window size = 64-16, time shift = 32-8



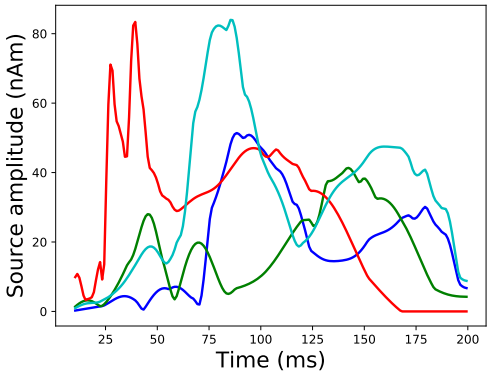
(B) STFT: window size = 64-16, time shift = 32-8



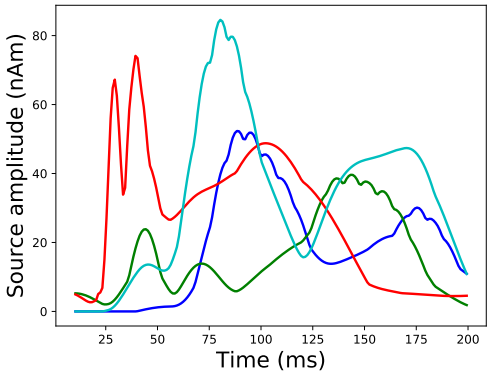
(C) MDCT: window size = 128-16, time shift = 64-8



(D) STFT: window size = 128-16, time shift = 64-8



(E) MDCT: window size = 128-64-16-8, time shift = 64-32-8-4



(F) STFT: window size = 128-64-16-8, time shift = 64-32-8-4

FIGURE 3.5.7: Comparison between MDCT and STFT using Somatosensory of the MIND dataset. MDCT is shown in the left column and STFT in the right column.

length. Figure 3.5.7 shows several mixtures of dictionaries for MDCT, but also for STFT if we set the time shift at half the window size. One can directly notice that it is harder to obtain both transient and smooth signals without any leakage between the time sources. MDCT is more sensitive to all the hyperparameters: the dictionary window length and the $\lambda_{space}/\lambda_{time}$. Even if we leave aside the trade-off that one needs to keep in mind between the size of the multi-scale dictionary and the computation time, by having multiple dictionaries concatenated together as in Figures 3.5.7-3.5.7e, MDCT is still not able to compete with STFT as in Figure 3.5.6.

3.6 Conclusion & Perspectives

In this chapter, we motivated the use of multi-scale dictionaries for the M/EEG inverse problem and presented our first contribution. The irTF-MxNE solver using a multi-scale dictionary allows to better capture the mixture of the MEG/EEG data. The non-convex optimization problem is solved by iteratively solving the convex weighted TF-MxNE problem using block coordinate descent combined with active set strategy to speed up the convergence.

The benefits of the multi-scale irTF-MxNE have been shown on simulated and MEG somatosensory data. Both experiments confirm that multi-scale irTF-MxNE improves the source estimates, in terms of reduced mixing of the time courses, smoothness and detection of both short transients and slower waves. The improvement regarding the active set size and amplitude bias is due to the non-convexity of the regularization methods. Hence, the multi-scale irTF-MxNE should be applied to data involving a mixture of signals, and when the aim is to acquire focal sources with non-stationary and smooth time courses.

Further work related to this chapter may address different points:

- The source localization in general is computed over an evoked response, *i.e.* in the MEG/EEG field, the evoked is the mean of several trials of the same experiment. The main purpose is to reduce the noise, *i.e.* increase the SNR of the signal. At a trial level, *i.e.* for a lower SNR, how can we improve the source localization?

- Optimization direction: the idea of incorporating screening techniques presented in a huge amount of papers can help to speed up the convergence, and so the reconstruction time [MGS17; MSG; FGS15; Ndi+15; Ndi+16; Ndi+17].

- Multi-scale irTF-MxNE improvement in terms of hyperparameter learning, *i.e.* estimation of the best parameters λ_{space} and λ_{time}. In the standard cases, these parameters are selected by cross-validation or sometimes by using the discrepancy principle. A key contribution in this direction would be to use Bayesian inference techniques to estimate those regularization parameters in

a composite norms setting. This problem is addressed in the following chapter but not in the TF domain.

Chapter 4

Bridges between Bayesian models and sparsity inducing norms

Parts of this chapter have been published in the following:

- **Y. Bekhti**, R. Badeau, and A. Gramfort, "Hyperparameter estimation in maximum a posteriori regression using group sparsity with an application to brain imaging," Signal Processing Conference (EUSIPCO), pp. 246-250, 2017.
- **Y. Bekhti**, F. Lucka, J. Salmon, and A. Gramfort, "A hierarchical Bayesian perspective on majorization-minimization for non-convex sparse regression: application to M/EEG source imaging," ArXiv preprint, (submitted).

4.1 Introduction - General concepts

This chapter presents a different perspective on the MEG/EEG inverse problem. It tries to bridge the gap between two communities both interested in sparse models for solving inverse problems. As mentioned several times so far in this thesis, sparsity has emerged as a key concept to solve inverse problems, not only the MEG/EEG inverse problem, but also tomographic image reconstruction, deconvolution, or inpainting. The idea is also well established to regularize high dimensional regression problems in the field of machine learning. There are mainly two routes to introduce sparsity to such problems.

The first route, embraced by the optimization community and frequentist statisticians, is to promote sparsity using convex optimization theory. This line of work has led to now mature theoretical guarantees [FR13] when using regularization functions based on ℓ_1 norm and other convex variants [Tib96a]. In particular, it has been popularized in the signal processing community under the name of compressed sensing [CW08] when combined with incoherent measurements.

There are however some limitations of sparsity promoting convex penalties based on the ℓ_1 norm. All the features (also called regressors, atoms or sources depending on the terminology of the community) involved in the solution form what is called the support of the solution. Convex penalties can fail to identify the correct support in the presence of highly noisy data, but also in low noise setups if the forward operator (referred to as design matrix in statistics) is poorly conditioned. Convex regularizations also lead to a systematic underestimation bias in the amplitude of the coefficients [Osh+06; CWB08; Cha07; SCY08; CHS17].

To address these limitations of ℓ_1-type models, reweighted schemes have been proposed [CWB08; GRC09; Rak11a; ZR11b; Str+16], of which the Adaptive Lasso [Zou06] is the most commonly used in the statistics community: Starting from the Lasso estimator, which amounts to regressing with a standard ℓ_1-norm as a regularizer (this estimator is sometimes referred to as Basis Pursuit Denoising (BPDN) [CDS98] in signal processing), the Adaptive Lasso solves a sequence of weighted Lasso problems, where at each iteration the weights are chosen such that the strongest coefficients are less and less penalized. From the optimization point of view, such an iterative scheme can be derived from so-called Majorization-Minimization (MM)

strategies [LHY00; SSW+10]. The idea behind MM is to minimize the objective function by successively minimizing upper bounds that are easier to optimize. Many well-known optimization approaches can be interpreted as instances of MM, e.g., simple gradient descent or proximal algorithms [CP11], expectation-maximization (EM) [DLR77], and difference-of-convex (DC) programming techniques [HT99]. More recently, re-weighted ℓ_1-norm schemes based on MM principle have been particularly popular to handle concave, hence non-convex regularizations such as $\ell_{0.5}$-quasi-norms or logarithmic functions. As such, these schemes are prone to converging to a local minimum determined by the initial, uniformly weighted ℓ_1-norm solution (i.e., the Lasso estimator) that constitutes the first iterate. This first route has been defined in more details in Chapter 2 and Chapter 3.

The second route to introduce sparsity formulates the regression problem in a Bayesian framework and uses Hierarchical Bayesian Modeling (HBM) [Mac03] for the inference. The common way to formulate HBMs is to consider the variance parameters of Gaussian prior models as additional random variables which have to be estimated from the data as well. Their prior distributions are referred to as hyper-priors. Plausible solutions to the regression problem that both fit data and the *a priori* assumption of sparsity are explicitly characterized as multiple distinct modes of the posterior distribution. This characterization is the Bayesian analogue to local minima in variational regression approaches when working with non-convex functionals. Different strategies to infer a point estimate for the parameters of interest from the *a posteriori* distribution then lead to different algorithmic frameworks, for instance Variational Bayesian approaches [Mac03; Jor+99; Sat+04; Fri+08; SB15], Sparse Bayesian Learning (SBL) approaches (also referred to as type-I or type-II maximum likelihood estimates) [Tip01; WR04; WN09a; ZR11b] and fully-Bayesian strategies [Cal+09; Luc+12a].

This chapter focuses on the later one for a non-standard type of HBM examined in [Luc14] that combines a non-Gaussian prior with an ℓ_1-type energy function with a specific Gamma hyper-prior. For this HBM, a simple alternating scheme to compute full maximum *a posteriori* (MAP) estimates leads to exactly the same sequence of problems solved by MM applied to $\ell_{1/2}$-type regularizations. With this observation made, it is natural to revisit and improve these MM schemes by leveraging the ability of the Bayesian framework to explore the modes of the posterior distribution by MCMC schemes [RC05; KS05]. This does not only mitigate the aforementioned initialization-dependence of MM, but more importantly, it offers insights into the structure and importance of potentially multiple plausible sparse solutions. Yet, the benefit comes at the cost of additional computational efforts.

This chapter is organized as follows: First, it presents in a unified perspective both routes to sparsity, i.e., reweighted ℓ_1 MM schemes and specific HBMs. We show that a particular optimization-based inference strategy recovers the MM algorithm. It then describes an HBM inference strategy based upon an MCMC sampling and shows on simulated and experimental MEG/EEG datasets how these

stochastic MCMC-based techniques do not only help to improve upon deterministic approaches but also help to reveal multiple plausible solutions to the inverse problem. This analysis leads to an Uncertainty Quantification (UQ) of the support recovery of non-convex sparse regression problems that provides very useful complementary information, in particular for very ill-conditioned and underdetermined applications like MEG/EEG source localization.

4.2 Lp hyper-models

In Bayesian statistics, a hyperprior is a prior distribution on a hyperparameter, that is, on a parameter of a prior distribution. Firstly, the use of a hyperprior allows one to express uncertainty in a hyperparameter: taking a fixed prior is an assumption, varying a hyperparameter of the prior allows one to do sensitivity analysis on this assumption, and taking a distribution on this hyperparameter allows one to express uncertainty in this assumption: "assume that the prior is of this form (this parametric family), but that we are uncertain as to precisely what the values of the parameters should be" [BS01].

A popular choice of hyperprior is the gamma distribution with α and β its corresponding parameters:

$$p(\lambda) = \frac{\beta^\alpha}{\Gamma(\alpha)} \lambda^{\alpha-1} \exp(-\beta\lambda) \mathbf{1}_{\mathbb{R}^+}(\lambda), \ \lambda \in \mathbb{R} \tag{4.1}$$

where Γ is the gamma function. In the following of this chapter, the hyperprior is always a gamma distribution.

4.3 Hyperparameter estimation in the variational formulation

This section investigates the estimation of the hyperparamter λ in the variational formulation. One can notice that hyperparameter setting is a classical statistics problem for which a number of solutions have been proposed. In signal processing, the Akaike information criterion (AIC) and Bayesian information criterion (BIC) criteria are quite popular techniques historically [Sch+78]. The SURE-based techniques [Ste81] have also been quite popular and recently explored for denoising and compressed sensing applications [LBU07; GD15]. In a standard supervised machine learning setup with independent and identically distributed (i.i.d.) observations, Cross-Validation (CV) is the reference approach. Also, the Bayesian approach suited for probabilistic models offers a principled way to estimate hyperparameters using hyperpriors that introduce softer constraints than solutions with fixed parameter values. This benefit yet usually comes at a price in terms of

computational cost. Finally, in a number of real scenarios, humans end up setting hyperparameters, as they can have some expert knowledge that can correct model mismatch.

In statistical machine learning, a hyperparameter typically aims at limiting overfitting by controlling the model complexity. In the particular case of regularized regression, classically a scalar parameter balances between the data fit and the penalty term. When using sparse regression, this parameter affects the sparsity of the solution, *i.e.* how many covariates or regressors are used.

With CV, some independent observations are left out of the inference and the hyperparameter values that yield the best prediction performance on this data are selected. A search for the best parameter can be done with a time consuming exhaustive grid-search, smooth optimization (see [Ped16] and references therein), sequential or even random search [Ber+11; BB12]. The CV approach however needs the i.i.d. assumption to be fulfilled, which is not always the case in practice, e.g. when working with signals or arrays of sensors as in the case of our application to brain imaging.

To keep it as a hierarchical Bayesian model problem and following a recent paper of Pereyra [PBDF15], we consider a HBM and propose to use a MAP estimation for the hyperparameters.

This thesis is particularly interested in the high-dimensional regression setting using Group-Lasso-like structured sparsity as seen so far. In the literature a number of approaches have been proposed and MAP estimates that boil down to penalized regression with smooth or non-smooth penalties are the standard approaches employed by neuroscientists [Hau+08; OHG09b; BVVN09; WN09b; GKH12b; Luc+12b; VS+09].

In a variational formulation, the value of the hyperparameter λ depends on the problem at hand, the noise level, and on the choice of regularization $\mathcal{P}(\mathbf{X})$. Finding a way to estimate the hyperparameter with minimal user intervention is therefore particularly important, as it makes a comparison between different models and regularization easier.

Recently *Pereyra et al.* [PBDF15] proposed a strategy for hyperparameter estimation in the context of MAP inference when the prior or the regularizer is a k-homogeneous function. The regularizer $\mathcal{P}$ is a k-homogeneous function if there exists $k \in \mathbb{R}^+$ such that:

$$\mathcal{P}(\eta \mathbf{X}) = \eta^k \mathcal{P}(\mathbf{X}), \quad \forall \mathbf{X} \in \mathbb{R}^{S \times T} \quad \text{and} \quad \forall \eta > 0.$$

The k-homogeneous condition is satisfied for all $\ell_{p,q}$ mixed norms. We focus on the estimation of the hyperparameters for hierarchical Bayesian models yielding convex $\ell_{2,1}$ ($\mathcal{P}(\mathbf{X}) = \|\mathbf{X}\|_{2,1}$) or non-convex $\ell_{2,0.5}$ penalties, which are respectively 1-homogeneous and 0.5-homogeneous. The non-convex penalization is solved using

iterative re-weighted convex optimization schemes, *i.e.* each iteration is a weighted $\ell_{2,1}$-norm as described in Section 4.4.1.

In [PBDF15], the fixed point strategy proposed is validated on an image denoising problem using an analysis prior, *i.e.* where the solution is not sparse but has a sparse representation in some transformed domain. This section illustrates and explains why the method from [PBDF15] cannot be used out-of-the-box when using a synthesis prior for an under-determined problem. A synthesis prior is when the solution itself is sparse.

4.3.1 Hierarchical Bayesian modeling and reformulation

The result shown in [PBDF15] and adapted to our problem and the notations used in this manuscript is the following. Using a joint MAP estimator of λ and $\mathbf{X}$, it states that $\hat{\lambda}$ should satisfy:

$$\hat{\lambda} = \frac{ST/k + \alpha - 1}{\mathcal{P}(\hat{\mathbf{X}}_{\hat{\lambda}}) + \beta} \quad , \tag{4.2}$$

where $\hat{\mathbf{X}}_{\hat{\lambda}}$ is the solution of Equation (2.13) (page 28) for $\lambda = \hat{\lambda}$. In [PBDF15], it is further suggested to set α and β to 1.

Looking at Equation (4.2), one can observe that if ST is big, which is the case for high dimensional problems, the numerator can significantly dominate the denominator, especially if the estimate $\hat{X}$ is very sparse. In practice using Equation (4.2) in this scenario results rapidly in huge values of λ and empty supports. This issue is much less critical when using an analysis prior for denoising as in [PBDF15], as the size of the unknown coefficients is in this case NT, where $NT \ll ST$.

As reported earlier, the update of the regularization parameter λ as in (Equation (4.2)) is not suitable for the synthesis prior $\mathcal{P}(\mathbf{X})$. The issue is due to the over-scaled numerator compared to the denominator. When the problem is important (as in [PBDF15]) - ST is big, whereas the support in $\mathbf{X}^\star$ is small - the estimated parameter λ then explodes, resulting in an empty support.

To overcome this problem, we propose to rewrite the objective function in such a way that we obtain the same solution $\mathbf{X}$ but with a multiplicative factor $\frac{\lambda}{ST}$. The new equivalent formulation can be written as:

$$\hat{\mathbf{X}} = \arg\min_{\mathbf{X}} \frac{ST}{2} \|\mathbf{M} - \mathbf{G}\mathbf{X}\|_{Fro}^2 + \lambda \mathcal{P}(\mathbf{X}) \tag{4.3}$$

Note that this is just a reparametrization of Equation (2.13). In practice, this boils down to multiplying $\mathbf{M}$ and $\mathbf{G}$ by $\sqrt{ST}$. However this only solves one difficulty in the parameter's update. Another disadvantage is that none of the parameters in Equation (4.2) takes into account the scale of $\mathbf{G}$. The next section explains

how to use results from convex optimization to properly calibrate the hyperprior parameters α and β given $\mathbf{M}$, $\mathbf{G}$ and $\mathcal{P}$.

4.3.2 Setting hyperpriors with a single hyperparameter

As in [PBDF15], Gamma hyperpriors are used to derive two iterative algorithms that simultaneously estimate a single hyperparameter λ and the entries of $\mathbf{X}$, yet the values of α and β are still to be defined. In [PBDF15], it is suggested to set α and β to 1, which turns out to be inappropriate for underdetermined inverse (deconvolution) problems as our MEG/EEG brain imaging problem of interest.

A first observation is that α and β should default to reasonable values and be insensitive to trivial changes in matrix $\mathbf{G}$ such as scaling, *i.e.* multiplying $\mathbf{G}$ by a scalar. This is the problem we investigate now.

In Equation (4.2), the numerator would not be affected by a rescaling of $\mathbf{G}$. However, the denominator that contains $\mathcal{P}(\mathbf{X}_{\lambda^\star})$ would. To make the estimation robust to changes of $\mathbf{G}$ such as scaling, one therefore needs to modify the numerator, hence make α a function of $\mathbf{G}$. Setting α to 1 independently of the problem, as in [PBDF15], is certainly inadequate.

In order to set the value of α, we propose to take advantage of the fact that if $\mathcal{P}(\mathbf{X}) = \|\mathbf{X}\|_{2,1}$, one can analytically compute λ_{max}, which is defined as the smallest regularization parameter for which the solution is zero [Bac+12]. It is given by:

$$\lambda_{max} = \|\mathbf{G}^\top \mathbf{M}\|_{2,\infty} = \max_i \|(\mathbf{G}^\top \mathbf{M})[i,:]\|_2. \tag{4.4}$$

Parameter λ can therefore be parametrized as a fraction, or a percentage, of λ_{max}. This allows us to have a good a priori guess on the peak of the gamma distribution. We set the peak, *a.k.a.* the mode, to $mode = \tau \times \lambda_{max}$, with $\tau \in [0, 1]$.

Once the mode is known, it is straightforward to fix the value of α: $mode = \frac{\alpha - 1}{\beta}$ for $\alpha \geq 1$. From now on we fix α as:

$$\alpha = mode \times \beta + 1 = \tau \times \lambda_{max} \times \beta + 1 \ . \tag{4.5}$$

Concerning the parameter β, for our specific MEG/EEG problem of interest we fix it so that 99% of the probability density of the gamma distribution is between 20% and 70% of λ_{max}. This is motivated by the fact that in our case solutions are expected to be extremely sparse, with only a handful of active brain regions. This is of course application specific.

4.3.3 Estimation of a vector of hyperparameters

The penalizations of the form $\mathcal{P}(\mathbf{X}) = \|\mathbf{X}\|_{2,\cdot}$ are separable in S groups of coefficients. As only a few groups are expected to be active, a natural idea is to penalize less the important groups. To do this, we propose to estimate one parameter per group of coefficients or row of $\mathbf{X}$ using the convex $\ell_{2,1}$ penalization. Let us now derive a result similar to Equation (4.2) but in the more general case of one λ per group.

Rewriting Equation (2.13) in the MAP framework leads to:

$$\mathbf{X}^\star = \arg\max_{\mathbf{X}} p(\mathbf{X}, \mathbf{M}|\lambda) = \arg\max_{\mathbf{X}} p(\mathbf{M}|\mathbf{X})p(\mathbf{X}|\lambda) \tag{4.6}$$

where $p(\mathbf{M}|\mathbf{X})$ is the likelihood function corresponding to the first term in Equation (2.13) and $p(\mathbf{X}|\lambda)$ is the regularization corresponding to the second term in Equation (2.13). This Bayesian formulation requires to compute the normalization factor $C(\lambda)$ in $p(\mathbf{X}|\lambda) = \exp(-\lambda\mathcal{P}(\mathbf{X}))/C(\lambda)$. Computing this constant $C(\lambda)$ in general is intractable as it involves an integration. Yet [PBDF15] showed that it admits an exact closed-form when the penalization is k-homogeneous as $C(\lambda) = D\lambda^{-ST/k}$ where $D = C(1)$ is a constant independent of λ [PBDF15].

We now propose a joint-MAP estimation with $\lambda \in \mathbb{R}^S$. We look for $(\mathbf{X}^\star, \lambda^\star) \in \mathbb{R}^{(S\times T)} \times \mathbb{R}^S$ which maximizes $p(\mathbf{X}, \lambda|\mathbf{M})$. A sufficient condition of optimality is given by:

$$(0_{(S\times T)}, 0_S) \in -\partial_{\mathbf{X},\lambda} \log p(\mathbf{X}^\star, \lambda^\star|\mathbf{M}) \tag{4.7}$$

i.e.

$$\begin{aligned} 0_{S\times T} &\in -\partial_{\mathbf{X}} \log p(\mathbf{X}^\star, \lambda^\star|\mathbf{M}), \\ 0 &\in -\partial_{\lambda_i} \log p(\mathbf{X}^\star, \lambda^\star|\mathbf{M}) \quad \forall i, \end{aligned} \tag{4.8}$$

where $\partial_{\mathbf{X},\lambda}$ is the set of subgradients (the subdifferential).

The optimization over $\mathbf{X}$ at iteration k satisfies Equation (4.3):

$$\mathbf{X}^{(k)} = \arg\min_{\mathbf{X}\in\mathbb{R}^{S\times T}} \frac{ST}{2}\|\mathbf{M} - \mathbf{G}\mathbf{X}\|_{Fro}^2 + \sum_i \lambda[i]^{(k-1)}\|\mathbf{X}[i,:]\|.$$

The next step is to optimize over $\lambda_i, \forall i$. Equation (4.8) leads to:

$$0 \in -\partial_{\lambda[i]} \log p(\mathbf{X}^{(k)}, \mathbf{M}|\lambda) - \partial_{\lambda[i]} \log p(\lambda). \tag{4.9}$$

Using $p(\mathbf{X}^{(k)}, \mathbf{M}|\lambda) = p(\mathbf{M}|\mathbf{X}^{(k)})p(\mathbf{X}^{(k)}|\lambda)$, one has that:

$$-\partial_{\lambda[i]} \log p(\mathbf{X}^{(k)}, \mathbf{M}|\lambda) = -\partial_{\lambda[i]} \log p(\mathbf{X}^{(k)}|\lambda). \tag{4.10}$$

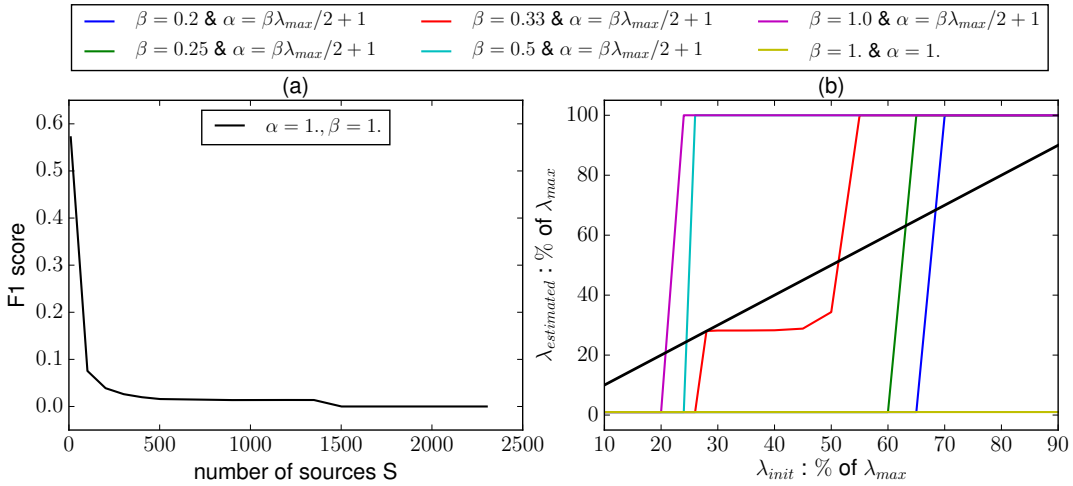


FIGURE 4.3.1: (a) Source identification results for different numbers of sources measured with F1 score using $\alpha = 1$ and $\beta = 1$. The higher the number of regressors, the worse the performance is. (b) Estimated λ as a function of λ_{init} for different values of a and b. The red curve for $\beta = 0.33$ gives the best plateau, which demonstrates that (a, b) shall be carefully adjusted.

We then use the normalization factor $C(\lambda)$ which gives:

$$-\partial_{\lambda[i]} \log p(\mathbf{X}^{(k)}, \mathbf{M}|\lambda) = \|\mathbf{X}[i,:]\| + \partial_{\lambda[i]} \log C(\lambda) \tag{4.11}$$

and

$$\partial_{\lambda[i]} \log C(\lambda) = \frac{-ST}{k\lambda[i]}. \tag{4.12}$$

Regarding the second term in Equation (4.9), Equation (4.1) yields $-\partial_{\lambda_i} \log p(\lambda) = -\frac{\alpha-1}{\lambda_i} + \beta$. Completing the derivations, the equation for each λ_i, $i \in [1 \dots S]$, reads:

$$\lambda_i^\star = \frac{ST/k + \alpha - 1}{\|X_{\lambda^\star}[i,:]\| + \beta} \quad . \tag{4.13}$$

4.3.4 Experiments

Simulation study

We generated a simulation dataset with $N = 302$ sensors, $T = 190$ time samples and $S = 1500$ sources. Four sources were randomly selected to be active with realistic waveforms obtained from the MIND dataset [Wei+07]. The linear forward operator $\mathbf{G}$ was a random matrix, whose columns were normalized to 1. Two levels of white noise were added to the simulation. We always used $\tau = 0.5$.

In order to illustrate the issue when using a synthesis prior for large problems, we run the estimation of the hyperparameter λ as suggested in [PBDF15] using the 0.5-homogeneous non-convex prior. Figure 4.3.1-(a) shows the F1 score [1] of

[1] The F_1 score is the harmonic mean of precision and recall: $F_1 = 2.\frac{precision.recall}{precision+recall}$

the source reconstruction (1 for good reconstruction and 0 for bad). The source estimation is failing for almost all the range of data size. Figure 4.3.1-(b) shows the results after reformulating the problem with different settings of α and β. One can notice that a setting as in [PBDF15] with $\alpha = 1$ and $\beta = 1$ always gives an estimated λ around 1% of λ_{max}, which is not promoting the sparsity we are looking for in this kind of setting. For this aim, we varied the values of β and computed α as defined before. Figure 4.3.1-(b) shows that for most values of β we have rather a too low estimation of $\lambda \approx 1\%$ or a too high $\lambda \geq 100\%$ resulting in zero source found active. Interestingly, setting $\beta = 1/3$ gives a plateau at $\hat{\lambda}$ close to $0.3\lambda_{max}$. This is an evidence of a clear fixed point for the iterative process $\lambda^{(t+1)} = f(\lambda^{(t)})$, where f is the update rule of λ in Equation (4.2). We use $\beta = 1/3$ from now on and its corresponding α.

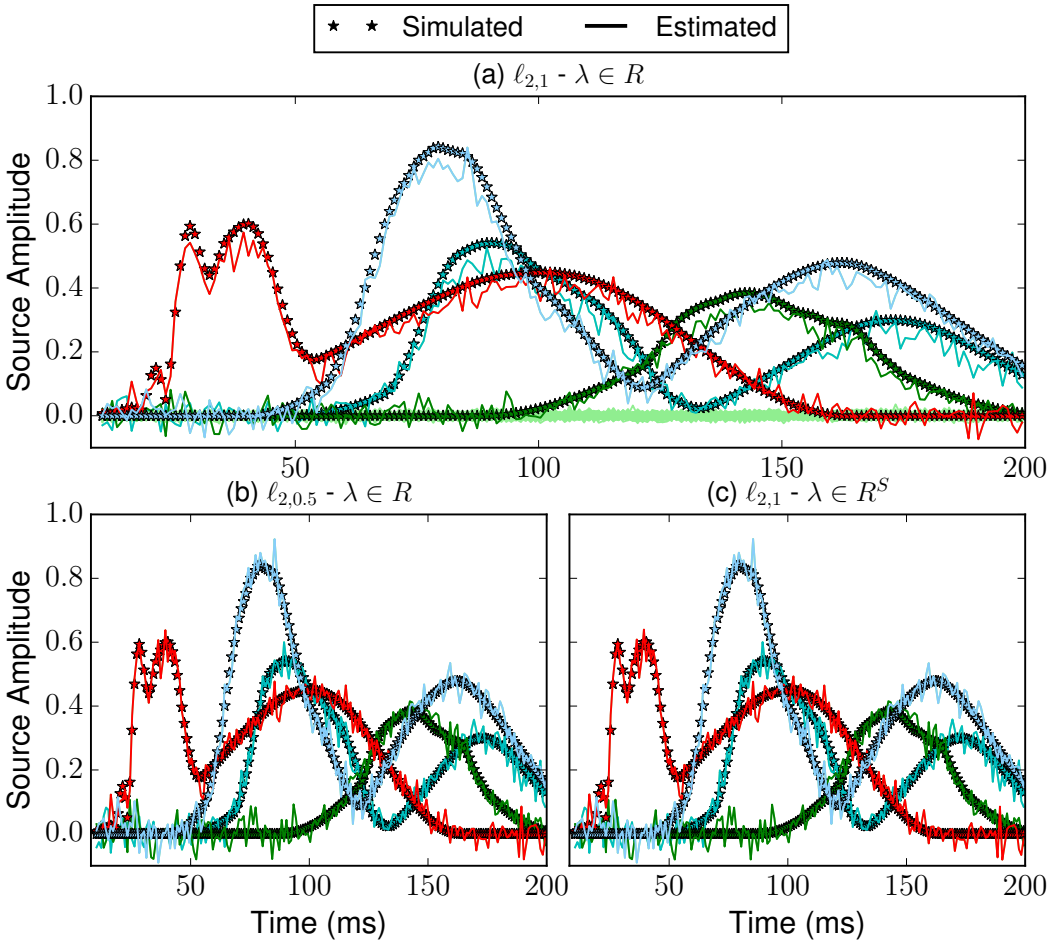


FIGURE 4.3.2: Source reconstruction on simulated data. (a): Source estimates obtained using $\ell_{2,1}$ with one λ. The solution is not sparse enough (zero sources in light green) and there is an amplitude bias between the exact amplitudes (stars) and the estimated ones (raw lines). (b): Good reconstruction of the four sources using $\ell_{2,0.5}$ and one λ, which is equivalent to the reconstruction using the $\ell_{2,1}$ norm with $\lambda \in \mathbb{R}^S$ (c). Each of the four sources is encoded with a different color.

Figure 4.3.2 represents the simulated sources with stars and the estimated ones

with plain lines. Figure 4.3.2-(a)-(b) display results with the $\ell_{2,1}$ and $\ell_{2,0.5}$ norms respectively, using one hyperparameter initialized to $\lambda = 0.5\lambda_{max}$. One can see that in Figure 4.3.2-(a), the $\ell_{2,1}$ norm recovers the four sources with an amplitude bias (the estimated amplitude is lower than the exact one), and that several sources shown in light green are almost flat around zero but still found as active sources. There is no way to reduce the support without losing one of the four simulated sources, *i.e.* the $\ell_{2,1}$ norm with one hyperparameter fails to recover the exact simulated sources. The $\ell_{2,0.5}$ norm in (b) estimates the exact four source amplitudes without amplitude bias thanks to the non-convexity [Str+16]. On the other hand, Figure 4.3.2-(c) shows the results for the convex penalty using one hyperparameter per source. It can be seen that it is qualitatively equivalent to the non-convex penalty.

The advantage of having one hyperparameter per source is to pick up only the sources involved in the measurement M and drop the extra almost-zero sources visible in Figure 4.3.2-(a) (light green). This extension produces sparser results and less amplitude bias without casting the problem as non-convex. This figure also suggests a link between the non-convex prior and one hyperparameter per source. As the non-convex prior is an iterative procedure estimating an internal weight to produce a better solution, the fact to have one hyperparameter per source can also be seen as a weight to define better source estimates. A more accurate study of this can be found in the next section 4.4.

Experimental results with MEG auditory data

We applied the estimation of a single hyperparameter and a hyperparameter per source using the convex $\ell_{2,1}$ penalty on a real open dataset (MNE sample dataset [Gra+14]). It corresponds to a dataset with $N = 305$ sensors, $T = 55$ time samples and $S = 7498$ sources. Figure 4.3.3 shows the source amplitudes of the two auditory sources and their positions in the brain when estimating a hyperparameter per source. When using a single hyperparameter on the convex norm $\ell_{2,1}$, multiple spurious sources are found as active which replicates the simulation on Figure 4.3.2-(a). These source estimates in Figure 4.3.3 correspond to the M100 peak (peak around 100 ms) generated in the vicinity of the bilateral auditory cortices in superior temporal gyri (the relevant auditory area).

4.4 Link between MM and special case of HBM

We start this section by recalling how majorization-minimization works when addressing variational formulations with concave, hence non-convex, regularization. It is followed by an introduction to hierarchical Bayesian models with Gamma hyper-priors. Then, we explain how these seemingly different approaches can lead to the exact same regression algorithm. From this, we detail how different

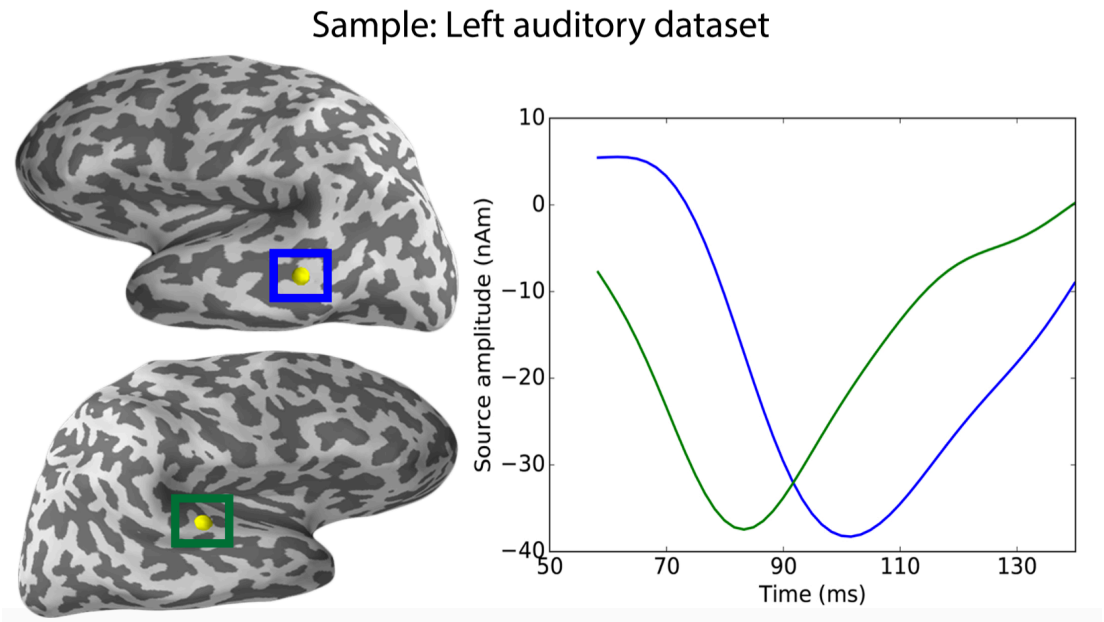


FIGURE 4.3.3: Source reconstruction on MEG auditory data (sample dataset [Gra+14]). Source amplitude of two sources (blue and green) in the right panel and their corresponding positions in the brain on the left.

