Building clinical biomarkers from cerebral electrophysiology Brain Age as a measure of neurocognitive disorders

David Sabbagh

université
PARIS-SACLAY

Building clinical biomarkers from cerebral electrophysiology: Brain Age as a measure of neurocognitive disorders

Construction de biomarqueurs cliniques à partir de l'electrophysiologie cérébrale : l'Age du Cerveau comme mesure des troubles neurocognitifs

Thèse de doctorat de l'université Paris-Saclay

École doctorale n° 580, Sciences et Technologies de l'Information et de la Communication (STIC)
Spécialité de doctorat : Traitement du signal et des images
Graduate School : Informatique et sciences du numérique
Référent : Faculté des sciences d'Orsay

Thèse préparée dans l'unité de recherche **Inria Saclay-Île-de-France** (Université Paris-Saclay, Inria), sous la direction de **Alexandre GRAMFORT**, Directeur de recherche, la co-direction de **Etienne GAYAT**, Professeur des universités - praticien hospitalier, le co-encadrement de **Denis A.ENGEMANN**, Chercheur

Thèse soutenue à Paris-Saclay, le 15 décembre 2021, par

David SABBAGH

Composition du jury

Sylvain CHEVALLIER Maître de conférences, HDR, Université de Versailles St-Quentin	Président
Fabien LOTTE Directeur de Recherche, Inria Bordeaux	Rapporteur & Examinateur
Karim JERBI Professeur agrégé, Université de Montréal	Rapporteur & Examinateur
Vadim NIKULIN Professeur, Max Planck Institute Leipzig	Examinateur
Alexandre GRAMFORT Directeur de Recherche, Université Paris-Saclay	Directeur de thèse

THESE DE DOCTORAT

NNT : 2021UPASG101

université PARIS-SACLAY

ÉCOLE DOCTORALE
Sciences et technologies de l'information et de la communication (STIC)

Title: Building clinical biomarkers from cerebral electrophysiology: Brain Age as a measure of neurocognitive disorders

Keywords: Machine learning, Neuroimaging, MEG, EEG, Riemannian manifold, Biomarker

Abstract: Neurodegenerative diseases are among the top causes of worldwide mortality. Unfortunately, early diagnosis is challenging as it requires a frequently too late indication of biomedical exam and dedicated laboratory equipments. It also often relies on research-based predictive measures suffering from selection bias. This thesis investigates a promising solution to tackle these problems: a robust method to build predictive biological markers from M/EEG brain signals, directly usable in the clinic, and validated against neurocognitive disorders following general anaesthesia.

In a first (theoretical) contribution [Sab+19], we benchmarked M/EEG regression models that could learn from between-channels covariance matrices as a compact summary of spatial distribution of power of high-dimensional brain M/EEG signal. Mathematical analysis identified different models supporting perfect prediction under ideal circumstances when the outcome is either linear or log-linear in the source power. These models are based on the mathematically principled approaches of supervised spatial filtering and projection with Riemannian geometry, and enjoy optimal prediction guarantees without the need of costly source localization. Our simulation-based findings were consistent with the mathematical analysis and suggested that these regression algorithms were robust across data generating scenarios and model violations. This study suggested that the Riemannian methods have the potential to support automated large-scale analysis of M/EEG data in the absence of MRI scans, which is one condition to be practically used in the clinic for biomarker development.

In a second (empirical) contribution [Sab+20], we validated our predictive modeling framework with several publicly available neuroimaging datasets and showed it can be used to learn the surrogate biomarker of brain age from research-grade M/EEG signals, without source localization and with minimal pre-processing. Our results demonstrate that our Riemannian data-driven method does not fall far behind the gold-standard source localization methods with biophysical priors, that depend on manual data processing, the costly availability of anatomical MRI images and specialized knowledge in M/EEG source modeling. Subsequent large-scale empirical analysis provided evidence that brain age derived from MEG captures unique information related to neuronal activity that was not explained by anatomical MRI. They also suggested that, consistent with simulations, Riemannian methods are generally a good bet across a wide range of settings with considerable robustness to different choices of preprocessing including minimalistic preprocessing. The good performance obtained on MEG was also reached with research-grade clinical EEG.

In a third (clinical) contribution [Sab+21, in prep.], we validated the concept of M/EEG-derived brain age directly in the operating rooms of Lariboisière hospital in Paris, from monitoring-grade clinical EEG during the particular period of general anaesthesia. We validated our EEG-based brain age measure against intra-operative complications and brain health in anaesthesia population with a potential link to postoperative cognitive dysfunctions, unveiling it as a promising clinical biomarker of neurocognitive disorders. We also showed that the drug critically impacts brain age prediction and demonstrated the robustness applicability of our approach across different types of drugs.