Bayesian inference strategies using MCMC sampling can more precisely explore the landscape of the posterior distribution of the HBM model and provide multiple possible solutions to the sparse regression problem compared to MM.

4.4.1 Majorization-Minimization: MM

Majorization-Minimization (MM) strategies consist in replacing a difficult optimization problem with a series of easier ones that are obtained by upper bounding the objective function, often by a convex majorant. In the context of inverse problems or high-dimensional statistics using sparsity constraints, MM has been successfully applied to address non-convex regularization terms. An example is the regression model with $\ell_{2,p}$-quasi-norms regularization over the groups when $0 < p < 1$: the desired estimate $\hat{\mathbf{X}}$ is defined as one of potentially multiple minimizers of:

$$\hat{\mathbf{X}} \in \underset{\mathbf{X} \in \mathbb{R}^{SO \times T}}{\arg\min} \frac{1}{2} \|\mathbf{M} - \mathbf{G}\mathbf{X}\|_{\text{Fro}}^2 + \lambda \sum_{i=1}^{S} \|\mathbf{X}[i]\|_{\text{Fro}}^p \ , \tag{4.14}$$

similarly to Equation (2.13) (page 28), where $\lambda > 0$ is the regularization parameter balancing the data fit and the penalty term, and $\mathbf{X}[i]$ defines the block of source i ($\mathbb{R}^T$ if $O = 1$ or $\mathbb{R}^{3\times T}$ if $O = 3$). One possible MM approach to solve Equation (4.14) with $p = 0.5$ would consist of minimizing a sequence of non-smooth convex surrogate functions where the non-convex regularization (irMXNE solver) is replaced

by a weighted $\ell_{2,1}$ norm similar to MxNE solver [Str+16]. In each iteration, the weights are derived from the current estimate of $\mathbf{X}$.

Due to the concavity of the non-decreasing function $\mathbf{X} \mapsto \sqrt{\|\mathbf{X}\|_{\text{Fro}}}$, it is upper bounded by its tangent and a first order Taylor expansion at the current estimate $\mathbf{X}[i]$ provides an upper bound that can be used to construct the non-smooth convex surrogate problem. By solving this sequence of surrogate problems, the value of the non-convex objective function is guaranteed to decrease. However, due to the non-convexity, only convergence towards a local minimum can be guaranteed.

For the problem in Equation (4.14) with $p = 0.5$, the k^{th} iteration of the MM scheme reads:

$$\mathbf{X}^{(k)} \in \underset{\mathbf{X} \in \mathbb{R}^{SO \times T}}{\arg\min} \frac{1}{2}\|\mathbf{M} - \mathbf{G}\mathbf{X}\|_{\text{Fro}}^2 + \lambda \sum_{i=1}^{S} \frac{\|\mathbf{X}[i]\|_{\text{Fro}}}{\mathbf{w}^{(k-1)}[i]}, \tag{4.15}$$

with:

$$\mathbf{w}^{(k-1)}[i] = 2\sqrt{\|\mathbf{X}^{(k-1)}[i]\|_{\text{Fro}}} \ .$$

As each weight $\mathbf{w}^{(k)}[i]$ is a non-decreasing function of $\|\mathbf{X}^{(k)}[i]\|_{\text{Fro}}$, sources with high amplitudes in one iteration will be less penalized in the next iteration and can better explain the data $\mathbf{M}$. Sources for which $\|\mathbf{X}^{(k)}[i]\|_{\text{Fro}} = 0$ at a certain iteration k are effectively pruned from the model for all following iterations. Using MM therefore leads to a solution that explains the data with fewer active locations i compared to a standard $\ell_{2,1}$ norm regularized solution.
Note that a default initialization consists in setting $\mathbf{w}^{(0)}[i] = 1, \forall i \in [1, \cdots, S]$ [Str+16].

To exploit existing fast solvers for the $\ell_{2,1}$ regularized problems [Str+16; Ndi+15], we reformulate the weighted subproblem and apply the weights by scaling the matrix $\mathbf{G}$ with a diagonal matrix $\mathbf{W}^{(k)} \in \mathbb{R}^{SO \times SO}$ given by:

$$\mathbf{W}^{(k)} = \text{diag}(\mathbf{w}^{(k)} \otimes \mathbf{1}_O) \ , \tag{4.16}$$

where $\mathbf{w}^{(k)} \in \mathbb{R}^S$, $\mathbf{1}_O \in \mathbb{R}^O$ is a vector of ones and $\otimes$ is the Kronecker product. Defining $\tilde{\mathbf{G}}^{(k)} = \mathbf{G}\mathbf{W}^{(k-1)}$, the reformulated problem reads:

$$\tilde{\mathbf{X}}^{(k)} \in \underset{\mathbf{X} \in \mathbb{R}^{SO \times T}}{\arg\min} \frac{1}{2}\|\mathbf{M} - \tilde{\mathbf{G}}^{(k)}\mathbf{X}\|_{\text{Fro}}^2 + \lambda \sum_{i=1}^{S} \|\mathbf{X}[i]\|_{\text{Fro}} \ . \tag{4.17}$$

The convergence of each weighted $\ell_{2,1}$ (MxNE) is controlled by monitoring the duality gap (see Section 2.5.3). For more details about convex duality of optimization with sparsity-inducing penalties, refer to [Bac+12].
As mentioned in Section 2.5.3, the minimum of the primal objective function $f_p(\mathbf{X})$ is bounded below by the maximum of the dual objective function $f_d(\mathbf{Y})$, *i.e.*
$f_d(\mathbf{Y}^\star) \leq f_p(\mathbf{X}^\star)$ where $\mathbf{X}^\star$ and $\mathbf{Y}^\star$ are the optimal solutions of the primal and the dual objective functions respectively. If strong duality holds, the duality gap defined as $\eta = f_p(\mathbf{X}) - f_d(\mathbf{Y})$ would be zero at the optimum.

Algorithm 4: $\ell_{2,p}$ MM ALGORITHM WITH $p = 0.5$ (ADAPTIVE LASSO) - ITERATIVE REWEIGHTED MXNE

input : $\mathbf{M}, \mathbf{G}, \lambda > 0, \mathbf{W}^{(0)} \succcurlyeq 0, \epsilon > 0, \tau > 0$ and K
for $k = 1$ *to* K **do**
  $\tilde{\mathbf{G}}^{(k)} = \mathbf{G}\mathbf{W}^{(k-1)}$
  Get $\tilde{\mathbf{X}}^{(k)}$ by solving Equation (4.17) at ϵ-precision as done in Algorithm 5.
  Update $\hat{\mathbf{X}}^{(k)} = \mathbf{W}^{(k-1)}\tilde{\mathbf{X}}^{(k)}$
  Update $\mathbf{W}^{(k)} = \text{diag}(\mathbf{w}^{(k)} \otimes \mathbf{1}_O)$ where $\mathbf{w}^{(k)}[i] = 2\sqrt{\|\hat{\mathbf{X}}^{(k)}[i]\|_{\text{Fro}}}$
  $\forall i \in [1, \cdots, S]$
  if $\|\hat{\mathbf{X}}^{(k)} - \hat{\mathbf{X}}^{(k-1)}\|_\infty \leq \tau$ **then**
    Break

Due to Slater's conditions [BV04], the strong duality holds for the MxNE subproblem and the gap can be used to check the convergence of Equation (4.15). Based on Fenchel-Rockafellar duality theorem [Roc97], one dual objective function associated with the primal objective function:

$$\begin{aligned} f_p(\mathbf{X}) &= \frac{1}{2}\|\mathbf{M} - \mathbf{G}\mathbf{X}\|_{Fro}^2 + \lambda\mathcal{P}(\mathbf{X}) \\ &= \frac{1}{2}\|\mathbf{M} - \mathbf{G}\mathbf{X}\|_{Fro}^2 + \lambda\sum_{i=1}^{S}\|\mathbf{X}[i]\|_{Fro} \end{aligned} \tag{4.18}$$

is given by:

$$f_d(\mathbf{Y}) = -\frac{1}{2}\|\mathbf{Y}\|_{Fro}^2 + Tr(\mathbf{Y}^\top\mathbf{M}) - \lambda\mathcal{P}^\star(\mathbf{G}^\top\mathbf{Y}/\lambda) \tag{4.19}$$

where $\mathcal{P}^\star$ is the Fenchel conjugate of $\mathcal{P}$, which is the indicator function of the associated dual form. For a full derivation, see [GKH12a]. Moreover, the Karush-Khun-Tucker (KKT) conditions of the Fenchel-Rockafellar duality theorem give a natural mapping from the primal to the dual space, which is given by a scaling of the residual $\tilde{\mathbf{Y}} = \mathbf{M} - \mathbf{G}\mathbf{X}$, as shown in [GKH12a].
In practice, we terminate the optimization scheme for solving MxNE when the estimate at the k^{th} iteration is ϵ-optimal with $\epsilon = 10^{-6}$ [Str+16].

For solving the weighted MxNE subproblems, the BCD scheme was used [Tse10], which for the problem at hand converges faster than FISTA (See Section 2.5.3). The subproblem per block has a closed-form solution, which involves applying the group soft-thresholding operator, the proximity operator associated to the $\ell_{2,1}$-mixed-norm [GKH12b; Str+16].

After convergence, we re-apply the scaling to $\tilde{\mathbf{X}}$ to obtain $\hat{\mathbf{X}}$:

$$\mathbf{X}^{\star(k)} = \mathbf{W}^{(k-1)}\tilde{\mathbf{X}}^{(k)} \quad . \tag{4.20}$$

The reformulation through Equation (4.17) and Equation (4.20) avoids any division by zero when $\mathbf{X}^{(k-1)} = 0$. The above procedure, which matches the strategy of the

Algorithm 5: MxNE with BCD and active set strategy

input : $\mathbf{M}, \mathbf{G}, \lambda > 0, \epsilon > 0$, and S
init : $\mathbf{X} = 0, \Gamma = \{\}, \eta = f_p(\mathbf{X}) - f_d(\mathbf{Y})$
for $i = 1$ *to* S **do**
  $\mu[i] = \|\mathbf{G_i}^\top \mathbf{G_i}\|^{-1}$
while $\eta \geq \epsilon$ **do**
  $\Gamma^\star = \{i \mid \|\mathbf{G}_i^\top(\mathbf{M} - \mathbf{G}\mathbf{X})\|_{Fro} > \lambda\}$
  $\Gamma = \Gamma \cup \Gamma^\star$
  Define $\mathbf{G}^\Gamma$ and $\mathbf{X}^\Gamma$ by restricting $\mathbf{G}$ and $\mathbf{X}$ to Γ
  $\mathbf{X}^\star_\Gamma \leftarrow$ Output of Algorithm 6 with $\mu, \mathbf{G}^\Gamma$, and $\mathbf{X}_0 = \mathbf{X}_\Gamma$
  $\mathbf{X}[\Gamma, :] = \mathbf{X}^\star_\Gamma$
  $\eta = f_p(\mathbf{X}) - f_d(\mathbf{Y})$

Algorithm 6: MxNE with BCD

input : $\mathbf{M}, \mathbf{G}, \mathbf{X}, \mu, \lambda > 0, \epsilon > 0$, and S
init : $\eta = f_p(\mathbf{X}) - f_d(\mathbf{Y})$
while $\eta \geq \epsilon$ **do**
  for $i = 1 \in S$ **do**
    $\mathbf{X}[i] \leftarrow$ Solve Equation (2.5.3) with $\mathbf{X}, \mu$, and $\mathbf{M}$
  $\eta = f_p(\mathbf{X}) - f_d(\mathbf{Y})$

Adaptive Lasso estimator [Zou06], is expressed as pseudo-code in Algorithm 4.

4.4.2 Hierarchical Bayesian Modeling

In this section, we formulate the inference problem defined by Equation (2.11) (page 27) and the regularization strategy with $\ell_{2,p}$-quasinorm from a Bayesian perspective [KS05; Luc14]: the Bayesian approach incorporates prior beliefs about the model parameters in terms of probability distributions. Under the Additive, White Gaussian Noise (AWGN) assumption, the likelihood of the model is given by:

$$p_{like}(\mathbf{M}|\mathbf{X}) \propto \exp\left(-\tfrac{1}{2}\|\mathbf{M} - \mathbf{G}\mathbf{X}\|^2_{\text{Fro}}\right) . \tag{4.21}$$

From Equation (4.14) we can construct the $\ell_{2,p}$ group prior as:

$$p_{prior}(\mathbf{X}) \propto \exp\left(-\lambda \sum_{i=1}^{S} \|\mathbf{X}[i]\|^p_{\text{Fro}}\right) = \prod_{i=1}^{S} \exp\left(-\lambda \|\mathbf{X}[i]\|^p_{\text{Fro}}\right) , \tag{4.22}$$

which leads to the following posterior probability density using the Bayes rule:

$$p_{post}(\mathbf{X}|\mathbf{M}) \propto \exp\left(-\frac{1}{2}\|\mathbf{M} - \mathbf{G}\mathbf{X}\|^2_{\text{Fro}} - \lambda \sum_{i=1}^{S} \|\mathbf{X}[i]\|^p_{\text{Fro}}\right) . \tag{4.23}$$

To extend Equation (4.22) to a hierarchical prior model [Mac03], the scalar λ has been replaced by a vector of hyperparameters $\gamma \in \mathbb{R}^S_+$ and for any $p \geq 1$ we

write the *conditional* $\ell_{2,p}$ *prior* as:

$$p_{prior}(\mathbf{X}|\gamma) = \exp\left(-\sum_{i=1}^{S}\left(\frac{\|\mathbf{X}[i]\|_{\text{Fro}}^{p}}{\gamma[i]} + \frac{OT}{p}\log(\gamma[i])\right)\right) \quad , \tag{4.24}$$

where the logarithmic term accounts for the terms of the normalization that depend on γ [Luc14]. A common choice for a hyper-prior on each $\gamma[i]$ is given by a *Gamma distribution* [Mac03; KS05; Cal+09; Luc+12a] with shape and scale parameters α and β:

$$p_{hyper}(\gamma) \propto \prod_{i=1}^{S}\gamma[i]^{\alpha-1}\exp\left(-\frac{\gamma[i]}{\beta}\right) \tag{4.25}$$

$$= \exp\left(-\sum_{i=1}^{S}\left(-\frac{\gamma[i]}{\beta} + (\alpha-1)\log(\gamma[i])\right)\right) \quad . \tag{4.26}$$

Then, the full posterior over both $\mathbf{X}$ and γ becomes:

$$p_{post}(\mathbf{X},\gamma|\mathbf{M}) \propto$$
$$\exp\left(-\frac{1}{2}\|\mathbf{M}-\mathbf{G}\mathbf{X}\|_{\text{Fro}}^{2} - \sum_{i=1}^{S}\left(\frac{\|\mathbf{X}[i]\|_{\text{Fro}}^{p}}{\gamma[i]} + \frac{\gamma[i]}{\beta} - (\alpha-1-\frac{OT}{p})\log(\gamma[i])\right)\right) \quad . \tag{4.27}$$

The question of how to best derive parameter estimates, in particular how to treat the two different types of parameters $\mathbf{X}$ and γ, distinguishes different HBM-based inference strategies. Variational Bayesian approaches [Mac03; Jor+99; Sat+04; Fri+08; SB15] and Sparse Bayesian Learning [Tip01; WR04; WN09a; ZR11b] approaches rely on approximating or marginalizing the full, joint posterior distribution (Equation (4.27)). In contrast, fully-Bayesian strategies [Cal+09; Luc+12a] work with it directly. The most popular one is the full maximum-a-posteriori (*full-MAP*) estimate which is defined as:

$$(\hat{\mathbf{X}}_{\text{MAP}}, \hat{\gamma}_{\text{MAP}}) \quad \in \arg\max_{(\mathbf{X},\gamma)\in\mathbb{R}^{SO\times T}\times\mathbb{R}_{+}^{n}}\{p_{post}(\mathbf{X},\gamma|\mathbf{M})\} \quad . \tag{4.28}$$

A common strategy to compute it is to minimize the *negative log posterior* $-\log p_{post}(\mathbf{X},\gamma|\mathbf{M})$ by alternating minimization over $\mathbf{X}$ and γ (*block coordinate descent* in optimization):

$$\mathbf{X}^{(k)} \in \underset{\mathbf{X}\in\mathbb{R}^{SO\times T}}{\arg\min}\left\{\frac{1}{2}\|\mathbf{M}-\mathbf{G}\mathbf{X}\|_{\text{Fro}}^{2} + \sum_{i=1}^{S}\frac{\|\mathbf{X}[i]\|_{\text{Fro}}^{p}}{\gamma^{(k-1)}[i]}\right\} \quad , \tag{4.29}$$

$$\gamma^{(k)}[i] \in \underset{\gamma[i]\in\mathbb{R}_+}{\arg\min} \left\{ \frac{\|\mathbf{X}^{(k)}[i]\|_{\text{Fro}}^p}{\gamma[i]} + \frac{\gamma[i]}{\beta} - (\alpha - 1 - \frac{OT}{p}) \log(\gamma[i]) \right\}, \quad \forall i \in [1, \cdots, S] \ . \tag{4.30}$$

Other fully-Bayesian estimates are defined as integrals of functions of $\mathbf{X}$ and γ with respect to the posterior distribution, *e.g.* first or second moment estimates. To compute these high dimensional integrals efficiently, only MCMC methods that draw correlated samples from the posterior distribution can be used [RC05; KS05]. A commonly used MCMC scheme for HBM is given by *blocked Gibbs sampling*, which alternates as:

$$\mathbf{X}^{(k)} \sim p_{post}(\mathbf{X}, \gamma^{(k-1)}|\mathbf{M}) \propto p_{post}(\mathbf{X}|\mathbf{M}, \gamma^{(k-1)}) \ , \tag{4.31}$$

$$\gamma^{(k)} \sim p_{post}(\mathbf{X}^{(k)}, \gamma|\mathbf{M}) \propto p_{post}(\gamma|\mathbf{M}, \mathbf{X}^{(k)}) \ . \tag{4.32}$$

Depending on the purpose of the study, here the main interest is not sampling the posterior distribution for computing the integral-based estimators, but we rather want to explore the different modes of this multi-modal distribution, each of which corresponds to parameters that are both sparse and likely to explain the data.
One can notice similar structures in Equations (4.29)-(4.30) and Equations (4.31)-(4.32): in each step, we make use of the conditional structure of the posterior: for γ fixed, we have to solve one SOT-dimensional $\ell_{2,p}$ optimization/sampling problems, while for $\mathbf{X}$ fixed, we have to solve S 1-dimensional optimization/sampling problems. These two steps will be described in more detail in the sections 4.5 and 4.6.

4.5 HBM optimization in the Bayesian formulation

The optimization problem defined in Equation (4.29) reduces to an $\ell_{2,p}$-norm regularized regression problem that can be solved as described in Section 4.4.1. For solving Equation (4.30), we compute the first order optimality condition for each i:

$$-\frac{\|\mathbf{X}^{(k)}[i]\|_{\text{Fro}}^p}{\gamma[i]^2} + \frac{1}{\beta} - \frac{(\alpha - 1 - \frac{OT}{p})}{\gamma[i]} = 0 \ , \tag{4.33}$$

For $\alpha \geqslant OT/p + 1$, the problem in Equation (4.30) is convex, and the positive root of Equation (4.33) is given by:

$$\gamma[i] = \beta \left(\nu + \sqrt{\nu^2 + \frac{\|\mathbf{X}^{(k)}[i]\|_{\text{Fro}}^p}{\beta}} \right), \qquad \nu := \frac{\alpha - 1 - OT/p}{2} \ . \tag{4.34}$$

Note that similar rules to update the noise level were considered in the Bayesian Lasso [PC08; Kyu+10] and the Scaled Lasso (see for instance [SBv10; Dal12]). A difference though is that the update we perform here is on the penalty term, whereas in the mentioned references, it was rather performed on the data-fitting term.

If we furthermore choose $\alpha = OT/p+1$, then $\nu = 0$ and most terms disappear; Equation (4.29) and (4.30) hence read:

$$\mathbf{X}^{(k)} = \arg\min_{\mathbf{X}\in\mathbb{R}^{SO\times T}} \left\{ \tfrac{1}{2}\|\mathbf{M}-\mathbf{G}\mathbf{X}\|_{\text{Fro}}^2 + \sum_{i=1}^{S} \frac{\|\mathbf{X}[i]\|_{\text{Fro}}^p}{\gamma^{(k-1)}[i]} \right\} , \tag{4.35}$$

$$\gamma^{(k)}[i] = \sqrt{\beta}\sqrt{\|\mathbf{X}^{(k)}[i]\|_{\text{Fro}}^p}, \quad \forall i = 1,\ldots,S , \tag{4.36}$$

which can be combined into the fixed point iteration:

$$\mathbf{X}^{(k)} = \underset{\mathbf{X}\in\mathbb{R}^{SO\times T}}{\arg\min} \left\{ \frac{1}{2}\|\mathbf{M}-\mathbf{G}\mathbf{X}\|_{\text{Fro}}^2 + \frac{2}{\sqrt{\beta}} \sum_{i=1}^{S} \frac{\|\mathbf{X}[i]\|_{\text{Fro}}^p}{2\sqrt{\|\mathbf{X}^{(k-1)}[i]\|_{\text{Fro}}^p}} \right\} . \tag{4.37}$$

If we compare Equation (4.37) with Equation (4.15), we see that we re-derived the MM algorithm for $p = 1$ as an alternating optimization scheme to compute the *full-MAP* estimate for a specific HBM, namely using a conditional $\ell_{2,1}$ group prior and a Gamma hyper-prior with $\alpha = OT + 1$ and $\beta = 4/\lambda^2$. Using $\mathbf{w}^{(0)}[i] := 1$ in the MM scheme corresponds to starting with $\gamma[i]^{(0)} := 1/\lambda = 2/\sqrt{\beta}$.

From previous work [Str+16] we know that due to the non-convexity, a good initialization of the weights $\mathbf{w}^{(0)}[i]$ in the MM algorithm is crucial for its performance, but aside uniform initialization, only heuristic initialization strategies were used, *e.g.* using the same re-weighting as in the sLORETA method [PM02]. In this thesis, we leverage the re-interpretation of the MM algorithm through the HBM framework to obtain multiple initializations in a systematic fashion, namely as samples drawn from the full posterior. In this way, we can not only reach better local minima, but more importantly, we can identify and characterize multiple possible sparse solutions. Such plausible solutions to the sparse regression problem in Equation (2.11) are the modes of the posterior distribution (Equation (4.27)) with different relative probability masses.

4.6 Posterior Sampling

As outlined in Equations (4.31) and (4.32) in Section 4.4.2, we sample the full posterior $p_{post}(\mathbf{X}, \gamma|\mathbf{M})$ by blocked Gibbs sampling, *i.e.*, we alternate between sampling the conditional distributions $p_{post}(\mathbf{X}|\mathbf{M}, \gamma^{(k-1)})$ and $p_{post}(\gamma|\mathbf{M}, \mathbf{X}^{(k)})$. The conditional $p_{post}(\mathbf{X}|\mathbf{M}, \gamma^{(k-1)})$ is a high dimensional distribution composed of a Gaussian likelihood and an $\ell_{2,p}$ prior, where our main interest here is $p = 1$. It was

Algorithm 7: BLOCK GIBBS SAMPLING SCHEME

input : $\mathbf{M}, \mathbf{G}, \mathbf{X}^{(-K_0)}, \gamma^{(-K_0)}, K_0, K, K_{SC}, K_{SS}, \alpha, \beta$
for $k = -K_0 + 1$ ***to*** K **do**
  Set $\mathbf{X}^{(k)} = \mathbf{X}^{(k-1)}$.
  for $k_{SC} = 1$ ***to*** K_{SC} **do**
    Draw a random permutation P of $\{1, \ldots, S\}$
    for $l \in P$ **do**
      Sample $\mathbf{X}^{(k)}[i,j] \sim p_{post}(\mathbf{X}[i,j]|\mathbf{X}^{(k-1)}[-(i,j)], \mathbf{M}, \gamma^{(k)}), \forall (i,j) \in [l]$ - via K_{SS} steps of **Slice Sampling** Algorithm 8.
  Sample $\gamma_i^{(k)} \sim p_{post}(\gamma_i|\mathbf{M}, \mathbf{X}^{(k)}), \forall i = 1, \ldots, n$ via **Accept-Reject** Algorithm 9.
return $\{\mathbf{X}^{(k)}, \gamma^{(k)}\}_{k=1}^{K}$

demonstrated in [Luc12] that *single component Gibbs sampling* (*SC Gibbs*) is an efficient MCMC technique to sample such distributions. For the specific $\ell_{2,p}$ priors used here, *slice sampling* can be used to perform the sub-steps in SC Gibbs sampling, namely the sampling of the one-dimensional single-component conditional densities. The resulting *Slice-Within-Gibbs* sampler was examined in [Luc16]. For completeness, the details of the implementation are given in Section 4.6.1.

Following Equation (4.27), the conditional $p_{post}(\gamma|\mathbf{M}, \mathbf{X}^{(k)})$ factorizes over groups i:

$$p_{post}(\gamma[i]|\mathbf{M}, \mathbf{X}^{(k)}) \propto \exp\left(-\frac{\|\mathbf{X}^{(k)}[i]\|_{\text{Fro}}^p}{\gamma[i]} - \frac{\gamma[i]}{\beta} + (\alpha - 1 - OT/p)\log(\gamma[i])\right) \quad . \tag{4.38}$$

For the case of $\alpha = OT/p + 1$, which is our main interest due to its connection to MM revealed in the previous section, Equation (4.38) reduces to:

$$p_{post}(\gamma[i]|\mathbf{M}, \mathbf{X}^{(k)}) \propto \exp\left(-\frac{\|\mathbf{X}^{(k)}[i]\|_{\text{Fro}}^p}{\gamma[i]}\right) \exp\left(-\frac{\gamma[i]}{\beta}\right) \quad , \tag{4.39}$$

which can be sampled with a simple accept-reject algorithm as described in Section 4.6.1. The complete procedure is described in Algorithm 7. Therein, K_0 refers to the burn-in size, *i.e.* the initial samples that are discarded, K to the sample size of the blocked Gibbs sampler and K_{SC}, K_{SS} to the sample sizes of the SC Gibbs and the slice sampler that carries out the sampling in the sub-steps.

4.6.1 Slice-Within-Gibbs Sampler for Parameter Update

Within the Algorithm 7, to update a group $\mathbf{X}[l]$, we need to sample from all the one-dimensional SC densities:

$$p_{post}(\mathbf{X}[i,j]|\mathbf{X}^{(k-1)}[-(i,j)], \mathbf{M}, \gamma^{(k)}), \qquad (i,j) \in [l] \quad , \tag{4.40}$$

where $\mathbf{X}[-(i,j)]$ refers to all the coefficients of matrix $\mathbf{X}$ except the term (i,j).

In order to implement this efficiently, we can precompute several terms and make use of the specific spatio-temporal group structure of the posterior. We first derive the part of the likelihood as in Equation (4.21) that depends on a given index pair $(i, j) \in [l]$:

$$\frac{1}{2}\|\mathbf{M} - \mathbf{G}\mathbf{X}\|_{\text{Fro}}^2 = \sum_{j'}^{T} \frac{1}{2}\|\mathbf{M}[:, j'] - \mathbf{G}\mathbf{X}[:, j']\|_2^2$$

$$\overset{j}{\propto}$$

$$\frac{1}{2}\|\mathbf{M}[:, j] - \mathbf{G}\mathbf{X}[:, j]\|_2^2 = \frac{1}{2}\|\mathbf{M}[:, j] - (\mathbf{G}[:, -i]\mathbf{X}[-i, j] + \mathbf{G}[:, i]\mathbf{X}[i, j])\|_2^2$$

$$\overset{i}{\propto}$$

$$\begin{aligned}
&\frac{1}{2}\|\mathbf{G}[:, i]\|_2^2\, \mathbf{X}[i, j]^2 + \mathbf{G}[:, i]^\top \left(\mathbf{M}[:, j] - \mathbf{G}[:, -i]\mathbf{X}[-i, j]\right) \mathbf{X}[i, j] \\
= &\frac{1}{2}\|\mathbf{G}[:, i]\|_2^2\, \mathbf{X}[i, j]^2 \\
&\qquad + \left(\left(\mathbf{G}^\top \mathbf{M}\right)[i, j] - \left(\mathbf{G}[:, i]^\top \mathbf{G}\right) \mathbf{X}[:, j] - \|\mathbf{G}[:, i]\|_2^2\right) \mathbf{X}[i, j] \\
&:= az^2 + bz \quad , \quad \text{with} \quad z := \mathbf{X}[i, j] \quad , \quad a := \frac{1}{2}\|\mathbf{G}[:, i]\|_2^2 \quad , \\
&b := \left(\mathbf{G}^\top \mathbf{M}\right)[i, j] - \left(\mathbf{G}[:, i]^\top \mathbf{G}\right) \mathbf{X}[:, j] - \|\mathbf{G}[:, i]\|_2^2
\end{aligned}$$

where $\overset{j}{\propto}$ means propotional for each j, and $:=$ assign the equation to a specific reformulation.

Note that $\|\mathbf{G}[:, i]\|_2^2$ and $\mathbf{G}^\top\mathbf{M}$ can be precomputed. The challenging part in the computation of b is to compute $\mathbf{G}[:, i]^\top\mathbf{G}$, as one typically does not want to precompute the $SO \times SO$ matrix $\mathbf{G}^\top\mathbf{G}$ and hold it in memory. However, to update all the TO components of the l^{th} group (*e.g.* in the visual evoked fields example, $T = 211$, $O = 3$) one only needs the $O \times SO$ matrix $\mathbf{G}[:, [l]]^\top\mathbf{G}$. Thus, we compute $\mathbf{G}[:, [l]]^\top\mathbf{G}$ at the start of updating group $\mathbf{X}[l]$ and hold it in memory throughout the bloc update. Besides this, the most costly operation to compute b is a dot product of vectors of size SO. Next, we derive the part of the prior (4.24) that depends on

Algorithm 8: SLICE SAMPLING

input : $p(z) \propto p_1(z)p_2(z), z, K_{(SS)}$
for $k = 1$ *to* K_{SS} **do**
  Draw y uniformly from $[0, p_2(z)]$ (vertical move).
  Determine $S_2^y := \{z \mid p_2(z) \geqslant y\}$
  Draw z from $p_1(z)\mathbb{1}_{S_2^y}(z)$ (weighted horizontal move).
return z as a sample of $p(z)$

$\mathbf{X}[i,j], (i,j) \in [l]$:

$$\sum_{l=1}^{S} \left(\frac{\|\mathbf{X}[l]\|_{\text{Fro}}^p}{\gamma l} + \frac{OT}{p} \log(\gamma[l]) \right)$$

$$\overset{\mathbf{X}[i,j],(i,j)\in[l]}{\propto} \quad \gamma[l]^{-1}\|\mathbf{X}[l]\|_{\text{Fro}}^p = \gamma[l]^{-1} \left(\sum_{(i',j')\in[l]} \mathbf{X}[i',j']^2 \right)^{p/2}$$

$$= \gamma[l]^{-1} \left(\mathbf{X}[i,j]^2 + \sum_{\substack{(i',j')\in[l] \\ (i',j')\neq(i,j)}} \mathbf{X}[i',j']^2 \right)^{p/2}$$

$$:= c\left(z^2 + d\right)^{p/2} \quad ,$$

with c and d defined and computed in an obvious way. Taken together, to update $\mathbf{X}[i,j]$, we have to sample from the one-dimensional density:

$$p(z) \propto \exp\left(-az^2 - bz\right) \exp\left(-c\left(z^2 + d\right)^{p/2}\right) =: p_1(z)p_2(z) \quad . \tag{4.41}$$

We take advantage of the fact that (4.41) factorizes in a Gaussian likelihood part $p_1(z)$ and a symmetric, log-concave prior part $p_2(z)$, and use a generalized form of slice sampling [Nea03; RC05] as described in more detail in [Luc16] and summarized in Algorithm 8. Determining S_2^y in our case is trivial:

$$p_2(z) \geqslant y \quad \Leftrightarrow \quad c\left(z^2 + d\right)^{p/2} \leqslant -\log(y) \tag{4.42}$$

$$\Leftrightarrow |x| \leqslant \left(\left(\frac{-\log(y)}{c} \right)^{2/p} - d \right)^{1/2} \tag{4.43}$$

Then, we use a slightly modified, more robust version of the fast table-based algorithm described in [Cho11] to sample from the truncated Gaussian distribution $p_1(z)\mathbb{1}_{S_2^y}(z)$. As initialization for z, we always choose the current value of $\mathbf{X}[i,j]$.

4.6.2 Accept-Reject Sampler for Hyperparameter Update

The conditional density (4.39) is of the type

$$p(x) \propto \exp\left(-\frac{c}{x}\right) \exp\left(-\frac{x}{\beta}\right) \quad , \qquad c, \beta \geqslant 0. \tag{4.44}$$

Note that the first factor is monotonically increasing with limit 0 for $x \searrow 0$ and limit 1 for $x \to \infty$, while the second factor is proportional to a simple exponential distribution (*cf.* Figure 4.6.1). We can therefore easily construct a dominating density $g(x) \geqslant p(x)$ to carry out accept-reject sampling [RC05, Section 2.3.2] to generate a sample $z \sim p$: we generate $y \sim g$, $u \sim \mathcal{U}_{[0,1]}$ and accept $z = y$ if $u \leqslant p(y)/g(y)$ and repeat otherwise. Choosing $g(x) = \exp(-x/\beta)$ would yield a valid sampling density but this choice becomes inefficient with increasing c. Therefore, we split the sampling density into two parts:

$$g(x) = \begin{cases} \hat{p} & \text{if} \quad x < \tilde{x} \\ \exp\left(-\frac{x}{\beta}\right) & \text{otherwise} \quad , \end{cases} \tag{4.45}$$

where $\hat{p} = \max_x p(x)$ is the maximal probability attained at $\hat{x} = \arg\max_x p(x) = \sqrt{\beta c}$ and $\tilde{x} = \beta c/\hat{x} + \hat{x}$ is the solution to $\exp(-x/\beta) = \hat{p}$ (*cf.* Figure 4.6.1). Sampling from (4.45) is then straightforward using $v, w \sim \mathcal{U}_{[0,1]}$: if one computes

$$G_{x \geqslant \tilde{x}} = \int_{\tilde{x}}^{\infty} g(x)\, dx = \beta \exp(-\tilde{x}/\beta), \qquad G_{x < \tilde{x}} = \int_0^{\tilde{x}} g(x)\, dx = \hat{p}\tilde{x} \quad , \tag{4.46}$$

then $v < G_{x \geqslant \tilde{x}}/(G_{x \geqslant \tilde{x}} + G_{x < \tilde{x}})$ determines that we are in the tail, $x > \tilde{x}$, where we can use a simple inverse cumulative distribution method to draw a proposal from $g(x)$ using w. If v determines $x \leqslant \tilde{x}$, then $x = w\tilde{x}$ is the proposal. For numerical precision, we only compute logarithms of probabilities and use that for $a > 0$, b$\geqslant 0$:

$$\log(a + b) = \log a + \log(1 + \exp(b - a)) \quad . \tag{4.47}$$

The whole sampling scheme is shown in Algorithm 9. We found the scheme to be efficient enough for all of our computations, *i.e.* the chosen $g(x)$ is close enough to $p(x)$ to result in an accepted sample after a few trials. If this had to become a problem, it would be easy to adaptively improve the dominating density.

Algorithm 9: ACCEPT-REJECT ALGORITHM FOR HYPERPARAMETER UPDATE

input : $c \geqslant 0, \beta > 0$
Set $\hat{x} = \sqrt{\beta c}$.
Set $\log \hat{p} = -c/\hat{x} - \hat{x}/\beta$.
Set $\tilde{x} = \beta c/\hat{x} + \hat{x}$.
Set $\log G_{x \geqslant \tilde{x}} = \log \beta - \tilde{x}/\beta$.
Set $\log G_{x < \tilde{x}} = \log \hat{p} + \log \tilde{x}$.
Set $\log G_{tot} = \log G_{x \geqslant \tilde{x}} + \log\left(1 + \exp\left(G_{x < \tilde{x}} - G_{x \geqslant \tilde{x}}\right)\right)$.
while *true* **do**
  Draw u, v, w uniformly from $(0, 1)$.
  if $\log v + \log G_{tot} < \log G_{x \geqslant \tilde{x}}$ **then**
    Set $W = \log w - \tilde{x}/\beta$.
    Propose $x = -\beta W$:
    if $\beta \log(u) < c/W$ **then**
      return x (acceptance)
  else
    Propose $x = w\tilde{x}$:
    if $\log u + \log \hat{p} < -c/x - x/\beta$ **then**
      return x (acceptance)

4.7 Experiments

4.7.1 Study of the different modes defining uncertainty maps of the MEG/EEG inverse problem

We now examine the benefits of our re-interpretation of the MM algorithm described in Section 4.4.1 as a specific way to compute a full-MAP estimate for a specific HBM as described in Sections 4.4.2 and 4.4.1. In particular, we investigate how using MCMC sampling of the posterior distribution as described in Section 4.6 can help getting better initializations for the optimization algorithm. We first present results for a simulated MEG dataset and then for two experimental MEG/EEG datasets.

Simulation study

We generated a realistic simulation based on a free-orientation ($d = 3$) source model with $n = 7498$ cortical locations and $m = 306$ MEG sensors. Two of these locations were selected to be active, one in each hemisphere. One of the sources had a deep ventral location in the inferior occipital gyrus (Figure 4.7.1-c), and the second one had a more superficial location in the motor cortex (Figure 4.7.1-a). Their corresponding waveforms are shown in Figure 4.7.1-b. When passed to the solvers, they are cropped between 40 to 180 ms to keep only the two peaks. This leads to $t = 43$ time samples.

The aim of the simulation is to answer two separate questions. First, we want to know whether we are able to find better source estimates using MCMC-derived

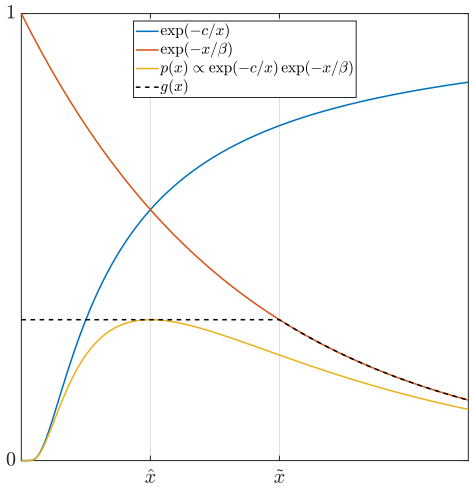


FIGURE 4.6.1: Sketch of the quantities used in the accept-reject sampling Algorithm 9.

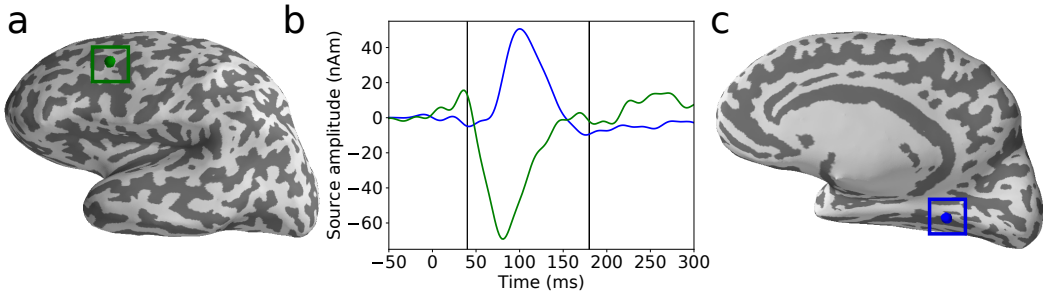


FIGURE 4.7.1: Simulated MEG dataset. a) and c) show superficial and deep sources (hidden in the medial view) locations, respectively. b) gives their corresponding waveforms color-coded by location.

initializations than with the uniformly initialized MM Algorithm 4. For this, we first run the MM Algorithm 4 using a uniform initialization, *i.e.*, $\mathbf{w}_i^{(0)} = 1, \forall i = 1 \in [1, \cdots, S]$, with $\lambda = 0.05\lambda_{max}$ where $\lambda_{max} = \max_{1 \leq i \leq S} \|(G^\top M)[i]\|_{\text{Fro}}^2$ is the smallest regularization value for which no source is found as active using an $\ell_{2,1}$ regularization [Ndi+15; Str+16]. As described above, this corresponds to computing a full-MAP estimate for the HBM with $p = 1$, $\alpha = OT + 1$, $\beta = 4/\lambda^2$ using the alternation scheme (4.35) initialized with $\gamma^{(0)}[i] = 1/\lambda, \forall i = 1 \in [1, \cdots, S]$.

Then, we sampled the corresponding posterior distribution given in Equation (4.27) using Algorithm 7 with $K_0 = 300$, $K = 900$, $K_{SC} = K_{SS} = 1$. From each $\gamma^{(k)}$ of the $K = 900$ obtained γ samples, we construct an initialization $\mathbf{W}^{(0)}$ for the MM Algorithm 4 by setting $\mathbf{w}^{(0)}[i] = \lambda\gamma^{(k)}[i], \forall i = [1, \ldots, S]$. Figure 4.7.2-c shows the histogram of the objective function values (computed with Equation (4.17)) obtained in this way. The vertical black bar shows the value of the objective function of the uniformly initialized MM solver and we can see that some initializations indeed lead to source estimates with a lower objective value. Figure 4.7.2-a and Figure 4.7.2-b show the locations of the estimated sources resulting from uniform and best MCMC-based initialization. For the artificial source in Fig. 4.7.2-a, both

results find the exact location, so they are superimposed. For the deeper source in Figure 4.7.2-b, neither result finds the exact position, but the MCMC-based initialization is closer. This means that the result did not only improve from an optimization point of view, but also judged by the quality criteria of the given application.

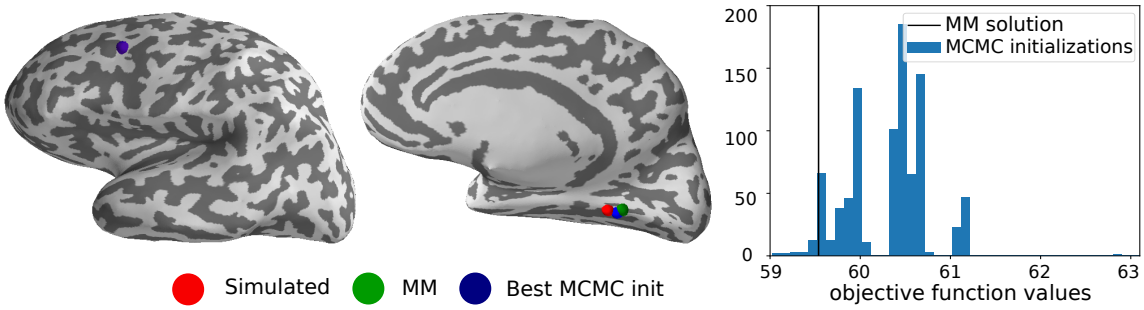


FIGURE 4.7.2: Location of simulated and estimated sources using the uniformly initialized MM solver (denoted as "MM") and best MCMC-based initialization in terms of objective function value. Left: estimation of the artificial source on the left hemisphere. Middle: estimation of the deep source on the right hemisphere. Right: histogram of the objective function value for 900 MCMC initializations. The uniform initialization used for the MM (black vertical line) is not very bad, meaning that the basic MM is able to recover a good source estimate for some configurations. See Figure 4.7.5 for a case where the basic MM fails.

Finding the correct support in a sparse under-determined regression problem like Equation (2.11) is inherently of combinatorial complexity. In our two approaches, this is reflected in the non-convexity of the objective function (Equation (4.14)) and the multi-modality of the joint posterior distribution (Equation (4.27)), respectively. The second question we want to investigate is whether the methods we developed here can reveal or quantify some of the ambiguity and uncertainty of this sparse support identification problem. Traditional uncertainty quantification (UQ) measures such as variance estimates of $\mathbf{X}$ or γ fail to do so as they cannot capture the multi-modality of the posterior in a satisfactory way. In addition, no sample $\mathbf{X}^{(k)}$ is exactly sparse: as the posterior distribution is a continuous probability density, the probability of the event $\{\mathbf{X}^{(k)}[i] = 0\}$ is zero, which means that the whole support of $\mathbf{X}^{(k)}$ is active with probability 1. Even a thresholded average of the support of $\mathbf{X}^{(k)}$ will only reveal the average probability of a location being part of the support. In source analysis, it is arguably more interesting to estimate which networks of sources from different brain areas have most likely produced a given data set, a question left open by these measures. Here, we propose to tackle this question in a different way.

Our procedure of initializing an MM iteration with a sample from the posterior distribution yields different local minima of Equation (4.14), *i.e.*, approximate solutions to Equation (2.11) that fulfill our *a priori* knowledge of a sparse support, but it also yields the relative frequencies with which these minima are found by the

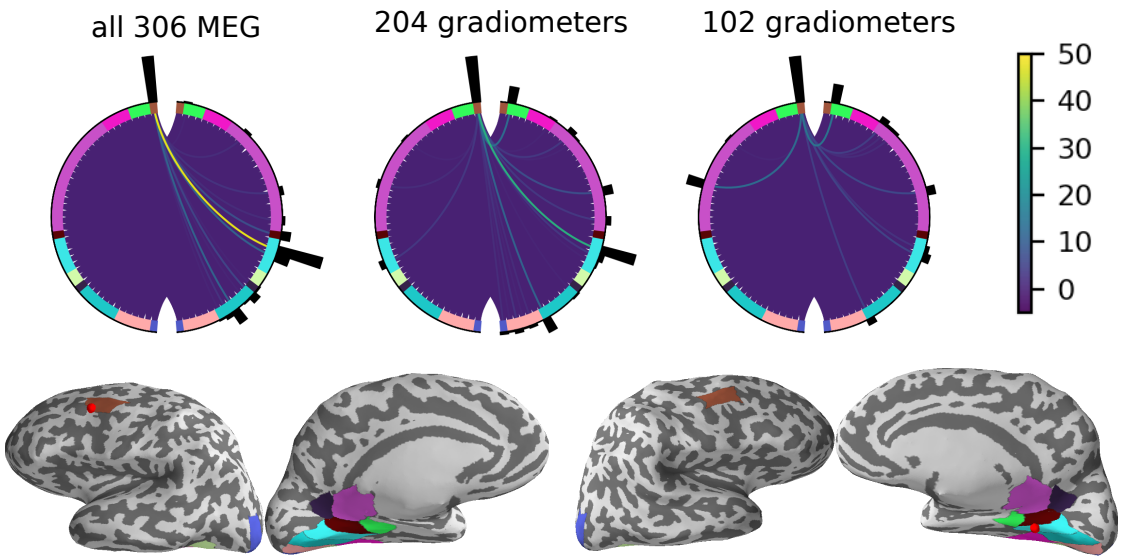


FIGURE 4.7.3: Source network analysis for simulated data: for a clearer presentation, the set of 900 initializations was thinned to the 100 that gave the lowest objective function (Equation (4.14)). The first row of sub-figures displays the support of these best local minima in the following way: each position in the circle represents a source location that was part of the support of at least one minimum for one sensor configuration. The black bar attached to each position corresponds to the relative frequency with which this source location appeared as part of the support. Two positions are connected by a line if they were simultaneously part of the support, and the color of this line corresponds to the relative frequency with which this happened. Note that the background of the circle is white, but densely covered by purple lines indicating rare connections. The positions are placed left or right, depending on which hemisphere they belong to. For symmetry, for each active source location, its counterpart on the other hemisphere was included in the graphic as well. In addition, the positions are grouped and colored based on a parcellation of the brain into anatomical regions (taken from an atlas). The second row of subfigures shows these regions in the brain and the simulated sources.

MM algorithm. If we assume that the division of $\mathbb{R}^{SO\times T}$ into attractors of the MM algorithm roughly overlaps with the division of $\mathbb{R}^{SO\times T}$ into modes of the marginalized posterior over $\mathbf{X}$ within the HBM framework, this relative frequency corresponds to the relative volume of the local minima of Equation (4.14). The latter is a better measure to compare different local minima than their depth (a local minimum that is deep but thin corresponds to an unstable source estimate). While a mathematically more profound and detailed analysis of this heuristic is left for future work, we examine here if this approach reacts to changes in the measurement design in the way we would expect. To do so, we switch from using all 306 MEG sensors to using only 204 gradiometers or one over two gradiometers (102 sensors). By reducing the number of sensors we increase the under-determinedness of the problem and the intuition is that it should lead to more variability among

the plausible sparse solutions. The graphical analysis presented and described in Figure 4.7.3 and Figure 4.7.4 confirms this. A first observation is that the superficial source in the premotor cortex was correctly identified as part of the support of every local minima when using the full 306 MEG sensors. It was however sometimes miss-localized when reducing the number of sensors (Figure 4.7.3). A second observation is that the spatial spread of these miss-localizations is smaller for this superficial source than it is for the deep source. This deep source in the ventral cortex is more difficult to find even with all sensors. Indeed, none of the 100 best initializations perfectly localized the deep simulated source. In general, we can clearly see how the ambiguity increases when decreasing the number of sensors, and how the distribution of source networks gets more fuzzy. However, our analysis also provides useful local measures of these phenomena.

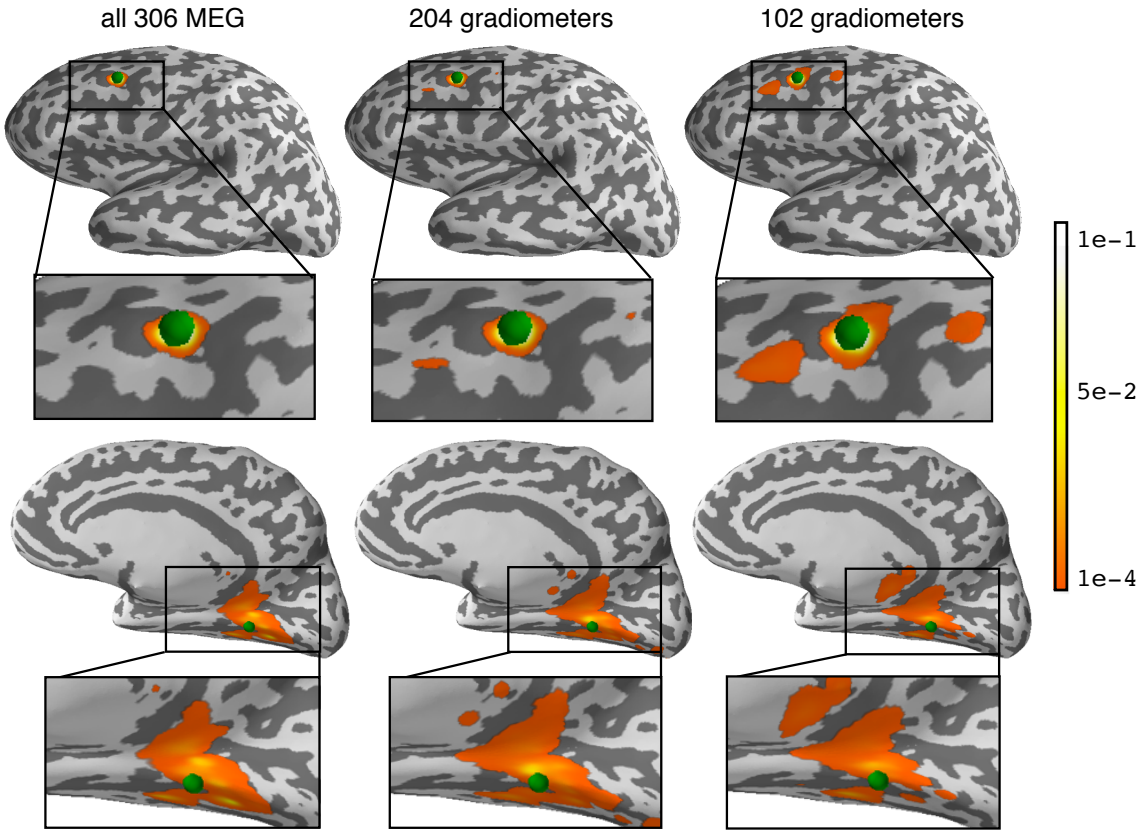


FIGURE 4.7.4: The support of the MM results based upon 900 MCMC-based initializations was extracted to build an uncertainty map. The relative frequencies with which each source location was part of the support was computed and plotted on the brain surface together with the two simulated sources (green dots). Each column corresponds to the results for each of the three sensor setups examined. The less the number of sensors and/or the deeper the source is, more uncertain the brain map is.

Experimental results with MEG auditory and visual data

We now repeat our analysis with two experimental open datasets. The first one is a recording of auditory evoked fields (MNE sample dataset [Gra+13b]). The second one contains visual evoked fields (visual condition of MNE sample dataset) for which source localization is a more difficult task due to the proximity between neural sources. The true nature of the underlying source network is also less clear for this second dataset.

Figure 4.7.5 shows the equivalent of Figure 4.7.2 for both datasets. Again, we see that a lower objective function value can be obtained using MCMC-based initializations. The auditory sample dataset is commonly assumed to be generated by two bilateral focal sources around the auditory cortices in the superior temporal gyrus of the temporal lobe. Due to the superficial nature of these sources and their large distance, the estimation of their position is regarded as a relatively simple task. Indeed, the histogram shows that using MCMC-based initializations does not help a lot to reduce the objective function compared to a uniformly initialized MM solution. In the case of the visual dataset, where several closed-by sources are active, the difference is however quite drastic. The majority of the MCMC-based initializations lead to lower values of the objective function. Looking at the source distribution plots on the brain for both datasets, one can also observe more complex source configurations for the visual data.

Next, we repeat the graphical source network analysis from Figure 4.7.3 for the two datasets. Figure 4.7.6 shows the results for the auditory dataset and three sensor configurations: all 364 EEG + MEG sensors, all 306 MEG sensors or one over two sensors resulting in 182 EEG + MEG sensors. One can see how adding EEG to MEG sensors reduces the ambiguity of the regression problem. The plots show fewer but more prominent modes, *i.e.* the posterior mass is concentrated on fewer stable source configurations. We also see that the locations of the most prominent modes shift. This is consistent with results of other studies on EEG-MEG combination [Mol+08; Luc14; Ayd+14] as EEG is sensitive to some sources that MEG is almost blind to, *e.g.* sources with a strong radial component. If we subsample the EEG+MEG sensors by only using every other location, the ambiguity and spatial spread of the recovered support increases. One can see that there is more activity in the dark green label, which corresponds to a brain area commonly not associated with auditory responses.The connections between source locations show that none of the found modes really stands out, *i.e.* is found much more often compared to the others. Most of the connections do not occur more than 200 times within the 900 samples, so they are part of the purple background of low frequency connections in the plots.

Figure 4.7.7 shows the same results for the more complex visual dataset. Compared to the auditory dataset, we see that even with all sensors, the ambiguity of the regression problem seems to be a lot higher compared to the auditory dataset: we see

that the posterior mass is distributed among many more source configurations. For the other two sensor configurations, we see similar effects as in the auditory data set. Nevertheless, it can be noticed that the large majority of identified sources with all MCMC initializations are on the right hemisphere. This is consistent with the known functional organisation of the visual cortex. Indeed, in this experimental condition the subject was presented with checker board flashes on the left visual hemifield which is known to primarily project onto the right hemisphere of the cortex.

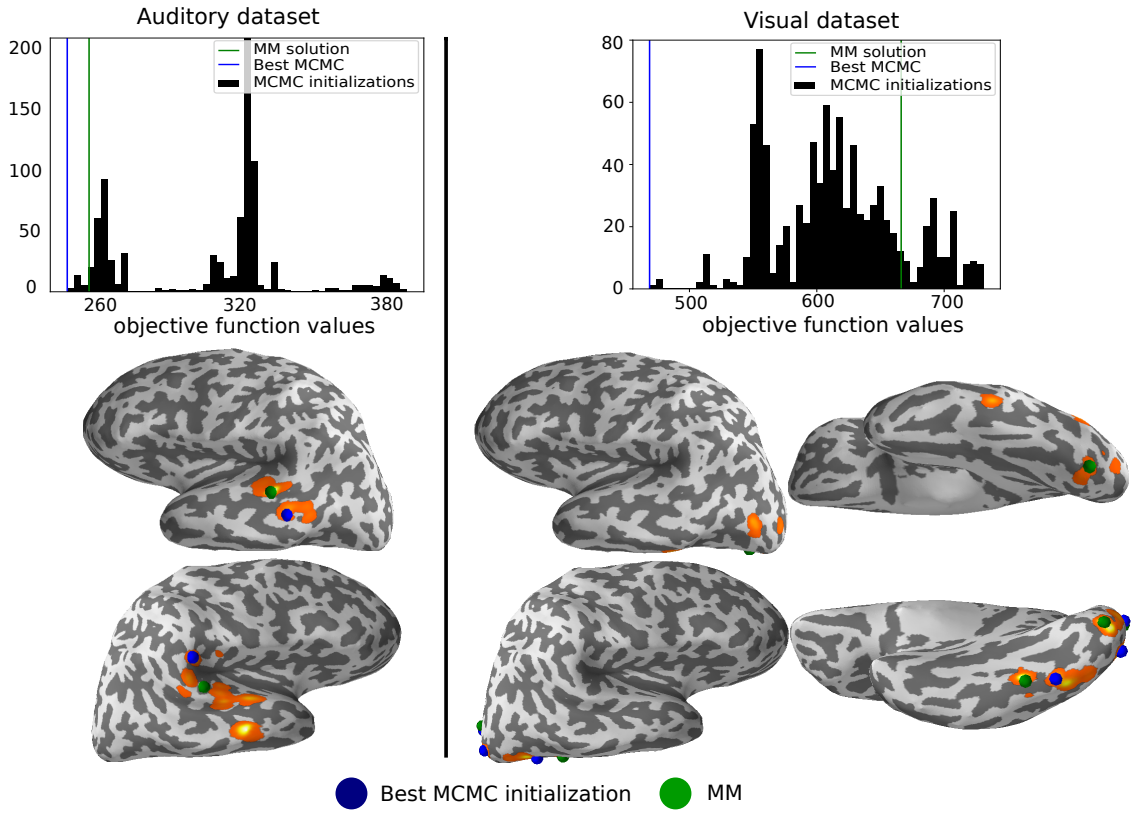


FIGURE 4.7.5: Histograms of the objective function value for 900 MCMC initializations for auditory and visual datasets (306 MEG sensors). The histogram for the visual dataset shows more MCMC initializations that outperform the uniform one in the MM solution. Under each histogram, these source configurations are shown on the left and right hemisphere.

4.8 Conclusion & Perspectives

Scientific literature relying either on frequentist or on Bayesian statistical inference often coexist in many fields ranging from machine learning, inverse problems, signal processing or computational biology. In this work, we started from an under-determined, ill-conditioned MMV / multi-task regression problem and examined two seemingly unrelated approaches - MM as an optimization technique

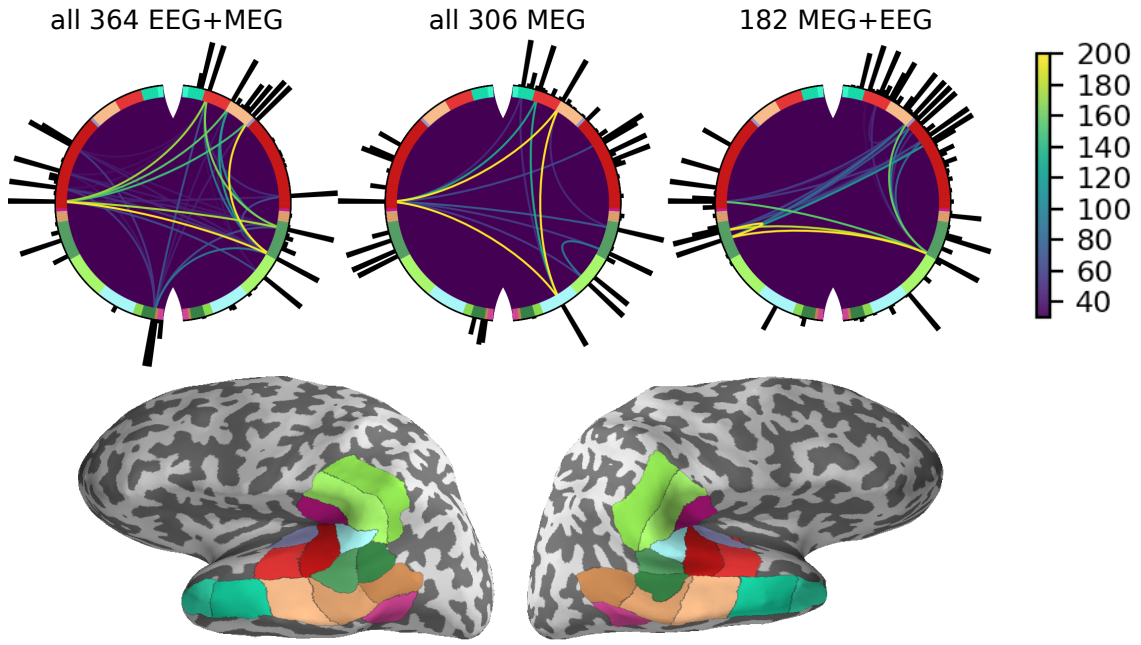


FIGURE 4.7.6: Source network analysis for auditory data. The figures are constructed in the same way as described in Figure 4.7.3 except that all 900 MCMC initializations are displayed.

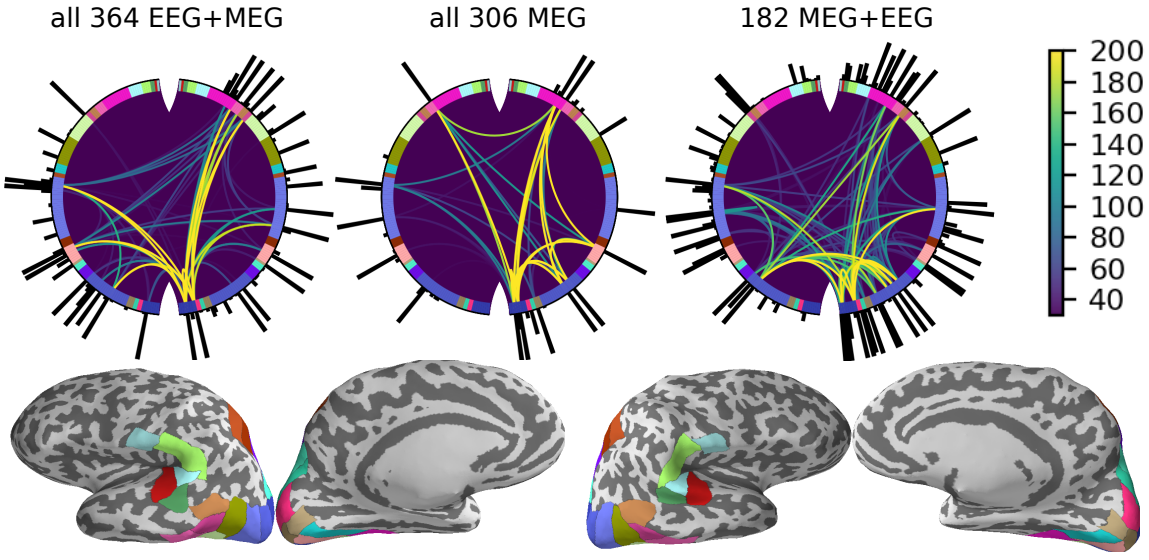


FIGURE 4.7.7: Source network analysis for visual data. The figures are constructed in the same way as described in Fig. 4.7.3 except that all 900 MCMC initializations are displayed.

for tackling non-convex optimization problems arising in frequentist regression, and HBM as a Bayesian prior modeling framework. We showed that one obtains the same algorithms, and therefore the same solutions, when considering some specific choices of models, parameters and inference strategies. In particular the parallel was done between the $\ell_{2,1/2}$-norm regularized regression by MM and the full-MAP estimation for $\ell_{2,1}$ hierarchical priors with specific Gamma hyper-priors. We further showed that this conceptual parallel can be exploited to improve the MM solution by providing well-informed algorithmic initializations.

For this, we first constructed a multi-layered Gibbs sampler for the joint posterior density of our HBM. Each sample is then used to initialize the MM step done with a state-of-the-art convex solver using block coordinate descent techniques and acceleration strategies based on active sets. The sampler used has also an efficient sub-sampler for $\ell_{2,1}$ priors at its core. Despite the multi-modality of the posterior, the MCMC scheme is able to jump rapidely between the different attractors of the MM scheme. Indeed, using each sample as an initialization to the MM computation, one ends up in many different local minima (*cf.* Figure 4.7.5, Figure 4.7.7). Therefore, this procedure allows us to reveal and explore different plausible source configurations in more details.

Based on this observation, we showcased how one can use the chain of local minima found by MCMC-initialized MM to analyze the variability of the different sparse solutions. Note that this is different from traditional and generic Bayesian uncertainty quantification techniques that use for example covariance estimates or credible sets derived from posterior samples [SVZ+15]. It is also different from methods developed specifically for parametric M/EEG source localization based on dipole fitting [FWK04; Dar+05]. These latter approaches cannot easily be transferred to sparse, non-parametric approaches. Using our developed techniques on simulations and actual data, one could observe that uncertainty in MEG/EEG is location specific and also source configuration specific. This is of course well-known by experts in this field, but here we provide a computational approach to visualize it and quantify it. This is an important incentive to develop such automated, data-dependent methods to quantify uncertainties in the context of MEG/EEG source imaging. In more conventional imaging methods such as Computer Tomography (CT) or MRI, the signal originates from weak tissue interaction with strong external fields and the forward operator $\mathbf{G}$ depends almost exclusively on the physical properties of the scanner. In this situation, uncertainty is usually distributed in a smooth, well-known way over the image domain. Artifacts as well as real anatomical features are also easy to distinguish for a trained radiologist. The situation for M/EEG is very different. The weak signals originate from endogenous activity, and they are very dependent on dataset specific factors such as source orientation, location and attenuation which all depend on the geometry of the head of the analyzed subject. That is also why the forward matrix $\mathbf{G}$ needs to be constructed for each individual patient, after fixing the electrical properties of the head issues,

which if wrong, increases the uncertainties.

When considering real data, the source to recover is often poorly understood, especially when it comes to pathological brain activity such as ictal or inter-ictal epileptic activity. In such a situation, providing a single source configuration as a result, together with an ad-hoc uncertainty quantification based on previous studies or acquired expertise, might not be an optimal use of the M/EEG data. Instead, providing multiple hypotheses together, along with a quantification of their uncertainty, can be more useful. Indeed for applications such as pre-surgical epilepsy diagnosis, where M/EEG recordings are one of several diagnostic modalities, each candidate source configuration can provide some evidence for or against a diagnostic hypothesis that could lead to a surgery decision. We therefore believe that extending the first steps we took here towards developing a consistent framework for interpreting and quantifying the multitude of potential results of sparse MEG/EEG source reconstruction approaches can have a significant impact on clinical settings.

Chapter 5

Benchmarking on Phantom datasets

5.1 Introduction

The previous chapters define various ways to solve the inverse problem of MEG/EEG brain imaging techniques. The evaluation of these solvers remain difficult due to the completely unknown ground-truth of the exact localization of the involved sources in each specific task. This limitation is primarily coming from the fact that the recording is done over the scalp and that multiple source configurations can lead to exactly the same measurements over the sensors. So the question is whether all these existing source localization techniques are able to accurately locate the positions and the orientations of current sources in the brain in a real acquisition scenario.

The typical way to answer this question is to perform simulations [LDB02; Mos+93; LBD98]. It consists in fixing the number and the location, orientation and amplitude of several dipoles in the brain, generate some simulated data corrupted by some additive noise [LDB02; Mos+93; LBD98; DM+02; Wol+06]. These simulations are unfortunately rarely realistic: they do not take into account the non-ideal nature of the sensors and the errors in the forward model, and they do not take into account the complex noise structure of real measurements. Inaccuracies in the computation of the forward operator are mainly due to approximations in the conductivity values in the head and/or the numerical errors associated with either spherical head approximations or BEM based on more realistic head geometries.

More sophisticated simulations might be investigated to overcome these issues, yet we propose here to use data collected from an artificial physical object in a real MEG machine. This has the advantage that the results can more closely reflect *in vivo* performance since they include factors that cannot readily be included in simulations, such as environmental noise.

In order to calibrate each MEG device, artificial objects that mimic the brain activity called "*phantoms*" are constructed by the manufacturers of MEG systems. They are based on the theoretical description in [Ilm85] producing realistic data corresponding to complex spatio-temporal current sources including realistic head geometries. In a typical phantom, from 4 to 32 independent current dipoles are present and MEG data are collected separately for each dipole. The true dipole locations and orientations, and the morphology of the brain, skull layers can be extracted from X-ray CT data [Lea+98]. One limitation is that such phantoms are not unsuitable for EEG, yet there exists some work on making EEG phantom devices [HSY16].

This chapter presents a new study to validate localization techniques using different publicly available phantom datasets. Other works have been done by [Haz+15; Lea+98; Bai+01] using also real-skull phantoms to investigate the performance of representative methods considering various head models.

The approaches considered in this chapter are mainly those described before in this thesis (see Chapter 2), namely: Dipole fitting (Section 2.3.1-page 18),

Gamma-Map, RAP-MUSIC, MxNE, irMxNE, TF-MxNE, irTF-MxNE. The methods not defined so far, will be briefly described in the next section.

5.2 Phantom datasets

5.2.1 Brainstorm-Elekta dataset

The description below was taken from the Brainstorm tutorial about the MEG phantom [Tad+11].

A current phantom is provided with Elekta for checking the system performance and can then be used for evaluation of the source localization (Figures 5.2.1 and 5.2.2). It contains 32 artificial dipoles and four fixed head-position indicator coils. The phantom is based on the mathematical fact that an equilateral triangular line current produces a magnetic field distribution equivalent to that of a tangential current dipole in a spherical conductor, provided that the vertex of the triangle and the origin of the conducting sphere coincide. For a detailed description of how the phantom works, see [Ilm85].

The phantom dipoles are energized using an internal signal generator which also feeds the HPI coils. An external multiplexer box is used to connect the signal to the individual dipoles. Only one dipole can be activated at a time. The location of the dipole is recorded relatively to the center of the sphere (0,0,0), where X is positive toward the nasion, Y is positive toward the left ear and Z is positive toward the top of the head.

The dataset has 32 dipoles, and is sorted into 3 different amplitude levels:

- Source with **2000 nAm**. This corresponds to an unrealistically strong 1000 nAm (2000 nAm peak-to-peak) dipole that gives the highest SNR of the experimental source.
- Source with **200 nAm**. This is a weaker dipole, closer to the range of amplitudes we can expect in raw data.
- Source with **20 nAm**. This represents some of the weakest sources we expect in evoked studies, which require averaging in order to be detected and estimated (*i.e.* generally cannot be seen in single trial analysis).

FIGURE 5.2.1: Reference phantom (RefPhantom) NM24058N (Serial number: 101861) provided by Elekta Oy, Helsinki Finland. 32 built-in simulated dipoles and four presetting head position indicator coils (HPI) (Figure taken from [Haz+15]).

5.2.2 MNE-Elekta dataset

The dataset has four dipoles named from 5 to 8, each with a different depth in the phantom head. The dataset contains dipoles with four different amplitudes: 2000 nAm, 200 nAm, 100 nAm and 20 nAm.

5.3 Methodology

5.3.1 Sphere models

The most commonly used head model assumes that it is made up of a set of nested concentric spheres, each with homogeneous and isotropic conductivity. Under this assumption, both the EEG and MEG problems have well-known closed form solutions [MLL99].

5.3.2 Preprocessing

The two Elekta phantom datasets have had a specific and identical preprocessing. First, we computed the forward operator with a single-shell sphere with origin (0., 0., 0.) and a head radius of 0.08, resulting in a discrete source space with 13782

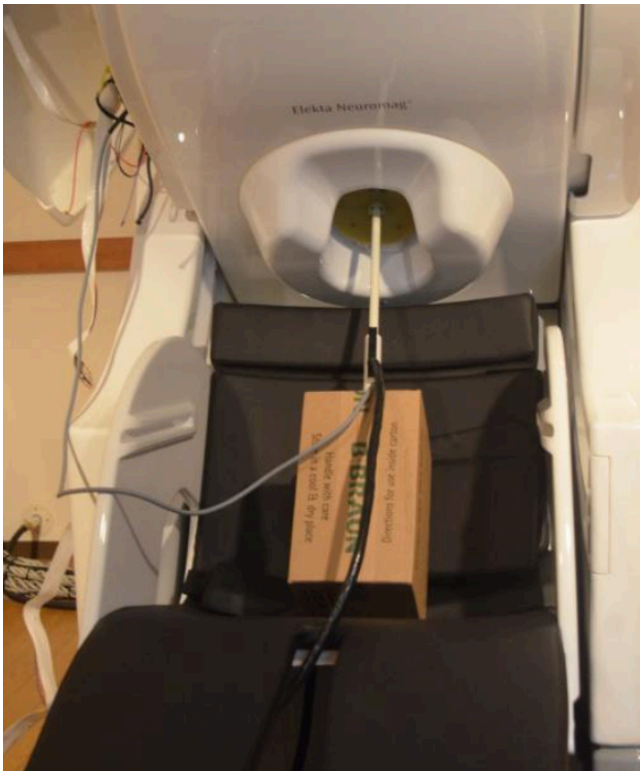


FIGURE 5.2.2: The phantom is carefully set into the sensor helmet of the probe unit and pushed against the helmet. HPI coil is fitted into outlet under the right gantry side cover (Figure taken from [Haz+15]).

location and free orientation ($O = 3$ orientations). Then we marked some channels as bad for each dataset:

- For Brainstorm-Elekta dataset, one channel has been marked as bad: "MEG2421".
- For MNE-Elekta dataset, three channels have been marked as bad: "MEG2233", "MEG2422", "MEG0111".

Second, Maxwell filtering was used to clean the data to to compensate for external magnetic interferences. The data was then low-pass filtered below 40Hz as we are interested only in frequencies around the simulated burst, which is at 25Hz. Once the filtering is done, we can construct the epochs and the evoked response from the filtered data from -100 ms to 100 ms. The covariance was computed on the baseline from -100 ms to 0. The sampling frequency for both datasets is of 1000Hz. The MNE-Elekta dataset resulted in 33 epochs before averaging, while the Brainstorm-Elekta dataset resulted in 20 epochs.

5.3.3 The selected solvers

Dipole fitting

Dipole fitting assumes that a small number of point-like ECDs can describe the measured topography. It optimises the location, the orientation and the amplitude

of the model dipoles in order to minimise the difference between the model and the measured topography. A good introduction to dipole fitting is provided by [Sch90]. A full description is done in Chapter 2-Section 2.3.1 (page 18).

γ-Map

γ-map offers a unifying view using a Bayesian perspective on some of the solvers presented here. As it was shown in Chapter 4, the Bayesian formulation of the MEG/EEG source localization consists in choosing priors/hyperpriors, estimation and inference procedure where often it boils down to alternating between the estimation of the source estimate and the estimation of the hyperparameter. The hyperparameter in this method is γ which represents covariance components. More details are in [WN09a].

RAP-MUSIC

Recursively Applied and Projected MUltiple SIgnal Classification (RAP-MUSIC) is an extension to MUltiple SIgnal Classification (MUSIC) by recursively estimating multiple sources [ML97; ML99b]. In other words, it consists in applying MUSIC successively after removing the contribution of the previously identified sources. In the same way matching pursuit algorithms are used for sparse signal decomposition over dictionary of atoms, the RAP-MUSIC method adopts a greedy strategy to select the relevant dipoles in a dictionary of sources.

MxNE | irMxNE

Mixed-Norm Estimates (MxNE) and Iterative Mixed-Norm Estimates (irMxNE) are the convex and the non-convex version of mixed-norms solvers using $\ell_{2,1}$-norm and $\ell_{2,0.5}$-quasinorm as regularizations. For more details, see Chapter 2-Section 2.5.2 (page 30) and [GKH12b; Str+16].

TF-MxNE | irTF-MxNE

Time-Frequency Mixed-Norm Estimates (TF-MxNE) and Iterative Time-Frequency Mixed-Norm Estimates (irTFMxNE) are the convex and the non-convex versions of the Time-Frequency mixed-norms using $\ell_{2,1} + \ell_1$-norm and $\ell_{2,0.5} + \ell_{0.5}$-quasinorm as regularizations. For more details, see Chapter 2-Section 2.5.2 and [Gra+13a; Bek+16].

All solvers except the irTF-MxNE are implemented in the MNE-python package [Gra+14; Gra+13b]. The dipole fitting, MUSIC, dSPM, MNE, sLORETA, γ-map have used the by default version. The regularization parameter of LCMV was set to $reg = 1$. The different hyperparameters of (ir)TF-MxNE were tuned on a grid search. The results below are shown after setting the window size to $wsize = 64$,

time shift to $ts = 4$, $\alpha_{space} = 40.$, and $\alpha_{time} = 1.$. While for the (ir)MxNE regularization hyperparameter was set to $\alpha_{space} = 40.$.

5.4 Experimental results

Three types of errors have been investigated for the different solvers, namely: the position or location error, the orientation error, and the amplitude error. All the solvers investigated here are implemented in the MNE-python package [Gra+13b; Gra+14].

The position error err_{pos} is represented in millimeters (mm), defining the distance between the exact location ($pos_{simulated} \in \mathbb{R}^3$) of the simulated dipole in the phantom head and the estimated location ($pos_{estimated} \in \mathbb{R}^3$) as in Equation (5.1). When the estimated source space contains multiple dipoles, the one having the biggest peak of amplitude is kept and compared to the exact dipole.

$$err_{pos} = 10^3 \|pos_{estimated} - pos_{simulated}\|_2 \tag{5.1}$$

The orientation error err_{ori} is represented in Radians (Rad), defining the angle between the exact orientation ($ori_{simulated} \in \mathbb{R}^3$) of the simulated dipole and the estimated one ($ori_{estimated} \in \mathbb{R}^3$) as in Equation (5.2). Same for the position error, the best dipole is kept when multiple ones are estimated.

$$err_{ori} = \arccos(|\langle ori_{estimated}, ori_{simulated} \rangle|) \tag{5.2}$$

Finally, the amplitude error err_{amp} is represented in penrcentage error (%), except for some solvers which give source estimates with statistical values and not electrical current values (example: dSPM). This error defines the relative difference between the peak of amplitude ($\max(amp_{estimated})$ with $amp_{estimated} \in \mathbb{R}^T$) of the estimated dipole and the peak of simulated dipole $amp_{simulated} \in \mathbb{R}$ (example 1000nAm, 200nAm, ... 20nAm) as in Equation (5.3).

$$err_{amp} = \frac{|\max(amp_{estimated}) - amp_{simulated}|}{amp_{simulated}} * 100 \tag{5.3}$$

5.4.1 Critical comparison of these MEG/EEG source localization

Figure 5.4.1 shows position errors obtained with most of the solvers for the four simulated dipoles (5 to 8) (see dataset 5.2.2) and for the different amplitude levels (20, 100, 200, 2000nAm). It shows less than 1mm error for the unrealistic high Signal to Noise Ratio (SNR) (peak-to-peak amplitude equal to 1000nAm), but also for 200nAM, and 100nAm. The location error gets worse with the very low SNR (20nAm).

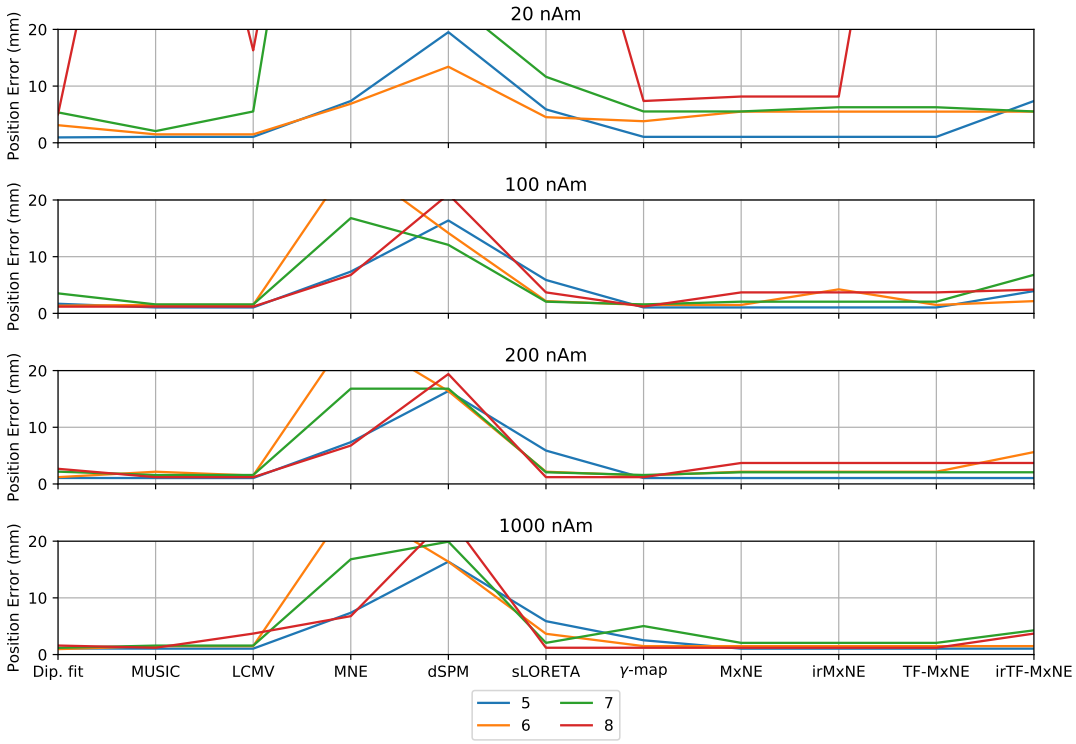


FIGURE 5.4.1: Comparison of the position error between most of the solvers for four different dipoles.

The dipole fitting approach is very suitable for localizing the neuronal activity when a small number of ECDs can describe the data. In this dataset 5.2.2, each dipole is recorded alone. The errors in orientations are displayed in Figures 5.4.2.

RAP-MUSIC is an approach based on the MUSIC technique which also performs very well when few ECDs are involved, especially when it is a dataset recording only one dipole at a time. It can be seen as very competitive, *w.r.t.* dipole fitting in Figures 5.4.1-5.4.2, showing the errors in position, and orientation respectively. However, for a deep source (dipole 8) combined with a very low SNR (20nAm), the red curve is outside of the box, meaning a location error bigger than 20mm. This is an issue with the signal subspace estimation, where the rank of the data covariance is not well estimated.

The γ-map which is a Bayesian formulation of the MEG/EEG inverse problem, performs worse than dipole fitting or RAP-MUSIC for very high and very low SNR. For SNRs in range of realistic data, its location error is upper bounded by 5mm depending on the depth of the studied dipole. The γ-map is doing worse for amplitude of 1000nAm compared to 100nAm or 200nAm, because it overestimates the noise when estimating the hyperparameter γ.

For MxNE and TF-MxNE, the errors shown in Figure 5.4.1 respectively demonstrate an equivalence or a slight improvement when using TF-MxNE compared to

MxNE, except for the deepest dipole 8 in red. This is explained by the fact that TF-MxNE is sensitive to the hyperparameters (λ_{space}, λ_{time}, window size, and time shift) depending on the SNR and the depth of the dipole. Here we tuned the hyperparameters on a grid search similar for all dipoles, although one might think that the hyperparameters depend on the easiness of the data (so on the SNR, and the depth of each dipole).

The same Figures 5.4.1- 5.4.2 also show an improvement when using the non-convex version in both irMxNE and irTF-MxNE compared to MxNE and TF-MxNE in localization and orientation, but the biggest difference is mostly seen in amplitude gain (*e.g.* Figure 5.4.6a for Brainstorm-Elekta dataset).

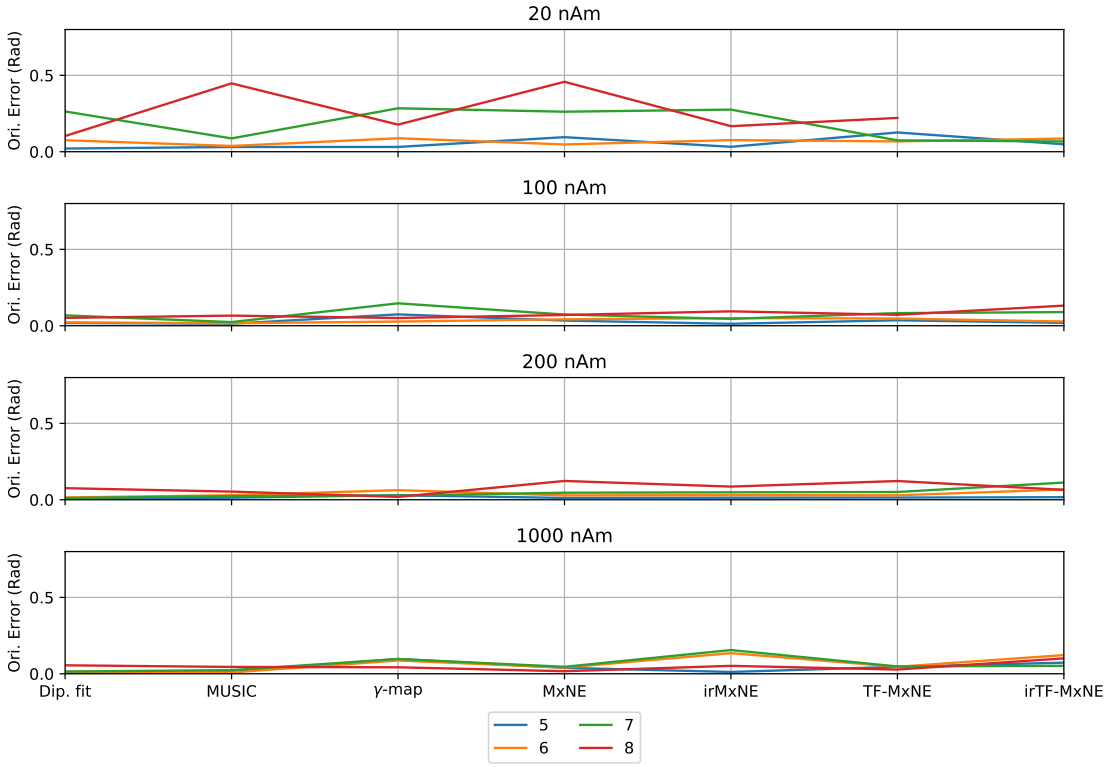


FIGURE 5.4.2: Comparison of the orientation errors for four different dipoles.

On the other hand, MNE and dSPM are surprisingly the methods giving the worst results for this dataset. One important argument is the fact that the study is biased as we know that the simulated phantom data is focal/sparse, while MNE and dSPM are not sparse methods. We always take the peak of amplitude and displays the best dipole for each method. sLORETA on the other hand is not a sparse method either, however, it performs much better than MNE and dSPM. The "center" of the pattern estimated with sLORETA is then closer to the exact dipole location compared to the center of dSPM or MNE.

The orientation error is comparable to the location error, where dipole fitting, MUSIC, γ-map, (ir)MxNE and (ir)TF-MxNE keep being the best performing methods. Dipole fitting is the best when simulating with only one dipole. Note however that dipole fitting does not suffer from the fixed grid resolution of 3.5 mm used by the other solvers. Figure 5.4.2 does not show orientation for LCMV, MNE, dSPM, and sLORETA as the orientation is not computed for source estimate. This is due to a current limitation of MNE-Python that only returns the magnitude of the sources when working with free orientation source spaces.

To make the source configurations harder, we considered doing linear combinations of the data produced by different dipoles. However, the locations of the dipoles are nearly the same except that the depth is different. When adding up this type of dipoles, they still have the same pattern in the sensor space, which makes it impossible to disentangle them.

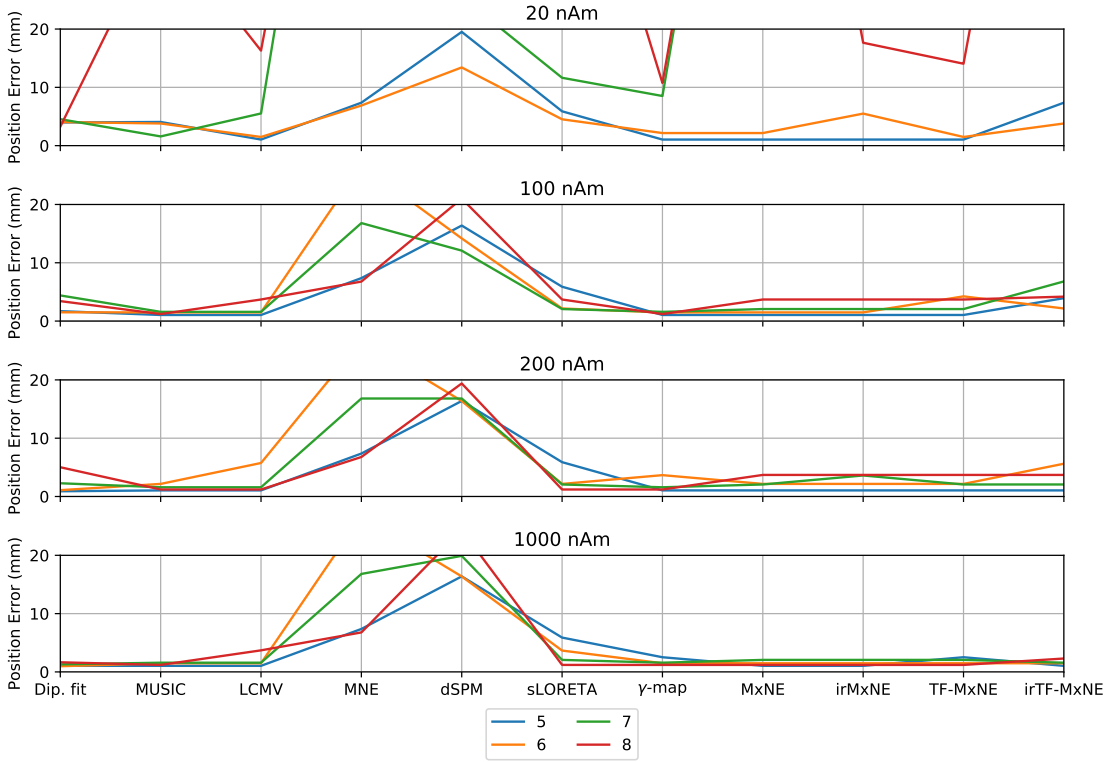
To investigate further the impact of each solver and the impact of a dataset with a high or a low SNR, we show in Figures 5.4.3a- 5.4.4a the results for the same solvers when we do not take all the epochs but only half of them. Those figures show a deterioration of dipole fitting when using the 20nAm, especially for the case where we keep only one epoch over four (Figure 5.4.4a). For 100nAm, 200nAm, and 1000nAm, the difference is small as the data corresponds already to very strong sources.

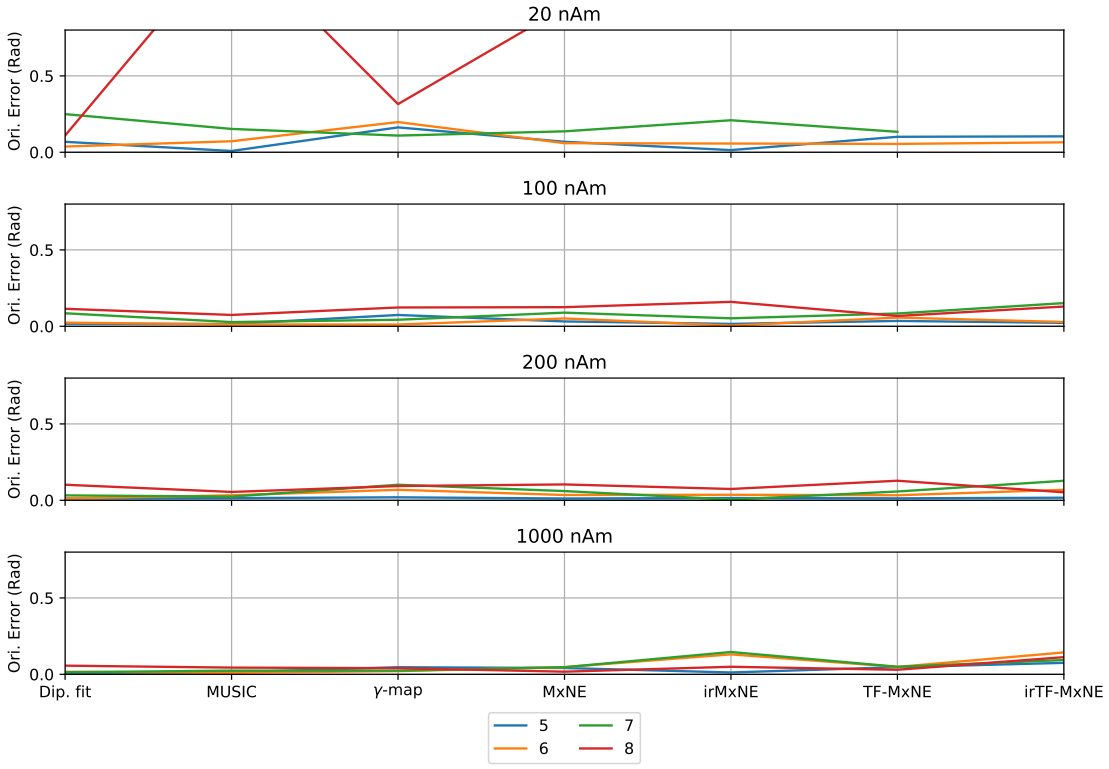
This analysis has been performed also on two other datasets which lead to the same conclusions (see Figures). The Brainstorm-Elekta dataset had 32 different dipoles, from which we take only four to display in Figure 5.4.5 for both amplitude and orientation errors.

Figure 5.4.6a shows the errors in amplitude for Brainstorm-Elekta dataset. The most important point in this figure is to see the difference between the convex MxNE | TF-MxNE and the non-convex irMxNE | irTF-MxNE in terms of amplitude bias. All dipoles basically improve their amplitude estimate when using the non-convex method (irMxNE | irTF-MxNE). For some cases, the irMxNE amplitude estimate is even better than the dipole fitting one and γ-map. Figure 5.4.6b shows the same effect of amplitude bias improvement even when we subsample the epochs, and take only one epoch over two.

The same analysis has been performed on the brainstorm CTF phantom dataset, which contains only one dipole with two different amplitude levels. The take home message remains the same, *i.e.*, dipole fitting is the most suitable solver when no more than one dipole is recorded at a time, however it performs very bad for low SNR. γ-map has been a very competitive solver compared to (ir)MxNE and (ir)TF-MxNE, but the iterative reweighted solvers remain the best for reducing the amplitude bias, and recovering almost the exact estimated amplitude (around 1% amplitude error).

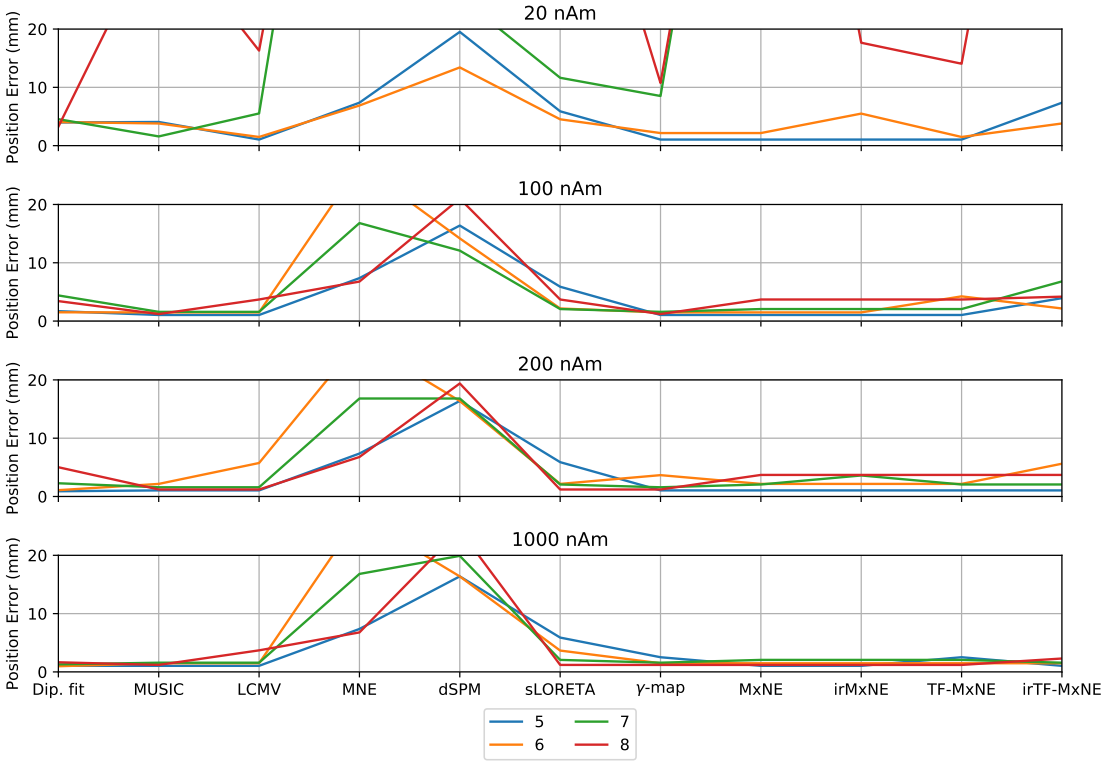
This study demonstrates the effectiveness of each method depending on the type of the dipole, *i.e.*, its depth, the level of its amplitude, and the SNR based on

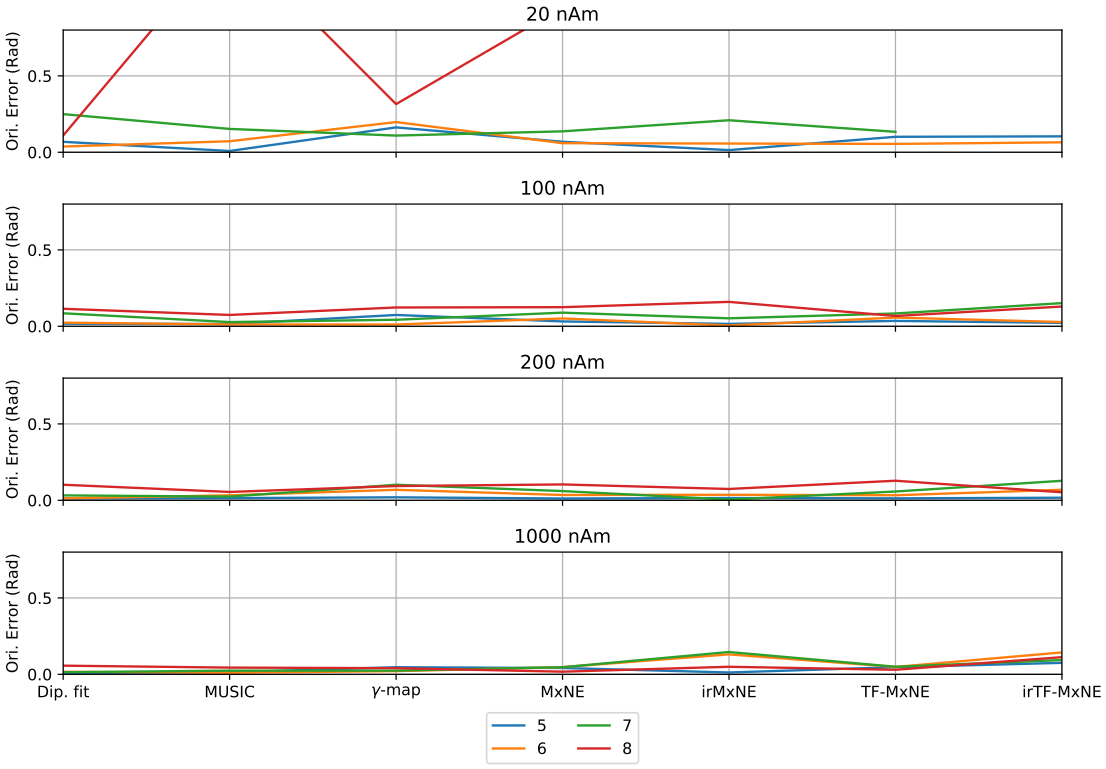




(B)

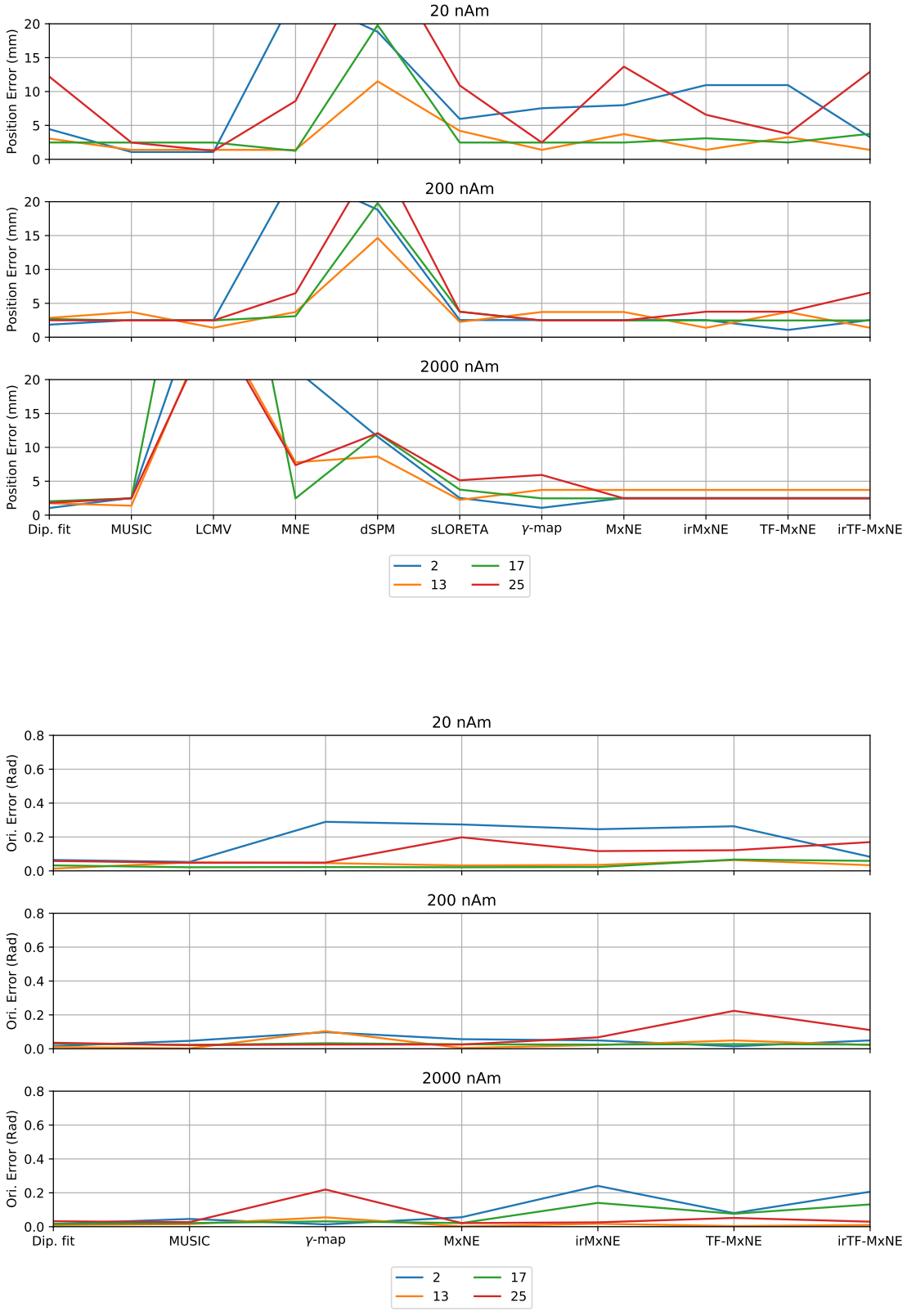
FIGURE 5.4.3: Comparison of the position and the orientation error between the solvers when taking only one over two epochs to reduce the SNR.





(B)

FIGURE 5.4.4: Comparison of the position and the orientation error between the solvers when taking only one over four epochs to reduce the SNR.



(B)

FIGURE 5.4.5: Comparison of the position and the orientation errors between the solvers for Brainstom-Elekta dataset. It shows 4 dipoles among the 32 for a good visibility. Nonetheless, it shows both the dipoles on the surface and the deepest ones.

the number of epochs. It specifically shows that there is no one best method for all use cases, but rather that the methods are complementary and that some of them works better with some specific conditions. Knowing the type of dataset one has in hand, is then primordial for a good source estimate recovery.

5.5 Conclusion & Perspectives

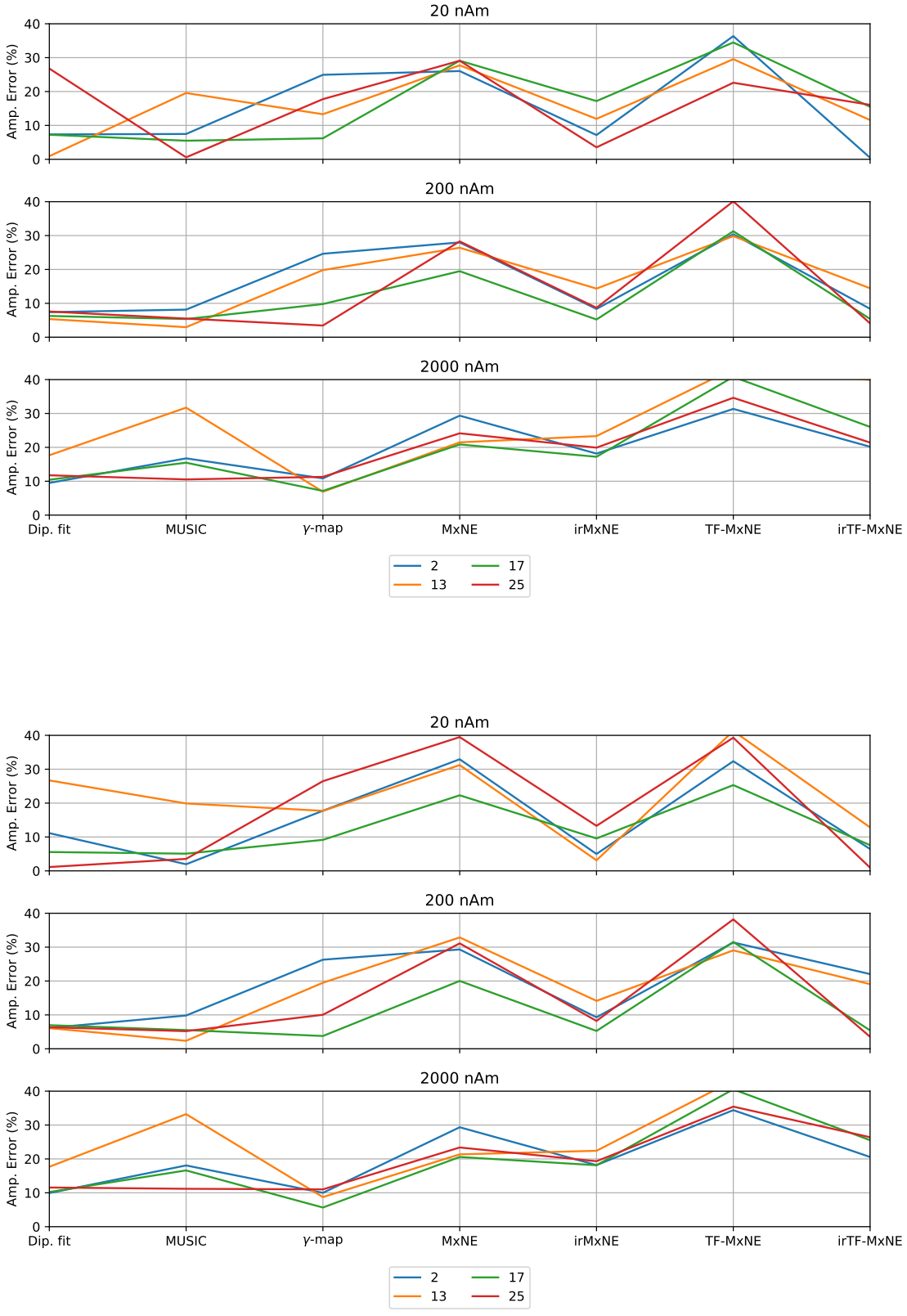
In this chapter, the main motivation was to present some results on various source localization techniques applied to phantom data. Being able to investigate "real" datasets with ground-truth is a big privilege to test the large list of existing methods for solving the MEG/EEG inverse problem.

Here we presented some of them, focusing on the approaches defined in this thesis. The conclusion would be that the dipole fitting is the most competitive and best method when having a focal dataset with only one dipole. Unfortunately here, we could not present a phantom dataset with two or more dipoles in the same recording, which would make the source localization more challenging.

A further work would be to investigate this aspect of multiple dipoles. The idea would be to confirm a better performance of convex and non-convex solvers (Variational or Bayesian formulation) compared to dipole fitting or MUSIC.

This chapter also did not show any hyperparameter tuning, or an automatic estimation as shown in Chapter 4, due to the "easiness" of the dataset when having only one dipole simulated. Indeed the location of the strongest dipole as extracted here for evaluation is barely affected by the choice of hyperparameters (provided they stay in reasonable ranges). This would not have been possible if we had two or more simulated dipoles, in which case hyperparameter setting would have been crucial. In this chapter, the hyperparameters to be selected were not having a big impact on the resulted source estimate, except for the TF-MxNE where several ones needed to be fixed. The size of the window size and the time shift of the dictionary were chosen as the best ones from a grid search.

In the near future, we plan to release all the code necessary to replicate these figures (data are already public). And we hope that this will foster new collaborations between researchers working on the MEG/EEG inverse problem. At least it should allow mathematicians and computer scientists working on this problem to more easily compare their methods to the state-of-the-art in the field.



(B)

FIGURE 5.4.6: Comparison of amplitude error between most of the solvers for 4 different dipoles amond the 32 existing in the Brainstom-Elekta dataset.

Chapter 6

Decoding visual motion from MEG

6.1 Introduction - Context

In natural environments, coherent motion is a vital sensory cue that helps the brain individuate objects in the world. Seminal neurophysiological work has described neurons in the middle temporal (MT) lobe of monkeys that were selective to the direction of motion and scaled to the coherence level of visual motion [Bri+92] . During a perceptual classification task, direction-selectivity can readily be decoded from the activity of neural populations in MT [JM06; Rus+06]. As visual motion processing relies on neural population codes, it is amenable to non-invasive functional human brain imaging such as fMRI or magnetoencephalography (MEG). Supervised learning techniques such as Multivariate Pattern Analysis (MVPA) are increasingly successful at characterizing where and when the neural analysis of stimuli such as visual orientation, motion direction or object classification is being realized [KT05; Wes+00; HR06; Pol11; CPO14; Hor+13; Hay15; War+16; RC16; Nak+03; Han+08; Ama+06].

In one of the earliest fMRI studies using MVPA, the direction of motion was successfully decoded from hMT+ (human analog of MT or V5) activity [BCD97]) but also, and surprisingly, from visual cortices V1, V2, V3 and V4 [KT06]. The successful decoding of visual motion in V1, V3 and hMT+ has since been reported several times [KT05; KT06; SB07; HV13; Kem+14]. In addition to the typical feedforward processing of visual information excepted in early visual cortices, the ability to decode visual motion from lower visual areas was interpreted as a marker of feature-based attention when required by the task [KT06] and an effect of top-down modulation of early feedforward processing for conscious perception [SB07]. However, whether brain decoding using MVPA captures the selectivity of neural populations or not has been a subject of debate on the interpretational weight given to decoding [Hay15; CW15; MKL15; PMH12]. Relevant to the current study, recent fMRI work has suggested that the sources of decoding in early visual areas may reflect the perceptual priors and biases of motion direction computation [VG14].

To disambiguate the functional role of different brain regions in motion selectivity, characterizing the temporal unfolding of pattern classification within and across visual regions could be helpful. Here, we exploited the temporal sensitivity of MEG to find the latency at which sufficient information had been integrated to reach a stable classification boundary [Mit+08; Ram+13; KD14]. 36 participants were recorded with MEG while performing a visual motion coherence discrimination task in which two intermingled clouds of visual dots (red and green random-dot-kinematograms) moved randomly on the screen until one of the clouds moved more coherently than the other one [Zil+14] (Figure 6.1.1A). Participants were asked to report which of the two populations became most coherent over time. Seven motion coherence levels were tested and a novel multivariate decoding approach combining ridge regression and a ranking metric was developed. Contrary

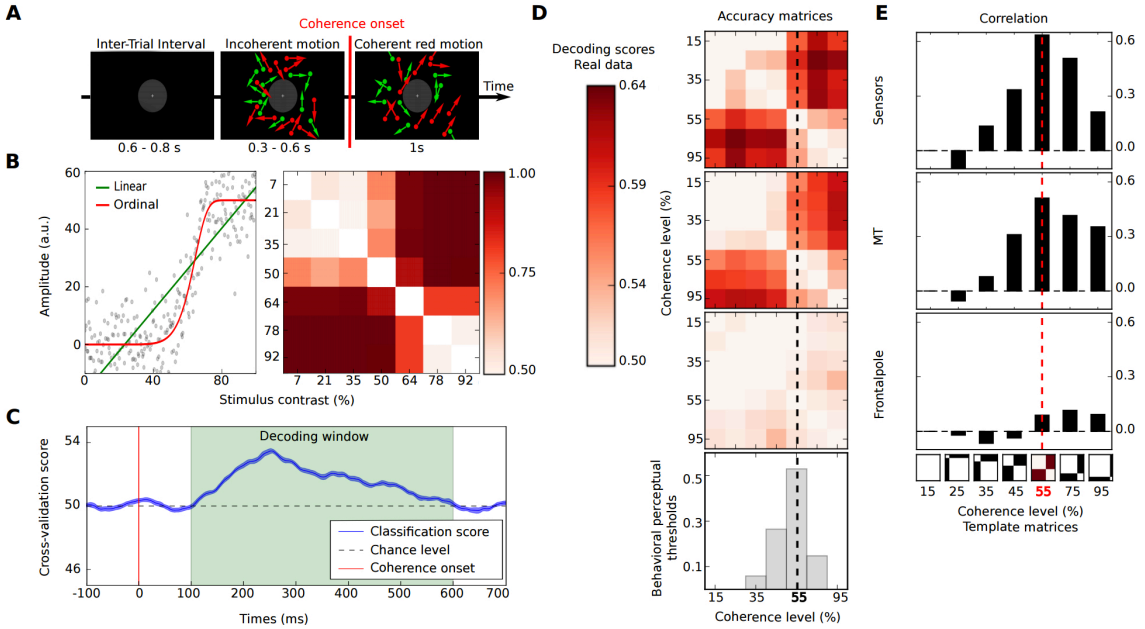


FIGURE 6.1.1: **Categorization Decoding.** A) One experimental trial in which participants discriminated which of the red or green cloud of moving dots was most coherent. B) Left: simulated data (gray) were best modeled by ordinal (red) than by a linear (green) fit. Right: similarity matrix providing a score of the decoding performance for each pairwise comparisons. C) Significant time-resolved decoding of visual motion coherence levels were found at 100 to 600 ms (green) post-stimulus onset. D) Grand-average (n=36) similarity matrices in sensors (top), hMT+ (middle) and frontal-pole (bottom) for the selected time window. Distributions of behavioral perceptual thresholds (gray histogram) and the mean (dashed line). E) Correlation scores between each template and similarity matrix (black histograms), and likeliest boundary decoded from MEG data (dashed red line).

to classical decoding approaches based on binary classifiers such as Support Vector Machines (SVM), a single decoder was estimated for all coherence levels, allowing robust parameter estimation despite high dimensional data. The ranking metric allowed taking into account the fact that visual motion coherence was an ordered variable [HGO00; Joa02]. This novel decoder was applied to brain activity recorded at the sensor level and to cortically-constrained source estimates. Using this decoding technique, we report the categorization of two separate brain states as a function of the degree of visual motion coherence. The categorization boundary matched participants' behavioral outcomes. Our results suggest that incorporating such decoding methods may be suitable to address questions relevant to predictive coding and perceptual decision-making.

6.2 Experimental design & Participants

Participants

Thirty-six participants took part in the study (16 females, mean 22.1 +/- 2.2 y.o.). All were right-handed, had normal hearing and normal or corrected-to-normal vision. Prior to the experiment, all participants gave a written informed consent. All methods were carried out in accordance with relevant guidelines and regulations and by NeuroSpin (Gif-sur-Yvette, France). The study was conducted in agreement with the Declaration of Helsinki (2008) and was approved by the Ethics Committee on Human Research at Neurospin (Gif-sur-Yvette, France).

Experimental design

The MEG session consisted of twelve experimental blocks alternating between rest and task [Zil+14]. Here, we solely focused on the main experimental task blocks in which participants' performance on a visual motion coherence task was being assessed. During the task, one trial started with the presentation of a fixation cross followed by two intermixed clouds of dots or Random Dot Kinematograms (RDKs) (red and green), whose motion was fully incoherent. After a variable interval of 0.3 to 0.6 s, one of the two RDKs became more coherent than the other one (Figure 6.1.1A). The participant had to determine by button press which of the red or green RDKs became more coherent. Seven possible levels of visual motion coherence were tested (15%, 25%, 35%, 45%, 55%, 75%, or 95%), randomly assigned to a colour and to a direction. Each participant was tested with 28 trials per visual coherence level.

Visual stimuli

The red and green RDKs were individually calibrated to isoluminance. In order to prevent local tracking of dots, a white fixation cross was located at the center of a 4° gray disk mask. RDKs were presented within an annulus of 4°-15° of visual angle. Dots had a radius of 0.2°. The flow of RDKs was 16.7 dots per deg$^2\times$ sec with a speed of 10°/s. During the first 0.3 to 0.6 s of a given trial, both RDKs were incoherent (0% of coherent motion). The duration of the incoherent phase was pseudo-randomized across each trial in order to increase the difficulty of the task by preventing participants' expectation of the temporal onset coherent motion. After the incoherent phase, one RDK became more coherent than the other one for one second in one specific direction. The direction of coherent dots was comprised within an angle of 45°-90° around the azimuth. At each frame, 5% of all dots were randomly reassigned to new positions and incoherent dots to a new direction of motion. Dots going into collision in the next frame were also reassigned a new direction of motion (more details in [Zil+14]).

Psychophysical analysis

The performance of each individual was averaged as a function of the seven degrees of visual motion coherence of the stimuli, irrespectively to the colour or direction of motion. The coherence discrimination threshold was set to 75% of correctness for each individual's data, as typically used in a two-alternative forced choice (2-AFC) paradigm, forcing participants to adopt the same decision criterion for all stimuli [GS66]. Here, the 75% detection threshold corresponds to chance level. They were then separately fitted to psychometric functions with the maximum-likelihood methodology (Psignifit [WH01]) which provided valid estimates of perceptual thresholds on a per individual basis (more details in [Zil+14]).

6.3 MEG pre-processing & source localization

All data pre-processing and source-imaging were done according to well accepted MEG guidelines [Gro+13]. Signal-Space-Separation (SSS) was performed on raw data using Maxfilter (Elekta-Neuromag [TS06]) to compensate for external magnetic interferences. MEG data were band-pass filtered (2 to 45 Hz), down-sampled to 250 Hz and epoched from -100 ms to 1000 ms relatively to the onset of RDK coherence. Trials that were contaminated by artifacts were rejected (e.g. peak-to-peak amplitude difference above 150 microvolts in EOG data) leaving 89% of trials considered to have an appropriate signal-to-noise ratio. The cortically constrained source reconstruction was done using the dSPM method following the guidelines of the MNE software [Gra+14]. The entire pre-processing was done using MNE [Gra+13c].

6.4 MEG decoding

Decoding generally consists in predicting a target variable y from one pattern of brain activity $x \in \mathbb{R}^p$ among all possible patterns or brain states. When the target can take a finite number K of possible values, like a multi-class classification problem, we have that $y \in \{1, \ldots, K\}$. Here, when x were MEG signals, p was the number of channels and time points used for the prediction. When x was the amplitude of cortical sources, p corresponded to the number of source locations. The first goal of this study was to estimate how well each pair of visual motion coherence level could be discriminated against each other. Considering that multi-class classification approaches do not take into account the ordinal nature of the target to predict, indeed predicting 1 instead of 7 is as bad as predicting 1 instead of 2 although the mistake is obviously smaller in the second case, we instead built a decoder which could yield a high pattern classification accuracy for distinguishable coherence levels, and a low pattern classification accuracy for nearby levels of visual motion coherence which were perceptually hard to differentiate (*cf.* next

two sections about the method).
The second goal of the study was to find whether separate categorical brain states (two or more) emerged following the presentation of the stimuli as a function of the seven levels of visual motion coherence. Specifically, the task of participants consisted in deciding whether the red or the green cloud of dots was most coherent as a function of coherence level. One working hypothesis was thus that at least one boundary delimiting a possible threshold between the neural activations induced by low *vs.* high coherent motion would be found during decoding.

To address this question, we opted out of a regression model estimated jointly for all levels of coherence, and combined it with a ranking metric adapted to discrete and ordered targets. Although an alternative approach could have consisted in testing the incoherent portion of the stimuli against each level of visual coherence, this would have lead to a strongly imbalanced training dataset (*i.e.* 196 incoherence trials for 28 trials per level of coherence) which is heavily problematic for MVPA classification approaches [HG09]. Specifically, with this formulation of the decoding, an inaccurate model which always predicts incoherence instead of coherence would have 85% of accuracy due to the imbalanced dataset. The ranking technique proposed here does not suffer from such class imbalance, considering that a single regression model was learnt for all coherence levels, and the ranking metric employed yielded 50% accuracy levels in spite of the low number of trials.

We now describe in detail the regression model employed.

Regression model

Due to the limited number of data points available for learning, and to the high dimensional nature of the neuroimaging data, we used a linear model following the standard approach in MVPA studies [KT05; Mit+08; Hay15]. The target values $y \in \mathbb{R}^n$, here provided for the n data points available for statistical inference, were derived from a linear combination of data, $y = X\omega$, where $\omega \in \mathbb{R}^p$ was a weight vector and X was a n-by-p data matrix. The value n here corresponded to the number of stimuli presentations, a.k.a. single trials or epochs. For each i^{th} observation, the target $y_i \in \{1, \ldots, K\}$ could take K different values: in this study, $K = 7$ corresponded to the seven levels of visual motion coherence defining the number of classes. Again, a multi-class classification approach could have been adopted, yet this strategy would have ignored that target values were ordered. For instance, decoding the 5th instead of the 2nd level of motion coherence is worse than predicting the 3rd level of motion coherence instead of the 2nd one. This is an information that a multi-class linear SVM model could not exploit. An SVM would also estimate $p \times K$ parameters instead of p, which would have naturally increase the risk of overfitting and reduced the interpretability of the results.

Instead, we chose a ridge regression method, and evaluated the predictive performance with a metric tailored for ordinal problems. The ridge regression model was defined as the solution to the convex optimization problem:

$$\hat{\omega} = \underset{\omega \in \mathbb{R}^p}{\arg\min} \|y - X\omega\|_2^2 + \lambda \|\omega\|_2^2. \tag{6.1}$$

The ridge regression model is a popular approach, whose practical success is due to fast estimation, robustness to noise and limited sensitivity to rough tuning of the parameter λ. Indeed, results obtained by ridge regression are known to be far less sensitive to the choice of λ parameter, compared to sparse estimators such as Lasso. In our experiments, λ was the same for all subjects [Var+17].

Decoding was performed on a per individual basis using all epochs. The 204 gradiometers and different time windows were tested: for example, for the time window ranging from 100 to 600 ms, the dimensions of the data were the number of samples $n = 196$ (at most 28 trials x 7 coherence levels) depending on the number of dropped epochs times the number of features $p = 204 \times 126 \sim 2.5 \times 10^4$, where the temporal window ranging from 100 ms to 600 ms contained up to 126 samples. The performance of the method was evaluated with a 10-fold stratified cross-validation which preserved the percentage of samples for each class or motion coherence level in each fold.

Decoding was also performed on source-reconstructed data in bilateral regions of interest (ROI), previously reported as being implicated in the task [Zil+14]. In source-space, the dimensions of the data were $n = 196$ at most and, for instance, $p = 126 \times 117 \sim 10^6$ depending on the size of the ROI (here, 117 dipoles in the ROI).

Following the estimation of the ridge regression model, a ranking metric was then employed to quantify the model performance while taking into account that the targets have a natural order.

Assessing decoding performance with pairwise ranking metric

Although ridge regression preserves the order of the target variables, it does not provide a relevant metric for the evaluation of the success rate of the decoder with an ordered set of categories. When using a linear regression model, the mean square error (MSE) is the natural performance metric. Yet, in high dimensional settings with a limited number of samples ($n \ll p$), as we are dealing with here, MSE is a poor metric. In order to reduce the variance of the estimated coefficients, high values of λ were used, causing a strong amplitude bias on the coefficients and a poor performance when measured using MSE. Performance evaluated with MSE was also affected in the presence of a bimodal state as illustrated in Figure 6.1.1B. Note that this strong bias problem is what motivates certain authors to use a Pearson correlation as a measure of performance rather than the MSE, although MSE is natural when using ridge regression [Kay+08].

In order to leverage the ordinal nature of the target values y, we quantified the performance in terms of ranking, where we tested the ability of the decoder to properly order pairs of samples, trials, based on the target to predict [HGO00; Joa02]. The ranking metric consisted in comparing the real values of y and the predicted ones. Let us consider two trials from the validation dataset with $y_i \neq y_j$ and where (y_i, y_j) denote their associated labels.

Let $\mathcal{P} = \{(i,j)\ s.t.\ y_i \neq y_j\}$ be the set of pairs with different labels. One quantifies the prediction accuracy Acc with the percentage of correct orderings for pairs of trials:

$$Acc = \#\{(i,j) \in \mathcal{P}\ \ s.t.\ \ (y_i - y_j)(y_i^{pred} - y_j^{pred}) > 0\} \tag{6.2}$$

For each pair of trials, there were two possible options and the chance level was therefore 50%. This quantity is related to Kendall's rank correlation metric [Kru58] which can be seen as a non-parametric correlation measure. To go beyond average accuracy, a key insight of this work was to inspect for which pair of trials the decoder made a mistake. For this, we thus defined a 7-by-7 similarity matrix M:

$$M_{y_i,y_j} = \frac{\#\{(i,j) \in \mathcal{P}\ \ s.t.\ \ (y_i - y_j)(y_i^{pred} - y_j^{pred}) > 0\}}{\#\{(m,n) \in \mathcal{P}, (y_m, y_n) = (y_i, y_j)\}} \tag{6.3}$$

Each $M_{i,j}$ was a value between 0 and 1 that told us how well we could distinguish the level i from the level j, 1 being the best; inversely, if the level i was similar or close to the level j, this decoding value would be close to chance level 0.5. The matrix was symmetric since comparing the levels i and j or j and i provides the same score. Such matrices, that can be seen as confusion matrices adapted for our pairwise ranking metric, are presented in Figure 6.1.1D.

Criteria for decoding categorization

Template matrices were defined for the discrete values of theoretically possible categorization into two brain states driven by the motion coherence levels, namely: 15%, 25%, 35%, 45%, 55%, 75%, or 95%. Each matrix had an on/off pattern at a given threshold (*e.g.* 55%) with values of 0.5 (off) or 0.65 (on) in order to make it comparable to decoding scores obtained in similarity matrices. An example is provided in the black matrices of Figure 6.1.1E. The correlation between the empirical matrices (fully based on MEG data) and all the possible template matrices, as defined above, provided the selection criterion to decode a categorization pattern at a specific threshold. Specifically, for each empirical similarity matrix, the template matrix yielding the highest correlation score was considered as a good predictor of the participants' motion coherence thresholds, eliciting the choice boundary from MEG data indicated as a dashed vertical line in Figure 6.1.1D.

6.5 Results & Discussion

Modeling of simulated data as a proof of concept

First, we modeled typical behavioral profiles observed during a perceptual discrimination task by using simulated data (*e.g.* ranging from 7% to 92% of coherence). The modeling allowed validating the use of an ordinal model which better fitted the data than a linear model (Figure 6.1.1B, left panel). As detailed above, the simulated trials were decoded using cross-validation by fitting a ridge regression to the training data and evaluating the performance of the model on all possible pairwise combinations of test trials. The similarity matrix (Figure 6.1.1B, right panel), which represents the predictive power in distinguishing two coherence levels, was evaluated with a 10-fold stratified cross-validation method. Each entry in the similarity matrix shows how similar each coherence level is to another one; alternatively, each entry can also be interpreted as how well one coherence level can be distinguished from another using a linear multivariate statistical model. All pairwise comparisons given in the similarity matrix built an anti-diagonal pattern: the lighter blocks in the similarity matrix were coherence levels for which no differences in brain responses could be captured, yielding a decoding score at chance level. Conversely, the darker blocks (red) captured high decoding accuracy scores for which brain responses highly differed between two coherent motion *e.g.*, brain responses to 7% coherent motion were highly distinguishable from those obtained during the presentation of 92% coherent motion. When comparing the neighboring levels 64% and 78% in Figure 6.1.1B-right panel, the high accuracy of decoding demonstrated a difference in brain activity patterns, reflecting a discontinuity in the activation profiles despite a progressive change in the visual motion coherence levels. The observed discontinuity or edge located between 50% and 64% of visual motion coherence revealed the presence of a categorical boundary.

Spatial selectivity of decoding categorization

The appropriate time window for best decoding performance was established using time-resolved cross-validation techniques [Ram+13]. The overall best decoding performance was obtained for latencies ranging from 100 ms to 600 ms post-motion coherence onset as illustrated in Figure 6.1.1C. The decoder was applied to MEG data in this time window on a per individual basis. Similarity matrices scored how well pairs of visual motion coherence could be distinguished, and then ordered, on the basis of brain activity. Figure 6.1.1D reports the similarity matrices computed on grand-average MEG data (n = 36 participants). Similarity matrices obtained for the MEG sensors (gradiometers) are reported in the top panel. Similarity matrices obtained for source-reconstructed estimates in the ROI hMT+ and in a control region “frontal pole” are provided in the middle and bottom panels, respectively.

The similarity matrices obtained in sensor and hMT+ data showed two distinct categories as an anti-block-diagonal pattern: two light blocks of decoding score at chance level (∼50%) for close coherence levels (low levels: 15%-45% against themselves, high levels: 55%-95% against themselves), and two dark blocks of decoding score nearing ∼65% for coherence levels that were apart, namely 15%-45% against 55%-95%. These results conform with the notion of perceptual categories, namely: visual motion coherence levels 45% and 55% were close from the point of view of the coherence level in visual stimulation, but distant in perceptual space with the former most likely classified as incoherent and the latter as coherent. The two brain states thus defined by the similarity matrix are compatible with categorical classification of the stimuli in this task. Specifically, visual motion coherence stimuli could either elicit a pattern consistent with not detecting the coherent signal in the display and not discriminating within the ensemble of stimuli whose coherence could not be detected (below the boundary) and detecting the coherent signal in the display but not discriminating within the ensemble of stimuli whose coherence could be detected (above the boundary).

To further investigate the link between brain activity at the single trial level and behavioral outcomes, we systematically compared the boundary delimited by the decoding approach with the perceptual threshold obtained from psychometric fits. The mean perceptual threshold was obtained in the task from the previous study [Zil+14] and shown here in the histogram over the 36 subjects (Figure 6.1.1D, bottom panel). The emerging categorical boundary at 45-55% of visual motion coherence in both sensors and hMT+ (but not frontal pole) matched well the mean perceptual threshold observed behaviorally (black dotted line; Figure 6.1.1D).

In order to establish a quantitative criterion for this observation, template matrices were constructed to model each theoretically possible perceptual threshold. Each template matrix was then correlated with each of the decoding similarity matrices obtained from empirical measurements (Figure 6.1.1E). The aim was to find the peak of the correlation between the template threshold and the emerging boundary. This procedure, which is similar in spirit to the Representational Similarity Analysis (RSA) approach [KMB08; CPO14], insured that the decoding similarity matrix was not forced to look like any specific template matrix. The quantitative metric confirmed our qualitative assessment (Figure 6.1.1D). Specifically, the peaks of the correlations were found for template matrices corresponding to a mean perceptual threshold of 55% in both MEG sensors and in source-reconstructed hMT+; the control ROI showed no selectivity.

Temporal accumulation selectivity of categorization decoding

The spatiotemporal sensitivity of source-reconstructed MEG data was exploited to test at which latency sufficient information had been integrated to reach a reliable

and stable classification pattern. To explicit the choice of the cumulative time window range, Figure 6.5.1A shows the grand average time course in response to the seven motion coherence levels over the 36 subjects in hMT+. As can be seen (Figure 6.5.1A) and as previously reported [Zil+14], main differences were located at these latencies although no clear categorization were visible in the time response. For this, scoring was established in a temporally cumulative manner from 100 ms post-motion coherence onset by adding the consecutive 50 ms time window to each previous one (Figure 6.5.1B) until 450 ms. The decoder was applied to sensors and to source estimates in the regions of interest as well as additional cortical sources known to be involved in the task [Zil+14], namely: hMT+ and the control region frontal pole but also the medial primary and secondary visual cortices (V1/V2), the intraparietal sulcus (IPS) and ventrolateral prefrontal region (VLPFC) (Figure 6.5.1, bottom left). In Figure 6.5.1B, in which all similarity matrices are reported, two brain categories of coherence levels seemed to emerge. As one of the focuses was to link the decoding to the behavioral data, the black dotted lines illustrated the known average perceptual threshold to find how well it fitted with the boundary found in the similarity matrices.

Using reverse-inference, we selected the template matrix which corresponded to the known mean perceptual thresholds of the 36 participants. We then computed the correlations in specific cortical regions to capture an anatomic and temporal discrimination. The correlation scores between the perceptual templates and the similarity matrices in the different cortical regions are provided in Figure 6.5.1C. The stability of the similarity matrices (Figure 6.5.1B) and the plateau of correlations between the template and the similarity matrices (Figure 6.5.1C) were first reached in hMT+ followed by V1/V2 in occipital regions, IPS and VLPFC. The latency of optimal decoding was consistent with seminal neurophysiology work suggesting functional selectivity of motion computation in hMT+ which may also be indicative of behavioral choice boundary [JM06; Rus+06; SB07; Bri+96]. Perceptual boundaries for motion coherence discrimination could also be decoded later on in regions implicated in the task (V1/V2, IPS and much later in VlPFC) but not in the control region. These observations suggest that the decoder was anatomically and temporally selective. Specifically, the sequence of decoding latencies suggests that the outcome of categorization computed in hMT+ may be forwarded downstream to V1/V2 – as a possible general mechanism contributing to plasticity - as well as VLPFC, as a likely consequence of perceptual decisions required by the task. The decision-related aspect was likely not encoded in low-level sensory areas, however the categorization pattern was still visible in hMT+ when appearing in VLPFC due to accumulation of evidence over the whole time range. Although one could argue that the emergence of these patterns over time are essentially due to longer integration windows for decoding, using sensors or hMT+ label yields a visible categorization pattern as early as 100ms.

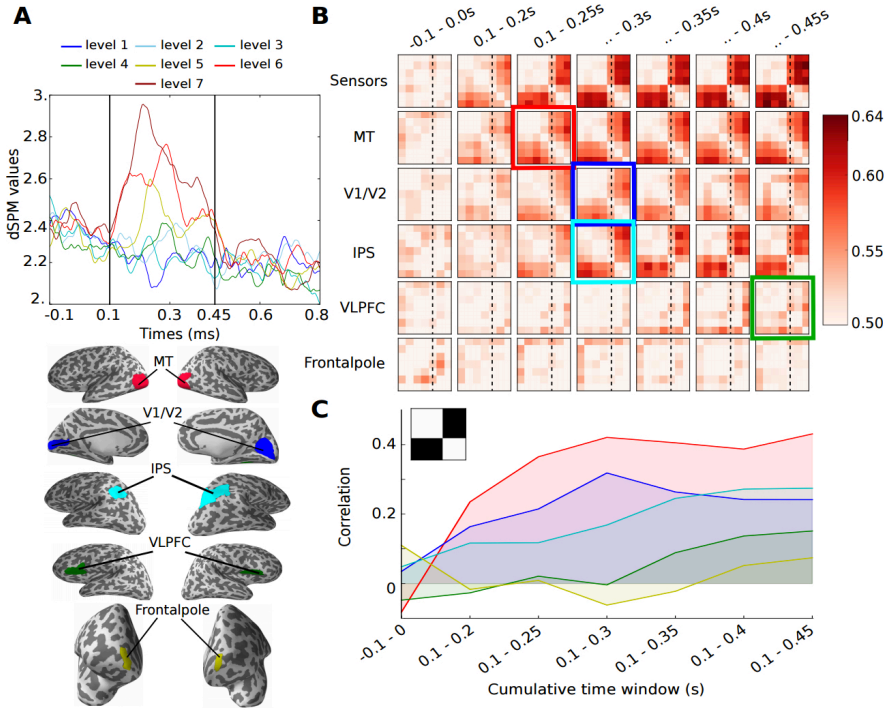


FIGURE 6.5.1: **Temporal-accumulation decoding.** A) Grand average hMT+ time courses in response to the seven motion coherence levels. B) Grand average similarity matrices (n = 36) in sensors, MT, V1/V2, IPS, VLPFC and frontal pole (top to bottom rows, respectively). Incremental decoding of the similarity matrices within the selected time window could be seen. Colored frames indicate the earliest decoding pattern capturing the perceptual thresholds (dashed lines) *e.g.* 250 ms for MT. C) Each similarity matrix was correlated with the template matrix optimally capturing perceptual thresholds. Correlations were cumulatively performed over the full time course of brain responses.

6.6 Conclusion

In this study, we showed that brain decoding could classify brain states as a function of visual motion coherence during a discrimination task. The categorical boundary partitioning two brain states was consistent with the participants' discrimination performance as indexed by their perceptual thresholds. Specifically, while the decoder was at chance level in discriminating between two motion coherence levels within the same perceptual category (within perceived or within non-perceived levels of visual motion coherence), the decoder performed well in discriminating brain activity in response to motion coherence levels across different categories (across perceived and non-perceived levels of visual motion coherence). We discuss below the implications and limitations of our findings.

In the visual motion coherence discrimination task used here, the intermixed clouds of dots (or RDKs) were identifiable by two distinct parameters: their color (red or green) and the increased degree of motion coherence in one cloud as compared with the other one. The task required participants to identify the colour of the coherent cloud of dots. Although the employed stimuli were quite typical for visual motion tasks, a couple requirements set this task apart. Firstly, the selective feature in the display was the coherence of motion irrespectively of the direction of motion. This differed from feature-based attention tasks in which the relevant feature is the direction of coherent motion [TT99; SS14]. Secondly, the task required the discrimination of two clouds of dots simultaneously presented and spatially intermingled; this was distinct from a previous decoding study in which the two populations were spatially segregated [SB07]. Nevertheless, and consistently with prior decoding work on visual motion processing [KT05; KT06; SB07; HV13; Kem+14] the earliest robust decoding of motion coherence was found in hMT+, as well as V1/V2. Thirdly, the color of the most coherent cloud of dots was randomized on every trial; as such, the color feature was orthogonal to the task requirement although the participants effectively classified their responses as "red" or "green". Accordingly, the successful decoding for any given pair of RDK coherence levels reported here (*cf.* cells in the decoding matrix being $> 50\%$) captured information about motion coherence *per se*, not its color or its direction.

The behavioral discrimination of continuous sensory information, such as coherent motion, requires the setting up of an internal criterion classifying sensory information into two or more categories [JM06; Bri+96]. Seminal work has shown that visual motion coherence at which neural activity reaches 50% of its maximum value can be estimated by means of a neurometric threshold [Bri+92]. A similar approach has been used on MEG source estimates in this task, revealing the extent to which the neurometric thresholds computed in the local brain area hMT+ could effectively reflect the participants' discrimination of visual motion coherence [Zil+14]. While perceptual thresholds can be derived using several analytical steps and fitting procedures, we have shown that a multivariate decoder can directly

capture the partitioning of brain activity as a function of the participants' performance by using a dedicated ranking metric associated with a template matrix correlated with the errors of the decoder when evaluated on left out test data. Our approach also showed that the partitioning of brain states fitting the perceptual thresholds at the population level could be found at different timings and at different cortical locations. Future ad-hoc investigations will focus on correlating individual perceptual thresholds and similarity matrices. Here, the individual similarity matrices were noisier, making it harder to interpret the outcomes. Hence, we compared conditions by taking the average over 36 participants, which still provided distinct categorization patterns.

Additionally, we found that the more sensory evidence accumulated over time, the more stable and robust the average similarity matrices became both in sensors and in brain regions. The first stable decoding pattern emerged in hMT+ ($\sim$250 ms), consistently with the known likelihood estimations and evidence accumulation of visual motion in this region and at this latency [Bri+92; JM06]. By 300 ms, a comparable decoding pattern was found in V1/V2, followed by IPS and by 450 ms by VLPFC. The early decoding latencies found in posterior regions and the later latencies found in frontal regions were overall consistent with decoding accuracies reported in perceptual decoding studies. Visual awareness can typically be decoded early in occipital regions and lately in frontal areas [And+16; MKL15; DGCK11; Sal+15; KPD16] . While the late decoding component is related to perceptual awareness, it can also reflect expectation, task requirements, and attentional selection [And+16; MKL15].

The observed spatiotemporal sequencing and stabilization of peak decoding in regions implicated in the task (but not others, *i.e.* control area frontal pole) suggest that the motion selectivity and choice probability computed in hMT+ could be passed on downstream to early visual cortices as well as to decision-related areas (IPS). Recent models of visual motion processing [Rus+06] and recent fMRI data [VG14] have suggested that perceptual priors in early visual cortices may be shaped on the basis of higher-levels computations. Both seminal and recent findings have suggested that attention and feature-selectivity may be crucial in the modulation of early sensory cortices [KT05; SB07; SS14]. Our MEG decoding results add to this literature by suggesting that selectivity to higher-order features computed in hMT+, such as motion coherence irrespective of direction or color may feedback to early visual cortices. These and other [HV13; Kem+14] results also suggest that the classification boundaries computed in hMT+ may have lasting effects for the analysis of visual motion. In particular, and consistently with previous literature [KT05; KT06; SB07; HV13; Kem+14], the latency of the categorization pattern across brain regions suggests the possibility that information relevant to perceptual boundaries from hMT+ feedbacks to V1/V2 consistently with

predictive coding models of visual processing [Rus+06; RB99] and learning theories [AH04; GLP09; SNW10; ROW10].

Nevertheless, it is noteworthy that in the context of perceptual categorization tasks such as the one employed here, the dissociation between the perceptual and the decisional components are difficult to disentangle [MKL15; KO13; HMU08]. Several studies have discussed the dissociation between perceptual processing and decision-making [PS06; PRS06; RPS09; ODK12; Wya+12; Lan+13; KO13; PMH12; MKL15]. For instance, a temporal dissociation between early sensory processing in occipital areas and decision-related processing in parieto-frontal regions have been shown to be increasingly pronounced over time [MKL15]. The perceptual thresholds used here to model the best fitting category do not readily dissociate between these two possibilities. Although the present study suggests that multivariate decoding can successfully retrieve perceptual thresholds, it is important to remain skeptical about the link between the information allowing decoding neural activity and its relationship to the computations effectively used to perform the task. For instance, brain activity categorized early on in hMT+ may contain top-down information feedback from decisional brain regions that may have helped the decoded categorization boundaries. However, three main aspects suggest that the decisional component may not be implicated here: firstly, the decision was made on the orthogonal feature color which was not used in the classifier as reported above. Secondly, the decoding in parietal cortices occurred much later than the stabilization observed in hMT+. Although response-locked analyses [KO13] could be used to disentangle the perceptual and decisional component, one limitation of the current decoder is that it is sensitive to any statistical differences in amplitude or in latency. Hence, analyzing the same time window sorted on the basis of the stimulus onset or of the response would not allow to draw stronger conclusions regarding the (perceptual or decisional) nature of the cortical representations enabling the categorization of brain states. Thirdly, recent evidence suggests that the inactivation of parietal regions are not decisive for motion categorization in monkeys [Kat+16].

To sum up, we presented a new MEG decoding technique that can capture the perceived categorization of continuous sensory information at the population level. Our results showed a sustainable pattern over time that correlated with the mean perceptual threshold and which successively implicated hMT+, V1/V2, IPS and VLPFC, consistently with general models of decision-making in motion categorization tasks [Maz+03]. Future work will aim at disentangling the perceptual analysis and the decisional components of perceptual decision-making tasks, as well as refining our approach to individual-level decoding.

Chapter 7

Conclusion & Perspectives

This thesis demonstrated various ways to solve the MEG/EEG source localization problem. It tackles specific challenges faced by current state-of-the-art techniques, and tries to improve them point by point:

- Promoting structured sparsity in the TF domain has proved useful for reconstructing non-stationary sources, although it needs to fix some parameters related to the Gabor transform, which are involved in the TF resolution. The first improvement proposed in this thesis was to tackle the choice of these parameters, which can be very detrimental for the analysis of brain waves with variable TF characteristics. It provides a new technique based on a multi-scale TF mixed norm allowing us to more accurately localize the source estimated in space and time (see Chapter 3).

- The formulation of the MEG/EEG inverse problem has been mostly written as a penalized regression, meaning that it needs to introduce a prior knowledge as a regularization term into the objective function. This results in adding a hyperparameter to the model which needs to be tuned. This thesis tackles this second challenge in two ways, both reformulating the problem as done in the Bayesian community. The Bayesian formulation allows to hierarchically add hyperparameters that are alternatively estimated with the main parameters of the model (the sources). The two main advantages are: the direct estimation of the hyperparameters, and the ability to use sampling in order to investigate the uncertainty of these solvers. These two points were presented in Chapter 4.

- An important step after developing any new technique is to validate it with a comparison with the other existing methods. This has for a long been a hard step as it is always hard to develop good and realistic simulations. For this aim, several studies have been investigating phantom datasets, which consists of real data recorded with a device mimicing a human head with focal sourdes. Chapter 5 shows a comparison of the solvers presented in this thesis on multiple phantom datasets.

This thesis was based on a long line of research started by my supervisor Alexandre Gramfort, and then his former PhD student Daniel Strohmeier, and was

meant to address several issues they encountered in the context of the MEG/EEG source localization problem. The points cited before were mainly the ones developed in this thesis, however several non-trivial ones remain in order to make best use of the available neuroimaging data:

- Although we found a way to solve the problem of source localization in the TF domain when having a mixture of brain signals in the data, it is still a non-trivial task to set the parameters of the multi-scale dictionary. As presented before, another possible way is to learn models that are good enough to capture the rich frequency content, and the morphology of the brain signal. The dictionary learning research line has given pretty nice results so far on electrophisiological signals [Jas+17b; Jos+06; BP16; Hit+17], which makes the technique completely autonomous and data-driven.

- The spatio-temporal techniques presented in this thesis are designed for the analysis of averaged evoked responses to ensure a descent SNR. Future work can be directed on how to make these techniques applicable on single trial data. A possible idea is to localize each trial separately by imposing an additional constraint onto the model, such as that the active set must be consistent over all trials [Str+12b; Str+12a].

- While invoking new constraints onto the model, another line of future work can be on the optimization side for solving the MEG/EEG inverse problem. In machine learning, various papers have been investigating mathematical and computational challenges to better tackle the inverse problem in general. One possible direction for the MEG/EEG inverse problem is to improve the computational complexity, because the proposed approaches need to be competitive in terms of running time. This results in research contributions that aim to accelerate the optimization algorithms; a practical example which was used in this thesis is the use of an active set. A more sophisticated approach would be to apply screening rules, *i.e.* find in advance the involved sources in order to compute the solution only for them, and avoid spending time on computing sources which will be inactive at the end [FGS15; MGS17; MSG; Ndi+16; Ndi+17].

- The proposed method in the TF domain still lacks an automatic model selection criterion to set the two regularization hyperparameters (one over space, the second over time). Chapter 4 presented a way to automatically set the hyperparameter for the mixed norm (MxNE) approach, which has only one regularization parameter over space. A future work could be to rewrite the problem for TF-MxNE, or investigate other model selection criteria.

- The novel methods and some of state-of-the-art approaches have been tested using three phantom datasets. A future work would be to investigate more

in depth this validation to compare their capabilities with a more sophisticated data, *i.e.*, two or more dipoles at a time, instead of only one dipole as presented here.

List of Figures

Abbreviations

AIC Akaike information criterion

BCD Block Coordinate Descent

BEM boundary element methods

BIC Bayesian information criterion

CV Cross-Validation

ECD Equivalent Current Dipole

EEG electroencephalography

EPSP Excitatory PostSynaptic Potentials

FISTA Fast Iterative Soft Thresholding Algorithm

HBM Hierarchical Bayesian Modeling

irMxNE Iterative Mixed-Norm Estimates

irTFMxNE Iterative Time-Frequency Mixed-Norm Estimates

ISTA Iterative Soft Thresholding Algorithm

LAD Least Absolute Deviations

Lasso Least Absolute Shrinkage and Selection Operator

MAP Maximum-a-posteriori

MCMC Markov Chain Monte-Carlo

MDCT Modified Discrete Cosine Transform

MEG magnetoencephalography

MM Majorization-Minimization

MNE Minimum Norm Estimate

MRI Magnetic Resonance Imaging

MUSIC MUltiple SIgnal Classification

MxNE Mixed-Norm Estimates

OLS Ordinary Least Square

RAP-MUSIC Recursively Applied and Projected MUltiple SIgnal Classification

SBL Sparse Bayesian Learning

SNR Signal to Noise Ratio

STFT Short Time Fourier Transform

TF Time-Frequency

TF-MxNE Time-Frequency Mixed-Norm Estimates

TFR Time-Frequency Representations

UQ Uncertainty Quantification

Résumé de Thèse: Contributions aux méthodes parcimonieuses pour la localisation de sources en MEG/EEG

Yousra Bekhti

June 21, 2018

Abstract

Cette thèse a développé des méthodes parcimonieuses pour la localisation de sources en magnétoencéphalographie (MEG) et l'électroencéphalographie (EEG).

Pour un champ électromagnétique donné, il y a un nombre infini de sources réparties à l'intérieur du cerveau qui aurait pu le créer. Cela signifie que le problème inverse est mal-posé, ayant de nombreuses possibles solutions. Cela nous contraint à faire des hypothèses ou des apriori sur le problème.

Cette thèse a étudié les méthodes parcimonieuses, i.e., seulement quelques sources focales sont activées lors d'une tâche précise. La première contribution est de modéliser le problème comme une régression pénalisée dans le domaine temps-fréquence avec un dictionnaire multi-échelle pour prendre en compte tous les aspects d'un signal cérébral. En ajoutant le terme de régularisation spatio-temporel, le modèle ajoute un hyperparamètre qui reste à optimiser. Ceci a constitué la seconde contribution de cette thèse où une estimation automatique des hyperparamètres a été mise en oeuvre.

La troisième contribution est de réduire l'écart entre les deux communautés qui formulent le problème inverse comme étant une régression pénalisée ou comme un modèle Bayésien. Cette thèse montre sous quelles hypothèses et sous quelle paramétrisation, on obtient une équivalence des deux formulations et comment profiter de cette nouvelle formulation Bayésienne pour quantifier l'incertitude de nos solutions.

La dernière contribution a eu pour but de valider les solveurs sur des données fantôme, c'est à dire des vraies données avec une réalité terrain pour pouvoir quantifier l'erreur de localisation en position, orientation, et amplitude.

Contents

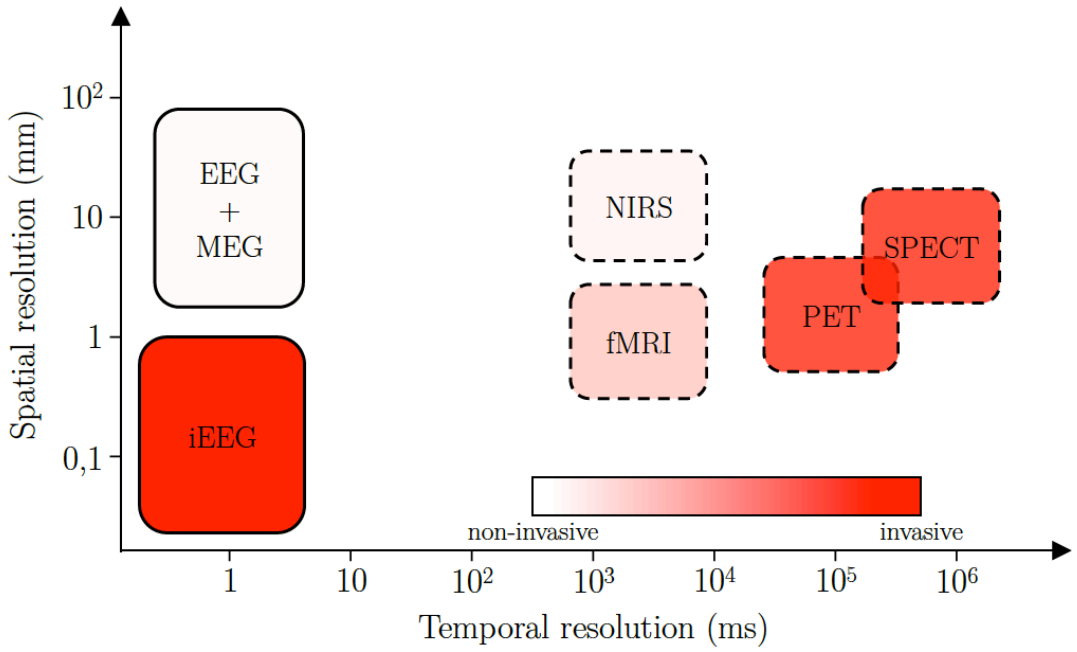


Figure 1: Vue d'ensemble des résolutions spatiales et temporelles de différentes méthodes de neuroimagerie fonctionnelle. Les approches directes (EEG, iEEG, MEG) sont indiquées par des boîtes pleines et des approches indirectes (IRMf, NIRS, PET et SPECT) par des tirets. Les couleurs des boîtes indiquent le degré d'invasivité.

1 Introduction

Un des plus grands défis du 20ème siècle est la compréhension du cerveau humain. Être capable de modéliser la façon dont le cerveau représente, analyse, traite et transforme l'information de millions de tâches différentes en un temps record est primordial pour les études cognitives et cliniques. Ces tâches peuvent aller du langage, de la perception, de la mémoire, de l'attention, de l'émotion, au raisonnement et à la créativité. Étudier le comportement du cerveau en chaque tâche et extraire des informations pour définir le réseau qui en est impliqué entraînera une meilleure compréhension de ses fonctions. Cela a été largement utilisé dans d'autres domaines tels que l'intelligence artificielle où les scientifiques et les ingénieurs essaient de mettre en œuvre les aspects qu'ils ont appris du cerveau humain dans les ordinateurs. Contrairement aux questions de sciences cognitives, dans le diagnostic clinique, comprendre comment une pathologie affecte le cerveau aide à trouver un remède ou un moyen d'améliorer la vie des patients. Par exemple, être capable de détecter l'autisme en bas âge de l'enfance aide les parents à fournir une éducation spécifique et un meilleur avenir.

Pour rendre ce balayage du cerveau possible, plusieurs technologies de pointe sont utilisées en fonction de la question que l'on se pose. Ces techniques diffèrent de leur degré d'invasivité et de leurs résolutions spatiales et temporelles, comme on peut le voir sur la Figure 1. Pour les différentes tâches que j'ai mentionnées ci-dessus, un aspect très important est le temps. Le cerveau est capable de traiter la plupart des tâches en une fraction de seconde, par exemple, de reconnaître une émotion, de percevoir un visage familier, etc. Dans cette thèse, pour étudier cette haute résolution temporelle du cerveau, j'ai été intéressée dans deux techniques d'imagerie cérébrale directe MEG et EEG.

MEG et EEG sont des techniques de neuroimagerie fonctionnelle pour cartographier l'activité cérébrale. Ils enregistrent respectivement les champs magnétiques et les courants électriques produits par l'activité électrique naturelle dans le cerveau au sein des neurones. Ils utilisent un ensemble de capteurs positionnés sur le cuir chevelu qui sont extrêmement sensibles à de minuscules changements dans le champ magnétique (mesurés par MEG) produits par de petits changements dans l'activité électrique (mesurée par EEG) dans le cerveau. C'est donc une mesure directe de

l'activité neurale. Ces techniques de mesure de l'activité cérébrales (MEG / EEG) ne sont pas nouvelles, mais ont été lancé à la fin des années 1960. Cependant, ce n'est que depuis le début des années 1990, avec l'introduction de grilles de détection haute densité couvrant toute la tête, que tout le potentiel de MEG a commencé à se réaliser. Le plus grand avantage du MEG et de l'EEG, comparé à l'IRMf, qui est beaucoup plus établi dans la recherche en neurosciences, est la résolution temporelle. Dans IRMf, l'activation neuronale est indirectement mesurée par des changements locaux dans le niveau d'oxygénation du sang, et une fenêtre de temps est généralement compressée dans un volume cérébral mesuré. Les autres techniques mentionnées dans la Figure 1 sont également des techniques d'imagerie cérébrale fonctionnelle indirectes.

En utilisant des magnétomètres / électrodes (capteurs) très sensibles, MEG et EEG fournissent un aperçu de l'activité cérébrale avec une haute résolution temporelle et spatiale. Ils permettent les mesures de l'activité en cours qui décrivent l'état des sources actives du cerveau à chaque milliseconde. Ce problème de calcul du résultat des mesures est appelé problème direct (*forward problem*). Le problème direct bioélectromagnétique décrit la relation entre une activité neuronale donnée dans le cerveau et les signaux MEG et EEG observables. Sa solution modélise mathématiquement l'activité neurale, la conductivité du volume et la configuration du modèle. Il nous permet de relier les potentiels et les champs externes à une distribution de courant interne par une solution stable et unique, ce qui est donc un problème bien posé.

Sa contrepartie, le problème inverse bioélectromagnétique, consiste à utiliser les mesures réelles pour déduire les paramètres (emplacements, amplitude, orientations) donnant la distribution des générateurs neuronaux. C'est un problème mal posé dans le sens de Hadamard [Had02] en raison de sa non-unicité et de sa grande sensibilité au bruit, ce qui rend sa solution instable. Le problème inverse est ce qu'on appelle le problème $n \ll p$ en apprentissage automatique, où on a beaucoup plus d'inconnues p à estimer que le nombre d'observations ou de variables n. Ce problème a des solutions infinies, principalement dues au petit nombre de capteurs (obervations n) présents en MEG et en EEG. Pa contre, même si MEG et EEG étaient mesurés simultanément avec un nombre infini de points au-dessus de la tête, l'information serait encore insuffisante pour calculer de façon unique la distribution de la source cérébrale qui a généré les signaux cérébraux mesurés. Ceci est dû au fait qu'il existe différentes combinaisons de sources capables de provoquer exactement les mêmes champs potentiels sur la tête. Ainsi, pour inférer l'activité neurale générant les données au niveau du capteur, différentes techniques de reconstruction de source peuvent être appliquées, qui utilisent typiquement des connaissances a priori sur l'état de l'activité cérébrale afin de réduire l'ensemble des solutions à une solution unique.

Dans cette thèse, je me suis intéressé aux modèles parcimonieux pour reconstruire et localiser les sources en MEG et EEG. Dans le but d'obtenir des solutions acceptables, faciles à interpréter, il ne suffit pas d'avoir un bon apriori sur les données, mais plusieurs questions doivent être posées:

1. Quelle est la meilleure façon de fixer la régularisation pour obtenir des solutions parcimonieuses facile à interpréter?
2. Comment peut-on estimer les hyperparamètres du modèle?
3. Comment quantifions-nous l'incertitude de ces modèles?
4. Comment comparer objectivement les différents solveurs de l'état de l'art?

Ces points définissent l'étendue de cette thèse. Elle tente d'aborder d'abord le problème des sources non stationnaires, c'est-à-dire, comment estimer une source qui a une explication neuroscientifique comme étant actif pendant une courte fenêtre de temps seulement, en étudiant une fenêtre plus longue. Cela implique la formulation du problème dans le domaine temps-fréquence, qui doit expliciter le dictionnaire de la décomposition. Deuxièmement, cette thèse tente de trouver un moyen d'estimer automatiquement l'hyperparamètre du modèle de régression pour faciliter la comparaison entre les solveurs. L'étape suivante consistait à réécrire le problème comme cela a été fait par d'autres communautés dans une formulation bayésienne. Ceci a ouvert la voie pour combler le fossé entre les mondes variationnel et Bayésien en écrivant leur équivalence sous une paramétrisation spécifique du même problème. L'avantage de la formulation bayésienne est la possibilité d'étudier la distribution postérieure, rendant possible une étude de l'incertitude de la

solution.

Cette thèse présente de nouvelles approches pour la reconstruction de sources en MEG / EEG. Elle peut être divisée en quatre projets principaux:

1. L'implémentation d'un algorithme largement connu pour le problème inverse MEG / EEG appelé *"Recursively Applied and Projected" (RAP) MUSIC* [ML99]. Le but étant d'avoir une comparaison avec un solveur de l'état de l'art basé sur une régularisation non-convexe qui favorise une plus grande parcimonie en se débarrassant de toutes les sources parasites. Ce travail a été publié dans le journal *IEEE Transactions on Medical Imaging (TMI)* [SBHG16].

2. L'amélioration d'un travail antérieur de Daniel Strohmeier sur la reconstruction de sources dans le domaine temps-fréquence, qui a abouti à l'introduction de l'algorithme TF-MxNE (Time-Frequency mixed-norm). La contribution aborde le problème du choix du dictionnaire utilisé pour décomposer les données lorsque l'on travaille dans le domaine temps-fréquence. Ceci consiste à permettre la possibilité d'utiliser des dictionnaires combinés pour rendre l'algorithme capable de trouver à la fois des formes d'onde transitoires et des ondes plus longues présentes dans le signal cérébral. Ce travail a été publié dans IEEE workshop on Pattern Recognition in NeuroImaging (PRNI) [BSJ+16].

3. Différentes lignes de recherche pour résoudre le problème inverse MEG / EEG ont donné différentes formulations. La formulation la plus fréquemment utilisée dans cette thèse est sous la forme d'un modèle de régression régularisé comme dans la plupart des problèmes d'apprentissage automatique. Avec ce genre de modèles, il faut trouver un bon compromis entre l'attache aux données, et le terme qui régularise le problème qui prend en compte toute hypothèse que l'on a sur le problème. Ce compromis est contrôlé par un paramètre externe généralement appelé hyperparamètre. Pour un exemple pratique, lorsque l'on utilise une régularisation parcimonieuse, si cet hyperparamètre est fixé à une petite valeur, la solution résultante ne sera pas suffisamment parcimonieuse et vice versa. Ainsi, la deuxième contribution de cette thèse a été alors de trouver un moyen automatisé d'estimer cet hypermaramètre sous certaines conditions du modèle. Ce travail a été publié dans la conférence européenne de traitement du signal (EUSIPCO) [BBG17].

4. Le plus grand inconvénient des solveurs parcimonieux réside dans le fait qu'ils donnent une solution sans aucune estimation de la variance ni aucun type d'intervalle de confiance. D'autres domaines d'application utilisent l'inférence bayésienne, principalement parce qu'elle permet d'estimer l'incertitude et que sa quantification est primordiale. Par conséquent, la troisième contribution de cette thèse est de réécrire le problème comme dans un monde bayésien, et tente de rapprocher cette formulation de ce qui a été présenté jusqu'à présent. Ce projet montre que sous certaines conditions, la formulation bayésienne et la formulation variationnelle sont équivalentes. Ensuite, il montre comment nous pouvons tirer parti de la distribution a posteriori pour extraire des cartes d'incertitude. Ce travail a été publié dans le journal Inverse Problems [BLSG18].

5. Le dernier projet de cette thèse était de tester et de valider nos solveurs et plusieurs autres qui sont les plus utilisés à ce jour pour les applications en neurosciences. Ceci est fait sur un jeu de données fantômes qui est un jeu de données simulé avec un environnement réaliste similaire à un vrai cerveau humain. Ce travail devrait être soumis bientôt à un journal.

6. Un projet supplémentaire sur le décodage cérébral est présenté à la fin de cette thèse. Ce travail présente une nouvelle approche basée sur une régression ridge avec une métrique spécifique prenant en compte le fait que la cible est ordonnée. L'approche est nouvelle en termes d'application aux données MEG. Ce travail a été soumis et est en cours de révision dans la revue Plos One [BGZvW17].

7. L'implémentation de certaines contributions est déjà sur le package MNE-Python [GLL+14, GLL+13], les autres devraient également être intégrés prochainement. Une autre contribution avec le projet d'un collègue est publiée dans Pattern Recognition in NeuroImaging [JER+16] et Neuroimage [JEB+17].

1.1 La structure de cette thèse

Se referer à ma thèse écrite en anglais pour plus de détails.

1. **Chapitre 2: Contexte et travaux liés au problème inverse MEG / EEG**
   Ce chapitre définit les bases et le contexte nécessaires pour ce qui sera présenté dans le reste de cette thèse. Il commence par donner l'origine des enregistrements MEG et EEG, c'est-à-dire, que mesurent réellement ces techniques? Il donne ensuite plus d'informations sur l'opérateur direct et comment il est calculé efficacement. À ce stade, je présente un état de l'art des problèmes inverses définissant les trois principales approches: les techniques de formation de faisceau ou de balayage, les méthodes basées sur l'image avec des modèles distribués, et les modèles de source parcimonieux. Ensuite, je présente quelques bases de la décomposition temps-fréquence, et compare plusieurs dictionnaires en donnant leurs avantages. Je termine ce chapitre par une section d'optimisation, en définissant différentes façons de régulariser le problème mal posé et ensuite comment les résoudre. Il donne également une comparaison entre plusieurs solveurs.

2. **Chapitre 3: Localisation de sources avec des dictionnaires multi-échelles**
   Ce chapitre est dédié à notre première contribution, à savoir la résolution du problème inverse dans le domaine temps-fréquence à l'aide d'un dictionnaire multi-échelle. La localisation de source dans le domaine temps-fréquence a déjà été étudiée en utilisant un dictionnaire de Gabor de façon convexe [GSH+13] et non convexe [SGH15]. Cependant, le choix d'un dictionnaire optimal reste non résolu. En raison d'un mélange de signaux, c'est-à-dire des signaux transitoires courts (juste après le début du stimulus) et des ondes cérébrales plus lentes, le choix d'un dictionnaire unique expliquant les deux types de signaux de manière parcimonieuse est difficile. Ce chapitre présente une méthode pour améliorer l'estimation de source en se basant sur un dictionnaire à plusieurs échelles, c'est-à-dire plusieurs dictionnaires avec différentes échelles concaténées pour s'adapter aux transitoires courts et aux ondes lentes en même temps. Les avantages de cette approche sont présentés en termes de détection des deux types de signaux, de lissage temporel, et non mixture entre les sources.

3. **Chapitre 4: Le lien entre les modèles bayésiens et les normes induisant la parcimonie**
   Ce chapitre présente les concepts de base de la formulation bayésienne du problème inverse MEG / EEG. Il vise également à expliquer le différent jargon pour lier les définitions variationnelles et bayésiennes. Ceci résulte dans la définition d'une équivalence entre les deux communautés sous certaines conditions en profitant de la formulation bayésienne qui permet d'étudier les multiples modes de la distribution postérieur. Les modes du postérieur définiront plusieurs solutions possibles au problème inverse, permettant alors l'obtention de cartes d'incertitude des estimations sources.

4. **Chapitre 5: Benchmarking sur des données fantômes**
   Ce chapitre est un chapitre de validation sur un jeu de données fantôme. Les données fantômes sont un ensemble de données obtenu en mesurant l'activité MEG / EEG avec une tête fantôme de crâne humain. Tous les aspects réels d'une tête sont simulés pour générer la même conductivité que celle attendue avec un vrai crâne. L'ensemble de données présenté dans ce chapitre comporte quatre dipôles simulés à différentes profondeurs. Avec la connaissance de la vérité terrain, ce chapitre étudie l'efficacité de chaque solveur en termes de localisation de source, d'orientation et d'amplitude.

5. **Chapitre 6: Décodage du mouvement visuel de MEG**
   Ce chapitre illustre un projet supplémentaire en dehors du sujet de problème inverse. Il est basé sur une application de l'apprentissage automatique aux neurosciences. L'objectif était de développer une approche efficace pour décoder l'activité cérébrale enregistrée avec MEG pendant que les participants discriminaient la cohérence de deux nuages de points entremêlés.

1.2 Publications durant cette thèse

Revue | Journal :

- **Y. bekhti**, and A. Gramfort, "Validation of dipole localization using phantom data in MEG source imaging," in preparation.

- **Y. Bekhti**, F. Lucka, J. Salmon, and A. Gramfort, "A hierarchical Bayesian perspective on majorization-minimization for non-convex sparse regression: application to M/EEG source imaging," Inverse Problems, 2018
- **Y. Bekhti**, A. Gramfort, N. Zilber, and V. van Wassenhove, "Decoding the categorization of visual motion with magnetoencephalography," Plos One, (submitted).
- D. Strohmeier, **Y. Bekhti**, J. Haueisen, and A. Gramfort, "The iterative reweighted Mixed-Norm Estimate for spatio-temporal MEG/EEG source reconstruction," IEEE Transactions on Medical Imaging, vol. 35, no. 10, pp. 2218-2228, 2016.
- M. Jas, D.A. Engemann, **Y. Bekhti**, F. Raimondo, and A. Gramfort, "Autoreject: Automated artifact rejection for MEG and EEG data," NeuroImage, vol. 159, pp. 417-429, 2016.

Conférence :

- **Y. Bekhti**, R. Badeau, and A. Gramfort, "Hyperparameter estimation in maximum a posteriori regression using group sparsity with an application to brain imaging," Signal Processing Conference (EUSIPCO), pp. 246-250, 2017.
- **Y. Bekhti**, D. Strohmeier, M. Jas, R. Badeau, and A. Gramfort, "M/EEG source localization with multi-scale time-frequency dictionaries," International workshop on Pattern Recognition in NeuroImaging (PRNI), pp. 1-4, 2016.
- M. Jas, D.A. Engemann, F. Raimondo, **Y. Bekhti**, and A Gramfort, "Automated rejection and repair of bad trials in MEG/EEG", International workshop on Pattern Recognition in NeuroImaging (PRNI), pp. 1-4, 2016.

2 Modélisation du problème

Notation
Dans une application MEG / EEG, l'opérateur direct $\mathbf{G}$ qui décrit la relation linéaire entre les mesures MEG / EEG $\mathbf{M} \in \Re^{N \times T}$ (N nombre de capteurs, T nombre d'instants temporels) et l'activation de source $\mathbf{X} \in \Re^{S \times T}$ (S est le nombre des emplacements possibles de sources). Le modèle linéaire lit alors: $\mathbf{M} = \mathbf{GX}$ où $\mathbf{G} \in \Re^{N \times T}$ est le gain ou la matrice de l'opérateur direct, une matrice de mélange instantanée connue, qui relie les signaux de source et de capteur.

Parmis toutes les formulation de l'état de l'art, on s'intéresse aux méthodes distribuées pour la localisation de sources. Ces méthodes permettent d'estimer les amplitudes d'un ensemble dense de dipôles distribués à des endroits fixes dans la surface ou le volume de la tête. Ces méthodes sont basées sur la reconstruction de l'activité électrique cérébrale au niveau chaque point d'une grille 3D de points de la solution, le nombre de points étant beaucoup plus grand que le nombre d'électrodes sur le cuir chevelu. Chaque point de la solution est considéré comme un emplacement possible d'une source de courant, il n'y a donc pas d'hypothèse a priori sur le nombre de dipôles dans le cerveau [Was08].

MNE (*Minimum norm estimate*) [HI94] et ses variantes [LWA+06, PMML94, DLF+00, PM+02b] résolvent le problème inverse MEG / EEG pour chaque point de temps séparément. Ils considèrent une régularité spatiale, mais ils ne prennent pas en compte la dimension temporelle des données MEG / EEG. De plus, ce sont des modèles denses, qui ne correspondent pas à l'hypothèse à laquelle on s'intéresse qui est la parcimonie. En soit, on cherche que quelques régions cérébrales focales qui sont impliquées dans une tâche cognitive spécifique. MNE ou dSPM [DLF+00], par exemple, auront tous les deux des sources différentes de zéro à chaque instant. Pour ce faire, plusieurs méthodes favorisant des configurations de sources focales parcimonieuses ont été proposées sur la base d'un relâchement de la norme 0. Une approximation populaire est ℓ_p-normes avec $0 < p \leq 1$:

$$\|\mathbf{X}\|_p = \Big(\sum_{s=1}^{S} \sum_{t=1}^{T} |\mathbf{X}[s,t]|^p \Big)^{\frac{1}{p}} \tag{1}$$

En supposant des données MEG / EEG blanchies spatialement, une estimation de source parcimonieuse peut être obtenue en résolvant le problème régularisé:

$$\mathbf{X}^\star = \underset{\mathbf{X}\in\Re^{SO\times T}}{\arg\min} \frac{1}{2}\|\mathbf{M}-\mathbf{G}\mathbf{X}\|_2^2 + \lambda\|\mathbf{X}\|p \quad \text{with} \quad \lambda > 0. \tag{2}$$

Le problème d'optimisation est non différentiable et n'a pas de solution directe. Par conséquent, des approches itératives doivent être appliquées pour résoudre le problème dans l'équation (2). Fixant p = 1 dans l'équation (2), le problème est connu sous le nom de Lasso [Tib96] dans les statistiques, Basis Pursuit Denoising (BPDN) [CDS98] en traitement du signal, et Norme minimale sélective (SMN) [MO95] en MEG / EEG. Cependant, l'application de la norme ℓ_1 pour l'orientation libre (O = 3) favorise la parcimonie même dans l'orientation à chaque emplacement source. Pour surmonter ce problème, MCE (*Minimum Current Estimate*) a été proposée par [UHS99] résolvant SMN en fixant l'orientation a priori en calculant d'abord une solution MNE.

Plusieurs autres approches peuvent être citées qui étudient la même idée d'avoir des estimations de sources focales. *Sparse Bayesian Learning (SBL)* (SBL) [Wip06], *Spatio-Temporal TOmographic NonNegative Independent Component Analysis* (STTONNICA) [VSVHSB+09], Normes Mixtes [OHG09], Champagne [OWA+12], Inférence Bayésienne Hiérarchique [LPBW12], ou les estimations de la norme mixte (MxNE) [GKH12]. Ces méthodes sont appelées spatio-temporelles car elles fonctionnent dans une fenêtre temporelle prédéfinie, mais elles ignorent complètement la corrélation temporelle. Cela peut être vérifié en déplaçant les colonnes de l'estimation de source: cela n'aura aucun effet sur l'estimation de la source elle-même.

Pour introduire une "vraie" contrainte spatio-temporelle dans le modèle, [ZR11b, ZR11a] a incorporé la corrélation temporelle pour améliorer les estimations de sources, *vector-based spatiotemporal minimum* ℓ_1-norme (VESTAL) [HDS+06] applique une projection temporelle pour réduire la sensibilité au bruit après l'utilisation de la norme ℓ_1. Le *Fast-VESTAL* [HHR+14] est une sorte de post-traitement à la méthode VESTAL. Cependant, cette étape de post-traitement suppose implicitement que les estimations de la source sont stationnaires. Pour surmonter ces problèmes, Ou et al. [OHG09] a proposé une approche pour reconstruire simultanément plusieurs instants temporels en appliquant la norme mixte $\ell_{2,1}$ pour imposer la régularité spatio-temporelle de groupe en tant que régularisation [GKH12, OHG09].

Time-Frequency Mixed-Norm Estimate (TF-MxNE) [GSH+13] a réutilisé la norme mixed $\ell_{2,1}$ (MxNE [GKH12]) dans le domaine temps-fréquence en ajoutant une seconde régularisation sur le temps $\ell_{2,1}+\ell_1$. Il multiplie la matrice de gain par un dictionnaire de fonctions de base spatiales. Ils obtiennent une matrice de gain modifiée, qui peut être utilisée pour estimer des sources spatialement étendues avec des formes d'onde temporellement lisses. Cette approche a également été étudiée par [CCHMV+15], en appelant la méthode Spatio-Temporary Unifying Tomography (STOUT).

Bien que ces méthodes spatiotemporelles améliorent la reconstruction de sources MEG / EEG, elles sont basées sur des pénalités convexes. Cela permet des algorithmes rapides avec une convergence globale garantie. Cependant, les estimations de source résultantes sont biaisées en amplitude et souvent sous-optimales en termes de récupération de support, c'est-à-dire de sources actives [CWB08]. Comme montré par example dans le domaine de *compressed sensing*, la promotion de la parcimonie en appliquant des pénalités non convexes, telles que des pénalités logarithmiques ou ℓ_p-quasi-normes avec $0 < p < 1$, améliore la reconstruction du support en termes de sélection, biais d'amplitude et stabilité [CWB08, Cha07, SCY08]. Plusieurs approches pour résoudre le problème d'optimisation non-convexe ont été proposées, y compris l'optimisation itérative ℓ_1-norm [CWB08]. [SHG14] a utilisé une approche itérative repondérée pour résoudre la pénalité composite non-convexe dans le domaine temps-fréquence.

3 Localisation de sources avec des dictionnaires multi-échelles

Cette section considère le problème variationnel dans le domaine temps-fréquence en fixant le terme de pénalisation comme un *Sparse group lasso* comme suit:

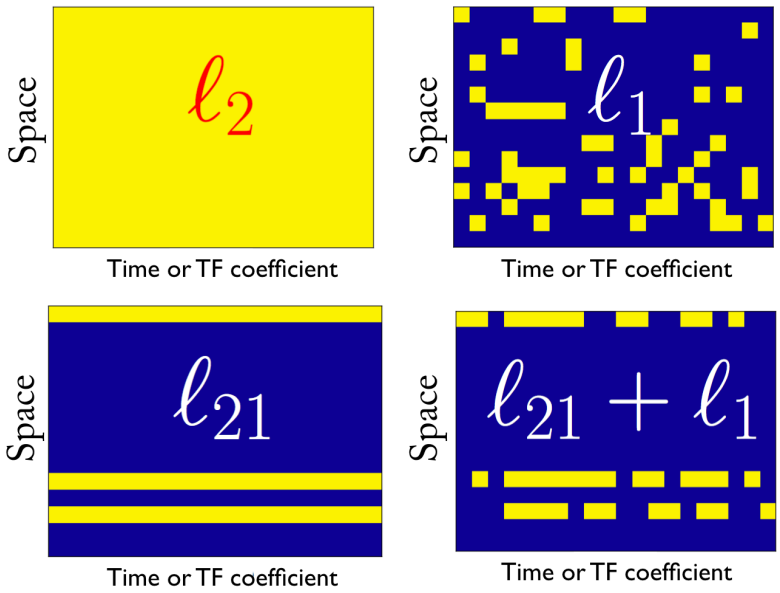


Figure 2: Modèles de parcimonie selon les différentes régularisations: ℓ_2 tous non-zéro, ℓ_1 dispersés et non structurés valeurs de non-zéro, $\ell_{2,1}$ structure de rangée de bloc, et $\ell_{2,1} + \ell_1$ (domaine TF) structure de rangée de bloc avec une parcimonie intra-rangée. La couleur jaune indique des coefficients non nuls.

Sparse Group lasso

$$\mathbf{X}^\star = \underset{\mathbf{X}\in\Re^{S\times T}}{\arg\min} \frac{1}{2}\|\mathbf{M} - \mathbf{G}\mathbf{X}\|_{Fro}^2 + \lambda_1\|\mathbf{X}\|_{2,1} + \lambda_2\|\mathbf{X}\|_1 \tag{3}$$

Si $\lambda_1 = 0$, c'est alors égal à la pénalité de lasso, et si $\lambda_2 = 0$, ça résulte en une pénalité de Group lasso.

λ_1 est un hyperparamètre sur l'espace et λ_2, un second hyperparamètre sur le temps. La figure 2 justifie ce choix. Elle montre comment $\ell_{2,1} + \ell_1$ permet de modéliser des sources non stationnaires qui ne peuvent pas être estimées avec ℓ_2 ou $\ell_{2,1}$ en raison de la norme non-parcimonieuse ℓ_2 favorisant le temps, tandis que l'estimation ℓ_1 est complètement dispersée et non structurée.

Cette section décrit la localisation de sources dans le domaine TF. Nous avons montré que la localisation de sources dans le domaine TF était une «vraie» approche spatio-temporelle prenant en compte la corrélation temporelle. Ces méthodes améliorent la reconstruction de sources transitoires et non stationnaires en promouvant la parcimonie structurée dans le domaine TF. Ces méthodes appliquent un *Sparse Group lasso* 3 sur les coefficients TF. TF-MxNE et STOUT appliquent une pénalité composite convexe, la somme d'une pénalité $\ell_{2,1}$ et d'une pénalité ℓ_1, sur la transformée de Gabor des séries temporelles. D'autre part, irTF-MxNE applique une pénalité composite non-convexe, la somme d'une pénalité de $\ell_{2;0.5}$-quasinorme et d'une pénalité de $\ell_{0.5}$-quasinorme sur les coefficients TF. Il a été démontré que les pénalités non convexes surpassent les approches convexes en termes de récupération de source et de biais d'amplitude [CWB08, Dea10]. Cependant, le choix d'un dictionnaire Gabor optimal pour décomposer les données reste difficile.

Le problème du choix du dictionnaire est spécialement rencontrée quand un mélange de signaux est disponible dans les données, par example, un court signal transitoire juste après le stimulus, et des ondes cérébrales plus lentes par la suite. Le choix d'un dictionnaire unique décrivant les deux signaux de manière parcimonieuse est difficile. Cette section montre les résultats obtenus après avoir incorporer un dictionnaire multi-échelle dans l'algorithme d'optimisation itératif repondéré, c'est-à-dire plusieurs dictionnaires avec différentes échelles concaténées pour s'adapter à des transitoires courts et des ondes lentes en même temps, tout en gardant une efficacité computationnelle. Le problème d'optimisation est résolu de la même façon que irTF-MxNE [SGH15], c'est-à-dire

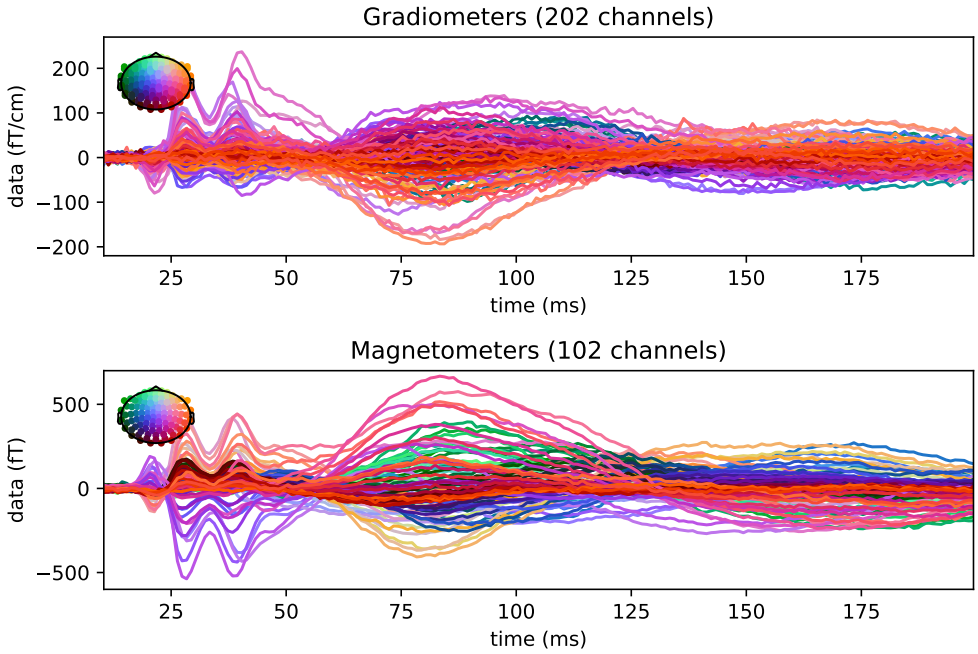


Figure 3: Réponse évoquée somatosensorielle après prétraitement et moyennage (données des gradiomètres et des magnétomètres). Le cercle en haut à gauche donne la position des capteurs sur la tête qui sont codés par couleur.

que chaque itération est une TF-MxNE pondérée, que nous résolvons en utilisant *Block coordinate descent* (BCD) [Tse10] et une stratégie d'ensemble actif [FHT10]. Nous démontrons l'intérêt du dictionnaire multi-échelle en termes de reconstruction de séries temporelles des sources et de démixage temporel des activations.

3.1 Résultats

Afin de démontrer l'intérêt de irTF-MxNE avec un dictionnaire multi-échelle comparé à irTF-MxNE de base, nous avons testé différents paramètres pour les différents solveurs sur un ensemble de données MEG: étude somatosensorielle du jeu de données MIND (voir détails dans [WHM+07]). La réponse évoquée est illustrée dans la Figure 3. On peut déjà remarquer ce mélange d'ondes cérébrales dans l'évoqué. Les vagues plus nettes juste après le début sont dues à un bon alignement des essais dont les informations ne sont pas perdues en prenant la moyenne. Ceci est principalement connu comme une réponse de la zone somatosensorielle primaire (S1) qui répond rapidement après une stimulation électrique indolore du nerf médian. Une vague plus longue qui vient plus tard autour de 70ms est clairement vue dans l'évoqué aussi. C'est ce qui fait de ces données un jeu de données difficile mais très bien pour tester le solveur multi-échelle.

L'estimation de la source a d'abord été réalisée en utilisant plusieurs solveurs: irTF-MxNE, ir-MxNE [SHG14] et dSPM [DLF+00]. En ce qui concerne irTF-MxNE, deux dictionnaires ont été testés (les deux dictionnaires STFT). Un dictionnaire avec une fenêtre de 64 échantillons et un décalage temporel de 4 échantillons, ce qui conduit à des séries temporelles lisses; et un dictionnaire avec une fenêtre de 16 échantillons et un décalage temporel de 2 échantillons, ce qui permet de capturer des sources transitoires courtes. Après l'inspection du résidu Dans la Figure 4, les résultats montrent qu'au moins quatre sources sont nécessaires pour capturer tous les composants évoqués.

Nous avons donc fixé les paramètres des solveurs irTF-MxNE pour obtenir seulement quatre sources tout en expliquant autant de variance que possible. Après cela, nous avons expérimenté avec un ensemble de paramètres différents et nous montrons deux d'entre eux, $\lambda_{time} = \lambda_2 = 1.5$ et $\lambda_{time} = \lambda_2 = 2.5$, pour démontrer leur impact sur la smoothness des différentes sources obtenues. Les paramètres ont été choisis de manière à réduire le résidu, c'est-à-dire à maximiser les données expliquées en ayant au moins quatre sources. Les figures 5 (a-b) représentent les quatre séries temporels obtenus avec irTF-MxNE en utilisant le dictionnaire de fenêtre courte pour les valeurs

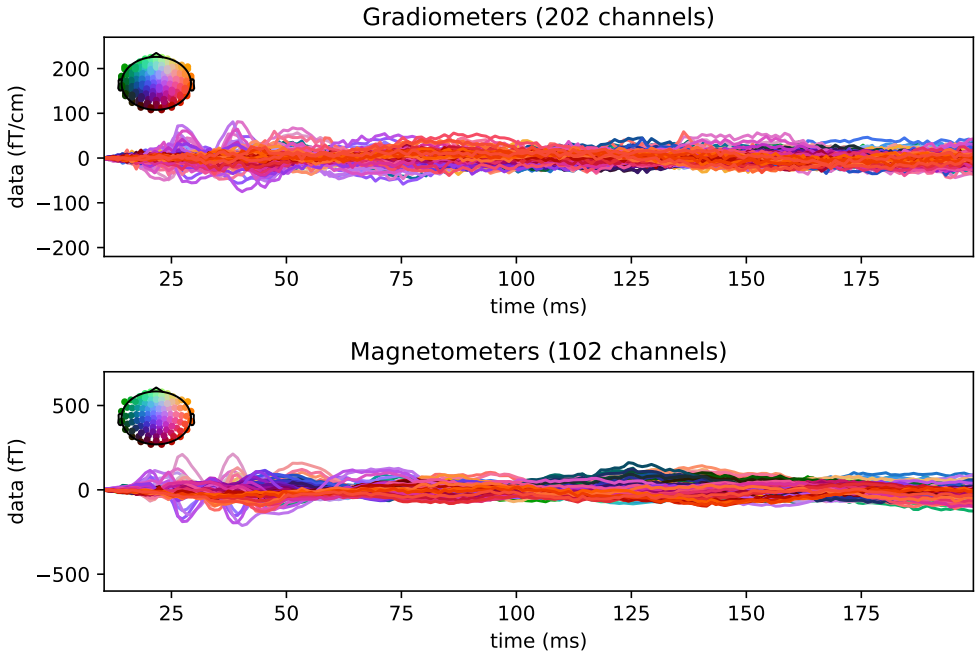


Figure 4: Résiduelle des données somatosensorielles après application de irTF-MxNE multi-échelle. Le cercle en haut à gauche donne la position des capteurs sur la tête qui sont codés par couleur

de λ_{time} sélectionnées.

Nous montrons que pour des valeurs élevées de λ_{time}(b), le solveur n'est pas capable de capturer la composante transitoire courte autour de 30 ms. Alors que pour une petite valeur de λ_{time} (a), le démixage n'est pas fiable puisque les estimations de la source bleue et de la source verte captent l'activité de la source rouge. De plus, les séries temporelles ne sont pas lisses. D'autre part, les figures 5 (c-d) représentent les quatre séries temporelles obtenues avec irTF-MxNE en utilisant le dictionnaire de fenêtre longue. La figure confirme que les deux paramètres ne sont pas capables de capturer l'effet transitoire après le stimulus, bien que les trajectoires temporelles soient lisses. Ces quatre sous-figures révèlent qu'il faut une courte fenêtre pour capturer l'effet transitoire du signal cérébral, alors qu'il doit avoir une longue fenêtre pour capturer les ondes longues et avoir des estimations de sources lisses. Ce résultat démontre comment une combinaison des deux dictionnaires est essentielle pour acquérir des estimations de sources avec une grande précision, mais les hyperparamètres doivent également être ajustés, comme le montre la figure 5, que leurs valeurs changent radicalement les résultats.

De plus, la Figure 5 (e) montre les amplitudes obtenues avec MxNE pour cinq sources. Quant à MxNE, il n'est pas possible d'obtenir les quatre sources pertinentes non mélangées (voir pour d'autres figures démonstratives [GKH12]). Nous remarquons que la source bleue clair des figures 5 (a) à (d) apparaît comme deux sources distinctes en (e): bleu clair et violet. Si nous augmentons le paramètre λ, nous augmentons le biais d'amplitude dû à la norme ℓ_1 du solveur. Si nous le fixons trop haut ($\lambda = 50$) nous obtenons quatre sources, mais la source bleue qui est pertinente pour l'étude serait supprimée et la source violette dupliquée serait conservée. Le dernier panneau Figure 5 (f) affiche les estimations de la source pour les valeurs dSPM correspondant aux quatre emplacements des sources obtenues avec irTF-MxNE. Ces sous-figures montrent qu'aucun des solveurs MxNE ou dSPM n'est capable d'obtenir des sources lisses sans aucune fuite entre les séries temporelles.

L'estimation de source a ensuite été obtenue en utilisant irTF-MxNE avec la combinaison des deux dictionnaires. La figure 6 montre la reconstruction de sources en utilisant le multi-échelle irTF-MxNE pour les paramètres de régularisation $\lambda_{space} = 28.5$ et $\lambda_{time} = 1.5$. L'emplacement de chaque source est marqué par une sphère dans la figure 6-gauche, et son amplitude dans le temps est codée en couleur dans le panneau de droite. Les résultats montrent une succession appropriée des sources. La source transitoire (rouge) est la seule source expliquant le champ lié à l'événement jusqu'à 48 ms. Cette source rouge correspond au cortex somatosensoriel primaire controlatéral

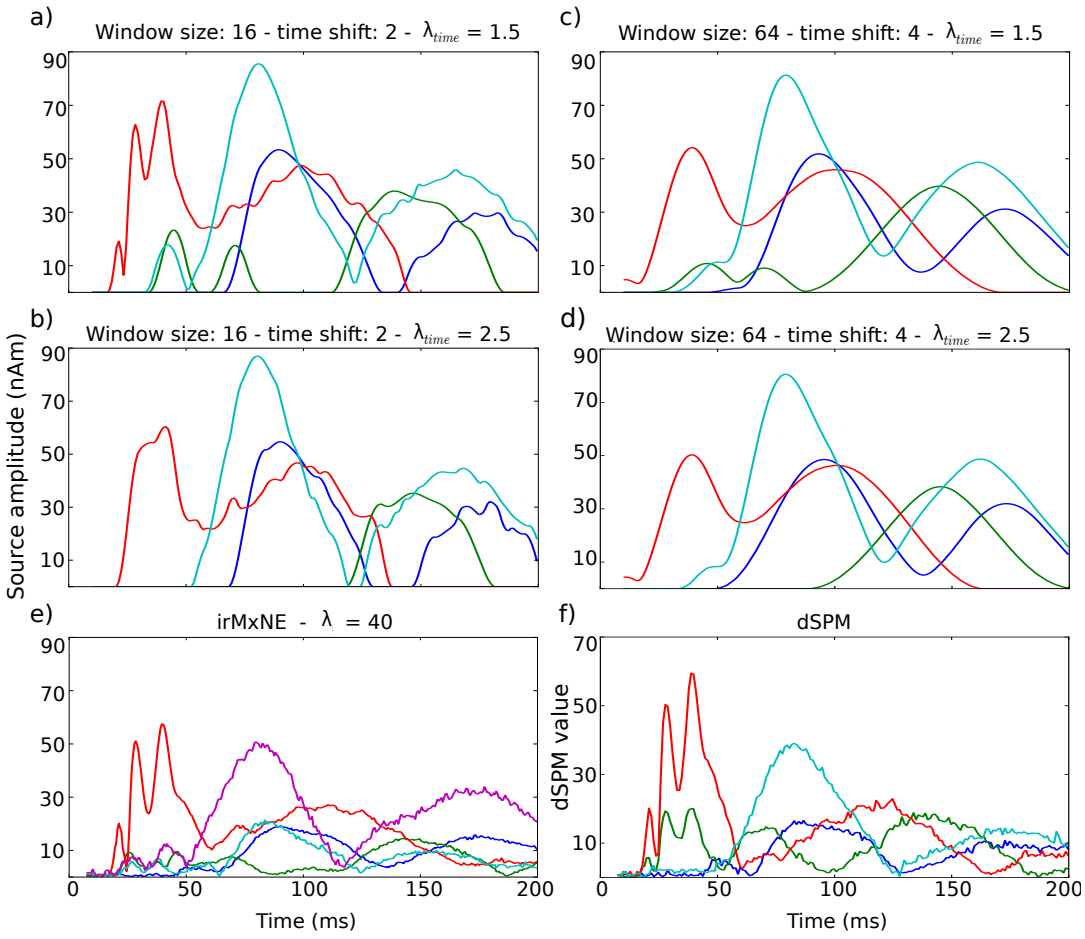


Figure 5: Reconstruction de source en utilisant des données somatosensorielles avec différents solveurs. (a) - (b) irTF-MxNE sur un dictionnaire de petite fenêtre avec $\lambda_{time} = 1.5$ et $\lambda_{time} = 2.5$ respectivement. (c) - (d) irTF-MxNE avec un dictionnaire de longue fenêtre avec $\lambda_{time} = 1.5$ et $\lambda_{time} = 2.5$ respectivement. De (a) à (d) $\lambda_{space} = \lambda_1 = 28.5$ (e) MxNE pour $\lambda = 40$ et (f) activation dSPM pour les quatre sources activées.

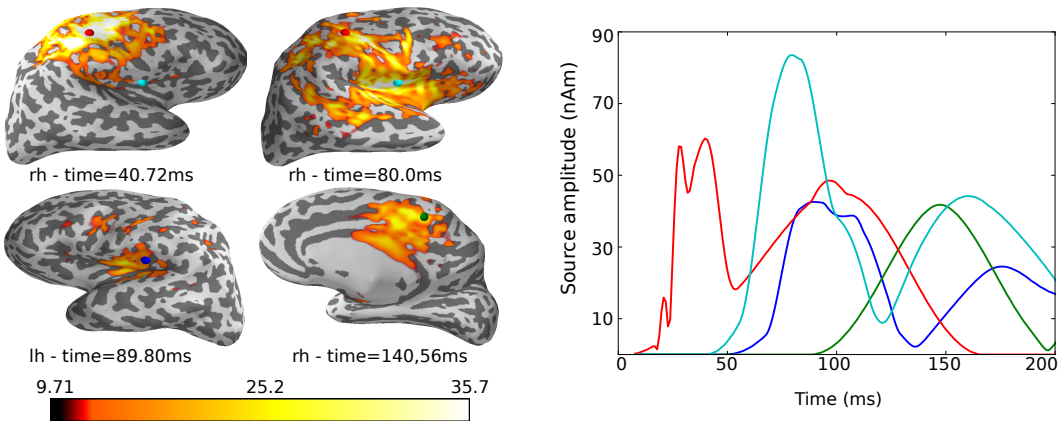


Figure 6: Reconstruction de sources en utilisant des données somatosensorielles avec irTF-MxNE multi-échelle. Le solveur estime quatre sources pour $\lambda_{space} = 28.5$ et $\lambda_{temps} = 1.5$. Les emplacements sources marqués avec des sphères dans l'hémisphère droit (droite) et gauche (gauche) et leur activation correspondante sont codés par couleur. La barre de couleurs est sur les valeurs dSPM qui n'a pas d'unités car ce sont des valeurs statistiques.

(cS1) situé dans le gyrus postcentral du lobe pariétal (hémisphère droit (rh)). La sphère rouge sur la vue latérale coïncide avec l'activation dSPM étalée autour de 40 ms. La deuxième source (bleu clair) correspond au cortex somatosensoriel secondaire (cS2), et se produit également avec l'activation dSPM autour de 80 ms. Environ 100 ms après le stimulus, des sources corticales supplémentaires sont activées, comme le cortex somatosensoriel secondaire ipsilatéral (iS2) (bleu-lh) et la paroi médiane controlatérale (vert-rh).

Les avantages de l'irTF-MxNE multi-échelle ont été démontrés sur des données somatosensorielles. Ces expériences confirment que l'irTF-MxNE multi-échelles améliore les estimations de sources, en termes de réduction de mélange, de régularité et de détection des ondes transitoires courtes et des ondes plus lentes. L'amélioration de la taille de l'ensemble actif et de l'amplitude est due à la non-convexité des méthodes de régularisation. Par conséquent, l'irTF-MxNE multi-échelles devrait être appliqué aux données impliquant un mélange de signaux, et lorsque le but est d'acquérir des sources focales avec des trajectoires temporelles non stationnaires et lisses.

4 Lien entre les approches déterministe et bayésienne

Cette section vise à établir un lien entre les approches déterministe et bayésienne exploitant un a priori parcimonieux dans le but de proposer une méthode hybride plus performante. Plus particulièrement, ça montre comment un modèle bayésien hiérarchique utilisant un échantillonneur MCMC peut reproduire à l'identique la solution d'un *lasso* adaptatif utilisant le principe de Maximisation-Minimisation (MM). Cette équivalence établie montre l'apport des approches stochastiques qui au prix d'un coût de calcul plus conséquent permettent d'estimer non seulement les hyperparamètres mais de mesurer également le degré d'incertitude des solutions calculées. Ceci consiste à initialiser l'approche déterministe MM à l'aide d'un échantillonneur MCMC de la distribution a posteriori. Une carte d'incertitude des différentes solutions obtenues est alors dérivée. Cette section décrit une des contributions méthodologiques majeures de cette thèse. Une estimation bayésienne de l'hyperparamètre d'un group-lasso, exploitant comme modèle la loi Gamma, est par ailleurs revisitée montrant l'intérêt d'estimer automatiquement l'hyperparamètre.

4.1 Estimation automatique de l'hyperparamètre

Cette section étudie l'estimation de l'hyperparamètre dans la formulation variationnelle. On peut remarquer que le réglage de l'hyperparamètre est un problème de statistique classique pour lequel un certain nombre de solutions ont été proposées. Dans le traitement du signal, le critère d'information Akaike (AIC) et le critère d'information bayésien (BIC) sont des techniques assez populaires historiquement [S+78]. Les techniques basées sur SURE [Ste81] ont également été très

populaires et explorées récemment pour les applications de débruitage et des applications de *compressed sensing* [LBU07, GD15]. Dans une configuration d'apprentissage automatique supervisée standard avec des observations indépendantes et identiquement distribuées (i.i.d.), la validation croisée (CV) est l'approche de référence. De plus, l'approche bayésienne adaptée aux modèles probabilistes offre une méthode raisonnée pour estimer les hyperparamètres utilisant des hyperpriors qui introduisent des contraintes plus faibles que les solutions avec des valeurs de paramètres fixes. Cet avantage a généralement un prix en termes de coût de calcul. Enfin, dans un certain nombre de scénarios réels, les humains finissent par définir des hyperparamètres, car ils peuvent avoir des connaissances d'experts qui peuvent corriger le décalge du modèle.

Dans l'apprentissage automatique, un hyperparamètre vise généralement à limiter le surajustement en contrôlant la complexité du modèle. Dans le cas particulier de la régression régularisée, classiquement un paramètre scalaire équilibre l'ajustement des données et le terme de la pénalité. Lorsque vous utilisez une régression parcimonieuse, ce paramètre affecte la parcimonie de la solution, c'est-à-dire le nombre de covariables ou de régresseurs utilisés.

Cette thèse est particulièrement intéressée par le paramètre de régression de haute dimension en utilisant la parcimonie structurée de type Group-Lasso comme vu jusqu'ici. Dans la littérature, un certain nombre d'approches ont été proposées et les estimations du MAP qui se résument à une régression pénalisée sont les approches standard utilisées par les neuroscientifiques [HNZ+08, OHG09, BVVN09, WN09, GKH12, LPBW12, VSVHSB+09].

Dans une formulation variationnelle, la valeur de l'hyperparamètre dépend du problème posé, du niveau de bruit et du choix de la régularisation $\mathcal{P}(\mathbf{X})$. Trouver un moyen d'estimer l'hyperparamètre avec une intervention minimale de l'utilisateur est donc particulièrement important, car cela facilite la comparaison entre différents modèles et régularisation.

Récemment, Pereyra et al. [PBDF15] a proposé une stratégie pour l'estimation de l'hyperparamètre dans le contexte de l'inférence MAP lorsque le prior ou la régularisation est une fonction k-homogène. La régularisation $\mathcal{P}$ est une fonction k-homogène s'il existe $k \in \Re^+$ tel que:

$$\mathcal{P}(\eta\mathbf{X}) = \eta^k \mathcal{P}(\mathbf{X}), \quad \forall \mathbf{X} \in \Re^{S \times T} \quad \text{and} \quad \forall \eta > 0.$$

La condition k-homogène est satisfaite pour toutes les normes mixtes $\ell_{p,q}$. Nous nous concentrons sur l'estimation des hyperparamètres pour les modèles bayésiens hiérarchiques produisant des pénalités convexes $\ell_{2,1}$ ($\mathcal{P}(\mathbf{X}) = \|\mathbf{X}\|_{2,1}$) ou non convexes $\ell_{2,0.5}$, qui sont respectivement 1-homogène et 0.5-homogène. La pénalisation non-convexe est résolue en utilisant des schémas d'optimisation convexes re-pondérés itératifs, c'est-à-dire que chaque itération est une norme $\ell_{2,1}$ pondérée.

Dans [PBDF15], la stratégie de point fixe proposée est validée sur un problème de débruitage d'image en utilisant un prior d'analyse, c'est-à-dire où la solution n'est pas parcimonieuse mais a une représentation parcimonieuse dans un domaine transformé. Cette section montre les résultats obtenus après adaptation de la méthode ayant un prior de synthèse pour un problème sous-déterminé. Un prior de synthèse est lorsque la solution elle-même est parcimonieuse.

4.1.1 Résultats

Nous avons généré un jeu de données pour simulation avec $N = 302$ capteurs, $T = 190$ échantillons de temps et $S = 1500$ sources. Quatre sources ont été sélectionnées au hasard pour être actives avec des formes d'onde réalistes obtenues à partir de l'ensemble de données MIND [WHM+07]. L'opérateur direct $\mathbf{G}$ était une matrice aléatoire, dont les colonnes ont été normalisées à 1. Deux niveaux de bruit blanc ont été ajoutés à la simulation.

La figure 7 représente les sources simulées avec des étoiles et les sources estimées avec des lignes simples. La figure 7- (a) - (b) affiche les résultats avec les normes $\ell_{2,1}$ et $\ell_{2,0.5}$ respectivement, en utilisant un hyperparamètre initialisé à $\lambda = 0.5\lambda_{max}$. On peut voir que dans la figure 7- (a), la norme $\ell_{2,1}$ récupère les quatre sources avec un biais d'amplitude (l'amplitude estimée est inférieure à l'amplitude exacte), et que plusieurs sources montrées en vert clair sont presque plat autour de zéro mais toujours trouvé en tant que sources actives. Il n'y a aucun moyen de réduire le support sans perdre l'une des quatre sources simulées, c'est-à-dire que la norme $\ell_{2,1}$ avec un hyperparamètre ne parvient pas à récupérer les sources simulées exactes. La norme $\ell_{2,0.5}$ de (b) estime les quatre amplitudes de source sans biais d'amplitude grâce à la non-convexité [SBHG16]. D'autre part, la figure 7 (c) montre les résultats de la pénalité convexe en utilisant un hyperparamètre par source. On peut voir qu'il est qualitativement équivalent à la pénalité non-convexe.

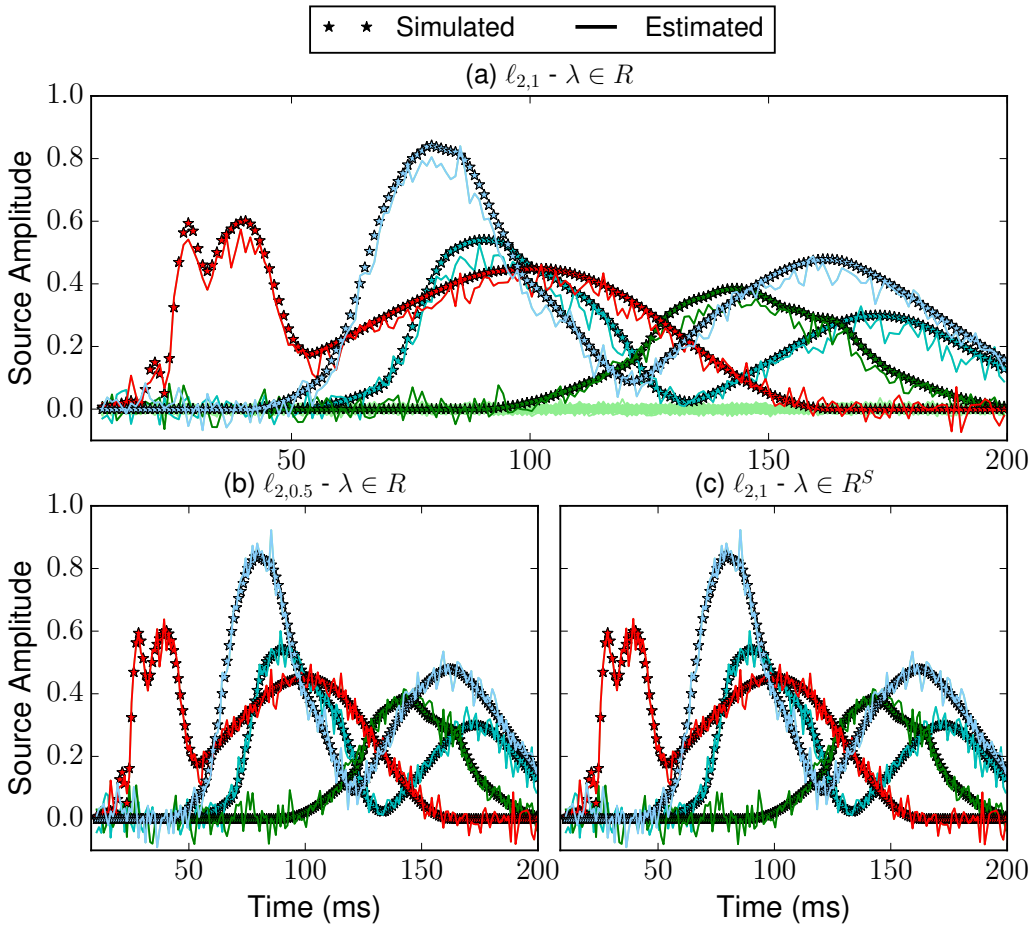


Figure 7: Reconstruction de sources sur des données simulées. (a): Estimations de sources obtenues en utilisant $\ell_{2,1}$ avec un λ. La solution n'est pas assez parcimonieuse (sources nulles en vert clair) et il existe un biais d'amplitude entre les amplitudes exactes (étoiles) et les estimations (lignes brutes). (b): Bonne reconstruction des quatre sources en utilisant $\ell_{2,0.5}$ et un λ, ce qui est équivalent à la reconstruction utilisant la norme $\ell_{2,1}$ avec $\lambda \in \Re^S$ (c). Chacune des quatre sources est codée avec une couleur différente.

L'avantage d'avoir un hyperparamètre par source est de ne prélever que les sources impliquées dans les données $\mathbf{M}$ et de laisser tomber les sources quasi-nulles supplémentaires visibles sur la figure 7- (a) (vert clair). Cette extension produit des résultats plus parximonieux et moins de biais d'amplitude sans que le problème soit non-convexe. Cette figure suggère également un lien entre le prior non-convexe et un hyperparamètre par source. Comme le prior non convexe est une procédure itérative estimant un poids interne pour produire une meilleure solution, le fait d'avoir un hyperparamètre par source peut également être considéré comme un poids pour définir de meilleures estimations de sources. Une étude plus précise de ceci est donnée dans la prochaine section.

4.2 Lien entre MM et HBM

Cette section dans la version longue de ma thèse montre comment dériver une paramétrisation du MM et HBM pour que les deux techniques soient totalement équivalentes. Ceci nouts aide à prendre avantages de chaque technique: l'optimisation du MM, et la connaissance de la distribution posterieur dans une formulation Bayésienne.

Du travail précédent [SBHG16], nous savons qu'en raison de la non-convexité, une bonne initialisation des poids $\mathbf{W}[0]$ dans l'algorithme MM est cruciale pour sa performance, mais mis à part une initialisation uniforme, seulement des stratégies heuristiques d'initialisation ont été utilisées, par exemple en utilisant la même repondération comme dans la méthode sLORETA [PM02a]. Dans

cette thèse, nous tirons parti de la réinterprétation de l'algorithme MM à travers le cadre HBM pour obtenir des initialisations multiples de manière systématique, à savoir des échantillons tirés du postérieur complet. De cette façon, nous pouvons non seulement atteindre de meilleurs minima locaux, mais plus important encore, nous pouvons identifier et caractériser plusieurs solutions parcimonieuses possibles. De telles solutions plausibles au problème de régression sparse sont les modes de la distribution a posteriori avec différentes masses de probabilité relative.

4.2.1 Résultats

Nous avons généré une simulation réaliste basée sur un modèle de source d'orientation libre ($O = 3$) avec $S = 7498$ emplacements corticaux et $m = 306$ capteurs MEG. Deux de ces endroits ont été sélectionnés pour être actifs, un dans chaque hémisphère. L'une des sources avait une localisation ventrale profonde dans le gyrus occipital inférieur (figure 4.7.1-c), et la seconde avait une localisation plus superficielle dans le cortex moteur (figure 4.7.1-a). Leurs formes d'onde correspondantes sont illustrées à la Figure 4.7.1-b. Lorsqu'ils sont passés aux solveurs, ils sont recadrés entre 40 et 180 ms pour ne conserver que les deux pics. Cela conduit à $T = 43$ échantillons de temps.

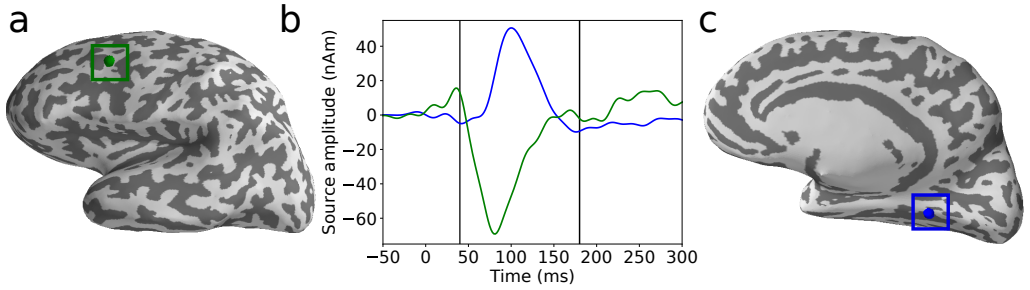


Figure 8: Jeu de données MEG simulé. a) et c) montrent des sources superficielle et profonde (cachées dans la vue médiane), respectivement. b) donne leurs formes d'onde correspondantes codées par couleur par emplacement.

Trouver le bon support dans un problème de régression sous-déterminé parcimonieuse est d'une complexité combinatoire. Dans nos deux approches, cela se reflète dans la non-convexité de la fonction objective et dans la multi-modalité de la distribution postérieure conjointe, respectivement. La deuxième question que nous voulons étudier est de savoir si les méthodes que nous avons développées ici peuvent révéler ou quantifier une partie de l'ambiguïté et de l'incertitude de ce problème d'identification de support parcimonieux. Les mesures traditionnelles de quantification de l'incertitude telles que les estimations de la variance de $\mathbf{X}$ ne parviennent pas à le faire, car elles ne peuvent pas capturer de manière satisfaisante la multimodalité du postérieur. De plus, aucun échantillon $\mathbf{X}^{(k)}$ n'est exactement parcimonieux: la distribution postérieure étant une densité de probabilité continue, la probabilité de l'événement $\mathbf{X}^{(k)}[i] = 0$ est nulle, ce qui signifie que tout le support de $\mathbf{X}^{(k)}$ est active avec la probabilité 1. Même une moyenne seuillée du support de $\mathbf{X}^{(k)}$ ne révélera que la probabilité moyenne d'un emplacement faisant partie du support. Dans l'analyse des sources, il est sans doute plus intéressant d'estimer quels réseaux de sources provenant de différentes régions du cerveau ont probablement produit un ensemble de données, question laissée ouverte par ces mesures. Ici, nous proposons d'aborder cette question d'une manière différente.

Notre procédure d'initialisation d'une itération MM avec un échantillon de la distribution postérieure donne différents minima locaux, c'est-à-dire des solutions approximatives au problème qui satisfont notre connaissance a priori d'un support parcimonieux. Si nous supposons que la division de $\Re^{SO\times T}$ en attracteurs de l'algorithme MM chevauche grossièrement à la division de $\Re^{SO\times T}$ en modes du postérieure marginalisée sur $\mathbf{X}$ dans le cadre HBM, cette fréquence relative correspond au volume relatif des minima locaux. Ce dernier est une meilleure mesure pour comparer différents minima locaux que leur profondeur (un minimum local qui est profond mais mince correspond à une estimation de source instable). Alors qu'une analyse mathématique plus profonde et plus détaillée de cette heuristique est laissée pour un travail futur, nous examinons ici si cette approche réagit aux changements dans le plan de mesure de la manière que nous attendrions. Pour ce faire, nous passons de l'utilisation de tous les 306 capteurs MEG à l'utilisation de seulement 204 gradiomètres ou un gradiomètre de plus de deux (102 capteurs). En réduisant le nombre de

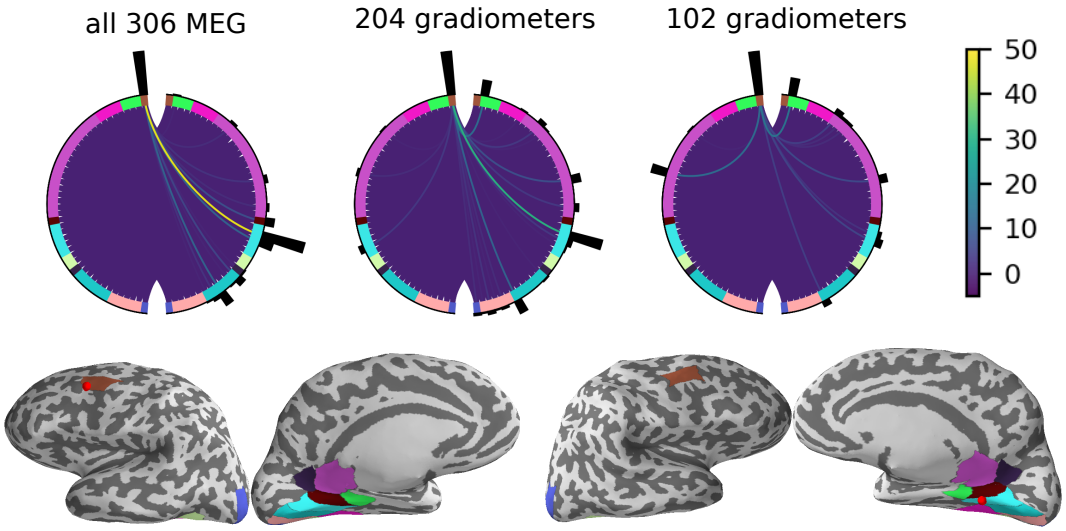


Figure 9: Analyse du réseau source pour les données simulées. La première rangée de sous-figures montre les 900 minima locaux de la manière suivante: chaque position dans le cercle représente un emplacement source qui faisait partie du support d'au moins un minimum pour une configuration de capteur. La barre noire attachée à chaque position correspond à la fréquence relative avec laquelle cet emplacement source est apparu comme faisant partie du support. Deux positions sont reliées par une ligne si elles faisaient simultanément partie du support, et la couleur de cette ligne correspond à la fréquence relative avec laquelle cela s'est produit. Notez que l'arrière-plan du cercle est blanc, mais densément couvert par des lignes violettes indiquant des connexions rares. Les positions sont placées à gauche ou à droite selon l'hémisphère auquel elles appartiennent. Pour la symétrie, pour chaque source actif, sa contrepartie sur l'autre hémisphère était également incluse dans le graphique. De plus, les positions sont regroupées et colorées en fonction d'une parcellisation du cerveau en régions anatomiques (issues d'un atlas). La deuxième rangée de sous-figures montre ces régions dans le cerveau et les sources simulées.

capteurs, nous augmentons la sous-détermination du problème et l'intuition est que cela devrait conduire à plus de variabilité parmi les solutions plausibles et parcimonieuses.

L'analyse graphique présentée et décrite dans la figure 9 et dans la figure 10 le confirme. Une première observation est que la source superficielle dans le cortex prémoteur a été correctement identifiée comme faisant partie du support de tous les minima locaux lors de l'utilisation des 306 capteurs MEG complets. Il était cependant parfois mal localisé lorsqu'il réduisait le nombre de capteurs (Figure 9). Une deuxième observation est que la propagation spatiale de ces localisations manquantes est plus petite pour cette source superficielle que pour la source profonde. Cette source profonde dans le cortex ventrale est plus difficile à trouver même avec tous les capteurs. En effet, aucune des 900 initialisations n'a parfaitement localisé la source profonde simulée. En général, nous pouvons clairement voir comment l'ambiguïté augmente quand on diminue le nombre de capteurs, et comment la distribution des réseaux devient plus floue.

Lorsque l'on considère des données réelles, la source à récupérer est souvent mal comprise, en particulier lorsqu'il s'agit d'une activité cérébrale pathologique telle qu'une activité épileptique ictale ou inter-ictale. Dans une telle situation, fournir une configuration de source unique en conséquence, avec une quantification d'incertitude ad hoc basée sur des études antérieures ou une expertise acquise, pourrait ne pas être une utilisation optimale des données MEG / EEG. Au lieu de cela, fournir plusieurs hypothèses ensemble, avec une quantification de leur incertitude, peut être plus utile. En effet, pour des applications telles que le diagnostic d'épilepsie pré-chirurgicale, où les enregistrements MEG / EEG sont l'une des modalités diagnostiques, chaque configuration de source candidate peut fournir des preuves pour ou contre une hypothèse diagnostique pouvant mener à une décision chirurgicale. Nous croyons donc que l'extension des premières étapes que nous avons prises ici pour développer un cadre cohérent pour interpréter et quantifier la multitude de résultats potentiels des approches de reconstruction de sources MEG / EEG parcimonieuses peut avoir un impact significatif sur les applications cliniques.

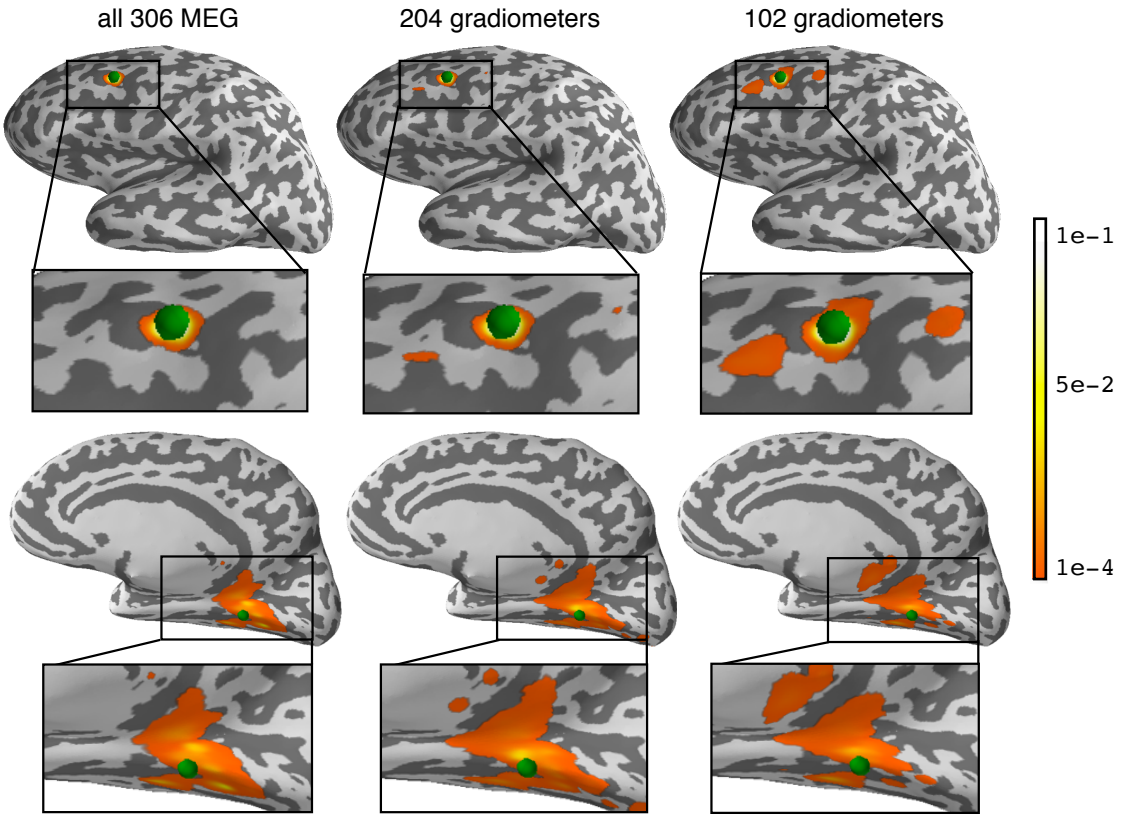


Figure 10: Le support des résultats MM basé sur 900 initialisations venant du MCMC a été extrait pour construire une carte d'incertitude. Les fréquences relatives avec lesquelles chaque source faisait partie du support ont été calculées et tracées sur la surface du cerveau avec les deux sources simulées (points verts). Chaque colonne correspond aux résultats de chacune des trois configurations de capteurs examinées. Moins le nombre de capteurs et / ou plus la source est profonde, plus la carte du cerveau est incertaine.

5 Benchmarking

Les précédentes sections définissent différentes façons de résoudre le problème inverse des techniques d'imagerie cérébrale MEG / EEG. L'évaluation de ces solveurs reste difficile en raison de la vérité terrain de la localisation exacte des sources impliquées dans chaque tâche spécifique est complètement inconnue. Cette limitation vient principalement du fait que l'enregistrement est effectué sur le cuir chevelu et que plusieurs configurations de sources peuvent conduire exactement aux mêmes mesures sur les capteurs. La question est donc de savoir si toutes ces techniques de localisation de sources existantes sont capables de localiser avec précision les positions et les orientations des sources actuelles dans le cerveau dans un scénario d'acquisition réelle.

La façon typique de répondre cette question est d'effectuer des simulations [LDB02, MSLL93, LBD98]. Il consiste à fixer le nombre et l'emplacement, l'orientation et l'amplitude de plusieurs dipôles dans le cerveau, générer des données simulées corrompues par un bruit additif [LDB02, MSLL93, LBD98, DMHWH02, WAT+06]. Ces simulations sont malheureusement rarement réalistes: elles ne prennent pas en compte la nature non idéale des capteurs et les erreurs du modèle direct, et elles ne prennent pas en compte la structure de bruit complexe des mesures réelles. Les imprécisions dans le calcul de l'opérateur direct sont principalement dues à des approximations des valeurs de conductivité dans la tête et / ou des erreurs numériques associées aux approximations de têtes sphériques ou BEM basées sur des géométries de tête plus réalistes.

Des simulations plus sophistiquées pourraient être étudiées pour surmonter ces problèmes, mais nous proposons ici d'utiliser des données collectées à partir d'un objet physique artificiel dans une véritable machine MEG. Ceci a l'avantage que les résultats peuvent refléter plus étroitement la performance *in vivo* puisqu'ils incluent des facteurs qui ne peuvent pas être facilement inclus dans des simulations, telles que le bruit environnemental.

Afin de calibrer chaque dispositif MEG, des objets artificiels imitant l'activité cérébrale appelés «fantômes» sont construits par les fabricants de systèmes MEG. Ils sont basés sur la description théorique de [Ilm85] produisant des données réalistes correspondant à des sources de courant spatio-temporelles complexes incluant des géométries de tête réalistes. Dans un fantôme typique, de 4 à 32 dipôles de courant indépendants sont présents et les données MEG sont collectées séparément pour chaque dipôle. Les vraies positions et orientations dipolaires, et la morphologie du cerveau, les couches de crâne peuvent être extraites des données CT de rayons X [LMS+98]. Une limitation est que ces fantômes ne sont pas inappropriés pour l'EEG, mais il existe un certain travail sur la fabrication de dispositifs fantômes EEG [HSY16].

Cette section présente une nouvelle étude visant à valider les techniques de localisation à l'aide de différents jeux de données fantômes accessibles au public. D'autres travaux ont été réalisés par [HAH+15, LMS+98, BRM+01] en utilisant également des fantômes de vrais crânes pour étudier la performance des méthodes représentatives en considérant divers modèles de tête.

Les approches considérées dans cette section sont principalement celles toutes décrites dans la version longue de ma thèse. Ici on va comparer: Dipole fitting, Gamma-Map, RAP-MUSIC, MxNE, irMxNE, TF-MxNE, irTF-MxNE.

5.1 Results

Trois types d'erreurs ont été étudiés pour les différents solveurs, à savoir: l'erreur de position ou de localisation en millimètre, l'erreur d'orientation en Radian et l'erreur d'amplitude en pourcentage. Tous les solveurs étudiés ici sont implémentés dans le paquetage MNE-python [GLL+13, GLL+14].

La figure 11 montre les erreurs en position obtenues avec la plupart des solveurs pour les quatre dipôles simulés (5 à 8) (voir l'ensemble de données dans ma thèse) et pour les différents niveaux d'amplitude (20, 100, 200, 2000nAm). Il montre une erreur inférieure à 1 mm pour le rapport signal / bruit (SNR) élevé irréaliste (amplitude peak-to-peak égale à 1000nAm), mais également pour 200nAM et 100nAm. L'erreur de localisation s'aggrave avec le très faible SNR (20nAm).

L'approche de *dipole fitting* est très appropriée pour localiser l'activité neuronale lorsqu'un petit nombre d'ECD (*Equivalent Current Dipole*) peut décrire les données. Dans cet ensemble de données, chaque dipôle est enregistré seul. Les erreurs d'orientation sont affichées dans les figures 12.

RAP-MUSIC est une approche basée sur la technique MUSIC qui fonctionne aussi très bien quand peu d'ECD sont impliqués, particulièrement quand c'est un ensemble de données enregistrant seulement un dipôle à la fois. Il peut être vu comme très compétitif comparé à *dipole fitting* dans les figures 11- 12, montrant les erreurs en position et en orientation respectivement. Cependant,

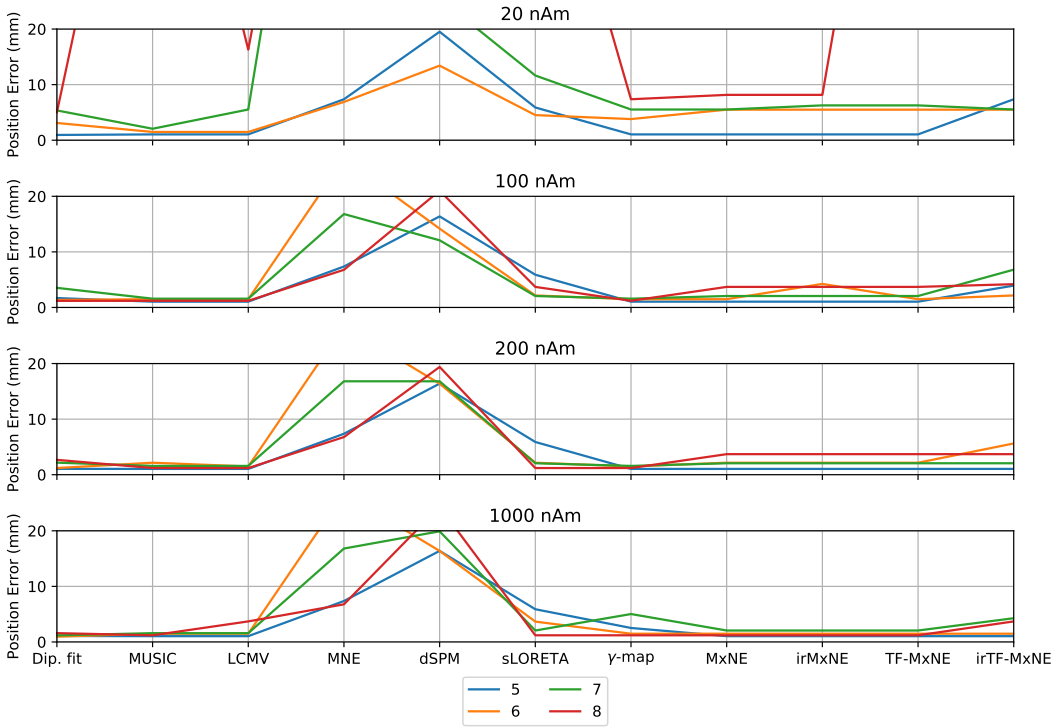


Figure 11: Comparaison de l'erreur en position entre la plupart des solveurs pour quatre dipôles différents. Le titre de chaque sous-figure donne le niveau du SNR (20nAm, 100nAm, 200nAm, ou 1000nAm).

pour une source profonde (dipôle 8) combinée à un très faible SNR (20nAm), la courbe rouge est en dehors de la boîte, ce qui signifie une erreur de localisation supérieure à 20mm. C'est un problème avec l'estimation du sous-espace de signal, où le rang de la covariance de données n'est pas bien estimé.

γ-map qui est une formulation bayésienne du problème inverse MEG / EEG, est moins performante que *dipole fitting* ou *RAP-MUSIC* pour un SNR très haut ou très bas. Pour les SNR dans une plage de données réalistes, son erreur de localisation est supérieure de 5 mm en fonction de la profondeur du dipôle étudié. *gamma*-map est pire pour l'amplitude de 1000nAm par rapport à 100nAm ou 200nAm, car elle surestime le bruit lors de l'estimation de l'hyperparamètre.

Pour MxNE et TF-MxNE, les erreurs illustrées dans la figure 11 démontrent respectivement une équivalence ou une légère amélioration lors de l'utilisation de TF-MxNE par rapport à MxNE, à l'exception du dipôle 8 le plus profond en rouge. Ceci s'explique par le fait que TF-MxNE est sensible aux hyperparamètres (λ_{space}, λ_{time}, taille de fenêtre et décalage temporel - *time shift*) en fonction du SNR et de la profondeur du dipôle. Ici, nous avons réglé les hyperparamètres en cherchant sur de grille (*grid-search*) similaire pour tous les dipôles, bien que l'on puisse penser que les hyperparamètres dépendent de la facilité des données (donc du SNR et de la profondeur de chaque dipôle).

D'autre part, MNE et dSPM sont étonnamment les méthodes donnant les pires résultats pour cet ensemble de données. Un argument important est le fait que l'étude est biaisée car nous savons que les données fantômes simulées sont focales / parcimonieux, alors que les méthodes type MNE et dSPM ne sont pas des méthodes focales. Nous prenons toujours le maximum d'amplitude et affichons le meilleur dipôle pour chaque méthode. En revanche, sLORETA n'est pas une méthode parcimonieuse, mais elle fonctionne beaucoup mieux que MNE et dSPM. Le «centre» de gravité estimé avec sLORETA est alors plus proche de l'emplacement exact du dipôle comparé au centre de dSPM ou de MNE.

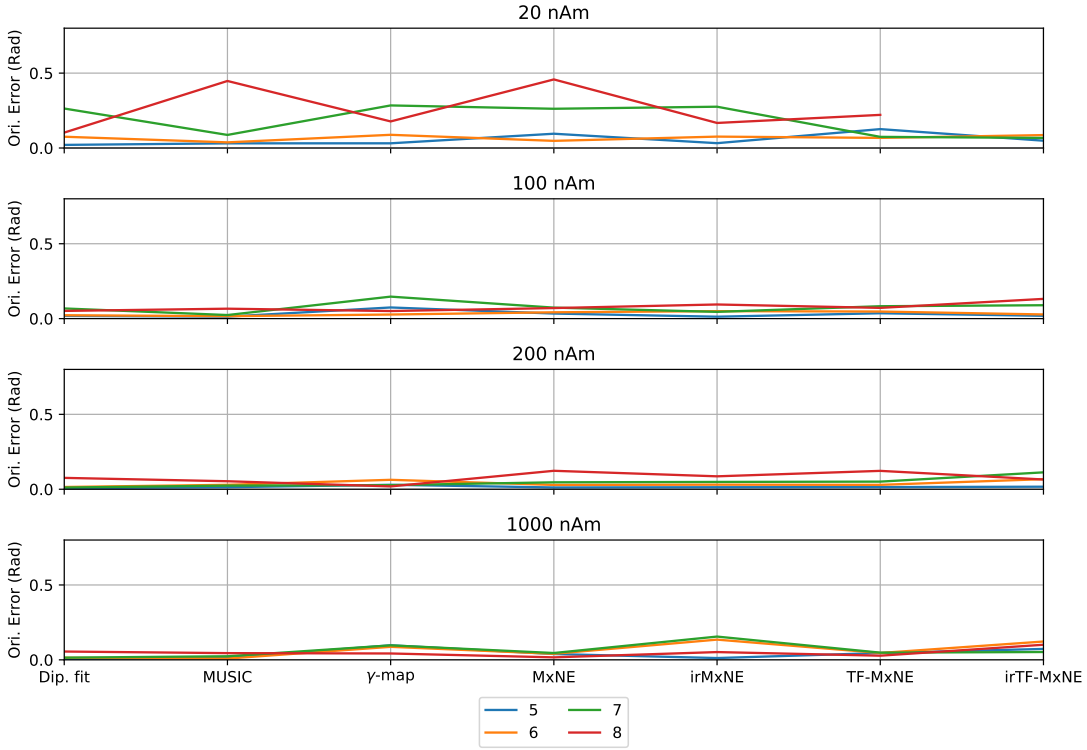


Figure 12: Comparaison des erreurs en orientation pour quatre dipôles différents.

6 Conclusion

Cette thèse a démontré différentes façons de résoudre le problème de localisation de source MEG / EEG. Il aborde les défis spécifiques rencontrés par les techniques de l'état de l'art, et tente de les améliorer point par point:

1. La promotion de la parcimonie structurée dans le domaine TF s'est avérée utile pour la reconstruction de sources non stationnaires, bien qu'elle doive fixer certains paramètres liés à la transformation de Gabor, qui sont impliqués dans la résolution de TF. La première amélioration proposée dans cette thèse a été d'aborder le choix de ces paramètres, ce qui peut être très préjudiciable pour l'analyse des ondes cérébrales avec des caractéristiques de TF variables. Il fournit une nouvelle technique basée sur une norme mixte TF multi-échelle permettant de localiser plus précisément la source estimée dans l'espace et le temps (voir chapitre 3 de la version longue de ma thèse).

2. La formulation du problème inverse MEG / EEG a été principalement écrite comme une régression pénalisée, ce qui signifie qu'elle doit introduire une connaissance antérieure comme un terme de régularisation dans la fonction objective. Cela entraîne l'ajout d'un hyperparamètre au modèle qui doit être réglé. Cette thèse aborde ce deuxième défi de deux façons, en reformulant le problème comme fait dans la communauté bayésienne. La formulation bayésienne permet d'ajouter de manière hiérarchique des hyperparamètres qui sont alternativement estimés avec les paramètres principaux du modèle (les sources). Les deux principaux avantages sont l'estimation directe des hyperparamètres et la possibilité d'utiliser l'échantillonnage pour étudier l'incertitude de ces solveurs. Ces deux points ont été présentés au chapitre 4.

3. Une étape importante après le développement de toute nouvelle technique est de la valider par une comparaison avec les autres méthodes existantes. Cela a longtemps été une étape difficile car il est toujours difficile de développer de bonnes simulations réalistes. Dans ce but, plusieurs études ont étudié des jeux de données fantômes, qui consistent en des données réelles enregistrées avec un appareil imitant une tête humaine avec des sources focales. Le

chapitre 5 présente une comparaison des solveurs présentés dans cette thèse sur plusieurs jeux de données fantômes.

4. Une liste de perspective et de futurs travaux est publiée à la fin de ma thèse.

References

[BBG17] Y. Bekhti, R. Badeau, and A. Gramfort. Hyperparameter estimation in maximum a posteriori regression using group sparsity with an application to brain imaging. In *2017 25th European Signal Processing Conference (EUSIPCO)*, pages 246–250, Aug 2017.

[BGZvW17] Yousra Bekhti, Alexandre Gramfort, Nicolas Zilber, and Virginie van Wassenhove. Decoding the categorization of visual motion with magnetoencephalography. *bioRxiv*, 2017.

[BLSG18] Yousra Bekhti, Felix Lucka, Joseph Salmon, and Alexandre Gramfort. A hierarchical bayesian perspective on majorization-minimization for non-convex sparse regression: application to m/eeg source imaging. *Inverse Problems*, 2018.

[BRM+01] S Baillet, JJ Riera, G Marin, JF Mangin, J Aubert, and L Garnero. Evaluation of inverse methods and head models for eeg source localization using a human skull phantom. *Physics in medicine and biology*, 46(1):77, 2001.

[BSJ+16] Yousra Bekhti, Daniel Strohmeier, Mainak Jas, Roland Badeau, and Alexandre Gramfort. M/eeg source localization with multi-scale time-frequency dictionaries. In *6th International Workshop on Pattern Recognition in Neuroimaging (PRNI)*, 2016.

[BVVN09] Andrew Bolstad, Barry Van Veen, and Robert Nowak. Space–time event sparse penalization for magneto-/electroencephalography. *NeuroImage*, 46(4):1066–1081, 2009.

[CCHMV+15] Sebastián Castaño-Candamil, Johannes Höhne, Juan-David Martínez-Vargas, Xing-Wei An, German Castellanos-Domínguez, and Stefan Haufe. Solving the eeg inverse problem based on space–time–frequency structured sparsity constraints. *NeuroImage*, 118:598–612, 2015.

[CDS98] S. S. Chen, D. L. Donoho, and M. A. Saunders. Atomic decomposition by basis pursuit. *SIAM J. Sci. Comput.*, 20(1):33–61, 1998.

[Cha07] Rick Chartrand. Exact reconstruction of sparse signals via nonconvex minimization. *IEEE Signal Process. Lett.*, 14(10):707–710, 2007.

[CWB08] Emmanuel J Candes, Michael B Wakin, and Stephen P Boyd. Enhancing sparsity by reweighted l_1 minimization. *Journal of Fourier analysis and applications*, 14(5-6):877–905, 2008.

[Dea10] Ingrid Daubechies et al. Iteratively reweighted least squares minimization for sparse recovery. *Comm. Pure Appl. Math.*, 63(1):1–38, 2010.

[DLF+00] Anders M Dale, Arthur K Liu, Bruce R Fischl, Randy L Buckner, John W Belliveau, Jeffrey D Lewine, and Eric Halgren. Dynamic statistical parametric mapping: combining fmri and meg for high-resolution imaging of cortical activity. *Neuron*, 26(1):55–67, 2000.

[DMHWH02] Jan Casper De Munck, Hilde M Huizenga, Lourens J Waldorp, and RA Heethaar. Estimating stationary dipoles from meg/eeg data contaminated with spatially and temporally correlated background noise. *IEEE Transactions on Signal Processing*, 50(7):1565–1572, 2002.

[FHT10] Jerome Friedman, Trevor Hastie, and Rob Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, 33(1):1, 2010.

[GD15] Chunli Guo and Mike E Davies. Near optimal compressed sensing without priors: Parametric SURE approximate message passing. *IEEE Trans. on Signal Processing*, 63(8):2130–2141, 2015.

[GKH12] Alexandre Gramfort, Matthieu Kowalski, and Matti Hämäläinen. Mixed-norm estimates for the M/EEG inverse problem using accelerated gradient methods. *Physics in medicine and biology*, 57(7):1937, 2012.

[GLL+13] Alexandre Gramfort, Martin Luessi, Eric Larson, Denis A Engemann, Daniel Strohmeier, Christian Brodbeck, Roman Goj, Mainak Jas, Teon Brooks, Lauri Parkkonen, and Matti S. Hämäläinen. MEG and EEG data analysis with MNE-Python. *Frontiers in Neurosci.*, 7(267), Dec. 2013.

[GLL+14] Alexandre Gramfort, Martin Luessi, Eric Larson, Denis A. Engemann, Daniel Strohmeier, Christian Brodbeck, Lauri Parkkonen, and Matti S. Hämäläinen. MNE software for processing MEG and EEG data. *NeuroImage*, 86:446–460, Feb. 2014.

[GSH+13] A. Gramfort, D. Strohmeier, J. Haueisen, M.S. Hamalainen, and M. Kowalski. Time-frequency mixed-norm estimates: Sparse M/EEG imaging with non-stationary source activations. *NeuroImage*, 70:410 – 422, 2013.

[Had02] Jacques Hadamard. Sur les problèmes aux dérivées partielles et leur signification physique. *Princeton university bulletin*, pages 49–52, 1902.

[HAH+15] OMAR Hazim, Alwani Liyan AHMAD, Noburo Hayashi, Zamzuri Idris, and Jafri Malin Abdullah. Magnetoencephalography phantom comparison and validation: Hospital universiti sains malaysia (husm) requisite. *The Malaysian journal of medical sciences: MJMS*, 22(Spec Issue):20, 2015.

[HDS+06] Ming-Xiong Huang, Anders M Dale, Tao Song, Eric Halgren, Deborah L Harrington, Igor Podgorny, Jose M Canive, Stephen Lewis, and Roland R Lee. Vector-based spatial–temporal minimum l1-norm solution for meg. *NeuroImage*, 31(3):1025–1037, 2006.

[HHR+14] Ming-Xiong Huang, Charles W Huang, Ashley Robb, AnneMarie Angeles, Sharon L Nichols, Dewleen G Baker, Tao Song, Deborah L Harrington, Rebecca J Theilmann, Ramesh Srinivasan, et al. Meg source imaging method using fast l1 minimum-norm and its applications to signals with brain noise and human resting-state source amplitude images. *Neuroimage*, 84:585–604, 2014.

[HI94] Matti S Hämäläinen and Risto J Ilmoniemi. Interpreting magnetic fields of the brain: minimum norm estimates. *Medical & biological engineering & computing*, 32(1):35–42, 1994.

[HNZ+08] Stefan Haufe, Vadim V Nikulin, Andreas Ziehe, Klaus-Robert Müller, and Guido Nolte. Combining sparsity and rotational invariance in EEG/MEG source reconstruction. *NeuroImage*, 42(2):726–738, 2008.

[HSY16] W David Hairston, Geoffrey A Slipher, and Alfred B Yu. Ballistic gelatin as a putative substrate for eeg phantom devices. *arXiv preprint arXiv:1609.07691*, 2016.

[Ilm85] RJ Ilmoniemi. The forward and inverse problems in the spherical model. *Biomagnetism: Applications & Theory*, 1985.

[JEB+17] Mainak Jas, Denis A. Engemann, Yousra Bekhti, Federico Raimondo, and Alexandre Gramfort. Autoreject: Automated artifact rejection for meg and eeg data. *NeuroImage*, 159(Supplement C):417 – 429, 2017.

[JER+16] M. Jas, D. Engemann, F. Raimondo, Y. Bekhti, and A. Gramfort. Automated rejection and repair of bad trials in meg/eeg. In *2016 International Workshop on Pattern Recognition in Neuroimaging (PRNI)*, pages 1–4, June 2016.

[LBD98] Arthur K Liu, John W Belliveau, and Anders M Dale. Spatiotemporal imaging of human brain activity using functional mri constrained magnetoencephalography data: Monte carlo simulations. *Proceedings of the National Academy of Sciences*, 95(15):8945–8950, 1998.

[LBU07] Florian Luisier, Thierry Blu, and Michael Unser. A new SURE approach to image denoising: Interscale orthonormal wavelet thresholding. *IEEE Trans. on image processing*, 16(3):593–606, 2007.

[LDB02] Arthur K Liu, Anders M Dale, and John W Belliveau. Monte carlo simulation studies of eeg and meg localization accuracy. *Human brain mapping*, 16(1):47–62, 2002.

[LMS+98] RM Leahy, JC Mosher, ME Spencer, MX Huang, and JD Lewine. A study of dipole localization accuracy for meg and eeg using a human skull phantom. *Electroencephalography and clinical neurophysiology*, 107(2):159–173, 1998.

[LPBW12] Felix Lucka, Sampsa Pursiainen, Martin Burger, and Carsten H Wolters. Hierarchical Bayesian inference for the EEG inverse problem using realistic FE head models: depth localization and source separation for focal primary currents. *Neuroimage*, 61(4):1364–1382, 2012.

[LWA+06] Fa-Hsuan Lin, Thomas Witzel, Seppo P Ahlfors, Steven M Stufflebeam, John W Belliveau, and Matti S Hämäläinen. Assessing and improving the spatial accuracy in meg source localization by depth-weighted minimum-norm estimates. *Neuroimage*, 31(1):160–171, 2006.

[ML99] John C Mosher and Richard M Leahy. Source localization using recursively applied and projected (RAP) MUSIC. *IEEE Trans. Signal Process.*, 47(2):332–340, Feb. 1999.

[MO95] Kanta Matsuura and Yoichi Okabe. Selective minimum-norm solution of the biomagnetic inverse problem. *Biomedical Engineering, IEEE Transactions on*, 42(6):608–615, 1995.

[MSLL93] John C Mosher, Michael E Spencer, Richard M Leahy, and Paul S Lewis. Error bounds for eeg and meg dipole source localization. *Electroencephalography and clinical Neurophysiology*, 86(5):303–321, 1993.

[OHG09] Wanmei Ou, Matti S Hämäläinen, and Polina Golland. A distributed spatio-temporal EEG/MEG inverse solver. *NeuroImage*, 44(3):932–946, 2009.

[OWA+12] Julia P Owen, David P Wipf, Hagai T Attias, Kensuke Sekihara, and Srikantan S Nagarajan. Performance evaluation of the champagne source reconstruction algorithm on simulated and real m/eeg data. *NeuroImage*, 60(1):305–323, 2012.

[PBDF15] Marcelo Pereyra, José M Bioucas-Dias, and Mário AT Figueiredo. Maximum-a-posteriori estimation with unknown regularisation parameters. In *Proc. EUSIPCO*, pages 230–234. IEEE, 2015.

[PM02a] R D Pascual-Marqui. Standardized low-resolution brain electromagnetic tomography (sLORETA): technical details. *Methods Find Exp Clin Pharmacol*, 24(Suppl D):5–12, 2002.

[PM+02b] Roberto Domingo Pascual-Marqui et al. Standardized low-resolution brain electromagnetic tomography (sloreta): technical details. *Methods Find Exp Clin Pharmacol*, 24(Suppl D):5–12, 2002.

[PMML94] Roberto D Pascual-Marqui, Christoph M Michel, and Dietrich Lehmann. Low resolution electromagnetic tomography: a new method for localizing electrical activity in the brain. *International Journal of psychophysiology*, 18(1):49–65, 1994.

[S+78] Gideon Schwarz et al. Estimating the dimension of a model. *Ann. Stat.*, 6(2):461–464, 1978.

[SBHG16] D. Strohmeier, Y. Bekhti, J. Haueisen, and A. Gramfort. The iterative reweighted mixed-norm estimate for spatio-temporal meg/eeg source reconstruction. *IEEE Trans. Med. Imag.*, 35(10):2218–2228, Oct 2016.

[SCY08] Rayan Saab, Rick Chartrand, and Ozgur Yilmaz. Stable sparse approximations via nonconvex optimization. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 3885–3888. IEEE, 2008.

[SGH15] Daniel Strohmeier, Alexandre Gramfort, and Jens Haueisen. MEG/EEG source imaging with a non-convex penalty in the time-frequency domain. In *Pattern Recognition in NeuroImaging (PRNI), 2015 International Workshop on*, pages 21–24. IEEE, 2015.

[SHG14] Daniel Strohmeier, Jens Haueisen, and Alexandre Gramfort. Improved MEG/EEG source localization with reweighted mixed-norms. In *PRNI, 2014 International Workshop on*, pages 1–4. IEEE, 2014.

[Ste81] Charles M Stein. Estimation of the mean of a multivariate normal distribution. *The annals of Statistics*, pages 1135–1151, 1981.

[Tib96] R. Tibshirani. Regression shrinkage and selection via the lasso. 58(1):267–288, 1996.

[Tse10] P. Tseng. Approximation accuracy, gradient methods, and error bound for structured convex optimization. *Math. Program.*, 125:263–295, Oct. 2010.

[UHS99] K. Uutela, M. S. Hämäläinen, and E. Somersalo. Visualization of magnetoencephalographic data using minimum current estimates. *NeuroImage*, 10(2):173–180, Aug. 1999.

[VSVHSB+09] Pedro A Valdés-Sosa, Mayrim Vega-Hernández, José Miguel Sánchez-Bornot, Eduardo Martínez-Montes, and María Antonieta Bobes. EEG source imaging with spatio-temporal tomographic nonnegative independent component analysis. *Human brain mapping*, 30(6):1898–1910, 2009.

[Was08] Wilhelmina Johanna Gerarda van de Wassenberg. Multichannel eeg. 2008.

[WAT+06] Carsten Hermann Wolters, Alfred Anwander, X Tricoche, D Weinstein, Martin A Koch, and RS MacLeod. 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, 2006.

[WHM+07] MP Weisend, FM Hanlon, R Montano, SP Ahlfors, AC Leuthold, D Pantazis, JC Mosher, AP Georgopoulos, MS Hämäläinen, and CJ Aine. Paving the way for cross-site pooling of magnetoencephalography (meg) data. In *International Congress Series*, volume 1300, pages 615–618. Elsevier, 2007.

[Wip06] David Paul Wipf. *Bayesian methods for finding sparse representations*. ProQuest, 2006.

[WN09] David Wipf and Srikantan Nagarajan. A unified Bayesian framework for MEG/EEG source imaging. *NeuroImage*, 44(3):947–966, 2009.

[ZR11a] Zhilin Zhang and Bhaskar D Rao. Exploiting correlation in sparse signal recovery problems: Multiple measurement vectors, block sparsity, and time-varying sparsity. In *28th Int. Conference on Machine learning (ICML)*, 2011.

[ZR11b] Zhilin Zhang and Bhaskar D Rao. Iterative reweighted algorithms for sparse signal recovery with temporally correlated source vectors. In *IEEE Int. Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3932–3935, 2011.