By combining concepts previously investigated separately, our contribution demonstrates the clinical relevance of EEG-brain-age in revealing pathologies of brain function and obtaining brain health assessments in situations where MRI scans cannot be conducted. It also provides early evidence that anaesthesia-based modeling has the potential to help biomarker discovery and eventually revolutionize preventive medicine.

université PARIS-SACLAY
ÉCOLE DOCTORALE
Sciences et technologies de l'information et de la communication (STIC)

Titre : Construction de biomarqueurs cliniques à partir de l'electrophysiologie cérébrale : l'Age du Cerveau comme mesure des troubles neurocognitifs

Mots clés : Apprentissage statistique, Imagerie cérébrale, MEG, EEG, Variétés riemanniennes, Biomarqueur

Résumé : Les maladies neurodégénératives figurent parmi les principales causes de mortalité dans le monde. Malheureusement, leur diagnostic précoce nécessite un examen médical prescrit souvent trop tardivement et des équipements de laboratoire dédiés. Il repose aussi fréquemment sur des mesures prédictives souffrant d'un biais de sélection. Cette thèse présente une solution prometteuse à ces problèmes : une méthode robuste, directement utilisable en clinique, pour construire des biomarqueurs prédictifs à partir des signaux cérébraux M/EEG, validés contre les troubles neurocognitifs apparaissant après une anesthésie générale.

Dans une première contribution (théorique) [Sab+19], nous avons évalué des modèles de régression capables d'apprendre des biomarqueurs à partir des matrices de covariance de signaux M/EEG. Notre analyse mathématique a identifié différents modèles garantissant une prédiction parfaite dans des circonstances idéales, lorsque la cible est une fonction (log-)linéaire en la puissance des sources cérébrales. Ces modèles, basés sur les approches mathématiques de filtrage spatial supervisé et de géométrie riemannienne, permettent une prédiction optimale sans nécessiter une coûteuse localisation des sources. Nos simulations confirment cette analyse mathématique et suggèrent que ces algorithmes de régression sont robustes à travers les mécanismes de génération de données et les violations de modèles. Cette étude suggère que les méthodes riemanniennes sont des méthodes de choix pour l'analyse automatisée à grande échelle des données M/EEG en l'absence d'IRM, condition importante pour pouvoir développer des biomarqueurs cliniques.

Dans une deuxième contribution (empirique) [Sab+20], nous avons validé nos modèles prédictifs sur plusieurs ensembles de données de neuro-imagerie et avons montré qu'ils peuvent être utilisé pour apprendre l'âge du cerveau à partir de signaux cérébraux M/EEG, sans localisation de sources, et avec un prétraitement minimal des données. De plus, la performance de notre méthode riemannienne est proche de celle des méthodes de référence nécessitant une localisation de sources et donc un traitement manuel des données, la disponibilité d'images IRM anatomiques et une expertise en modélisation de sources M/EEG. Une analyse empirique à grande échelle a ensuite permis de démontrer que l'âge du cerveau dérivé de la MEG capture des informations uniques liées à l'activité neuronale et non expliquées par l'IRM anatomique. Conformément aux simulations, ces résultats suggèrent également que l'approche riemannienne est une méthode pouvant s'appliquer dans un large éventail de situations, avec une robustesse considérable aux différents choix de prétraitement, y compris minimaliste. Les bonnes performances obtenues avec la MEG ont ensuite été répliquées avec des EEGs de qualité recherche.

Dans une troisième contribution (clinique) [Sab+21, en préparation], nous avons validé le concept d'âge cérébral directement au bloc opératoire de l'hôpital Lariboisière à Paris, à partir d'EEG de qualité clinique recueillis pendant la période de l'anesthésie générale. Nous avons évalué notre mesure de l'âge cérébral comme prédicteur de complications peropératoires liées aux dysfonctions cognitives post opération, validant ainsi l'âge du cerveau comme un biomarqueur clinique prometteur des troubles neurocognitifs. Nous avons également montré que le sédatif utilisé a un impact important sur la prédiction de l'âge du cerveau et avons démontré la robustesse de notre approche à différents types de médicaments.

Combinant des concepts précédemment étudiés séparément, notre contribution démontre la pertinence clinique de la notion d'âge du cerveau prédit à partir de l'EEG pour révéler les pathologies des fonctions cérébrales dans des situations où l'IRM ne peut pas être réalisée. Ces résultats fournissent également une première preuve que l'anesthésie générale est une période propice à la découverte de biomarqueurs cérébraux, avec un impact potentiel profond sur la médecine préventive.

Acknowledgement

This thesis is the result of an incredible intellectual journey that began 3 years ago when I first met my future PhD advisor, Denis, during this cold and snowy winter night on the Plateau de Saclay. We didn't know it by then but it was the beginning of a sensational scientific ride. I don't have the words to express my gratitude to him for his kindness during these 3 years, his daily care and his generosity. I was very lucky to meet him: his unfaltering enthusiasm made this thesis an unforgettable learning experience. Then two other advisors quickly joined the adventure.

To Alex I extend my deepest gratitude for welcoming me to his lab, and supporting me through the challenges of this project. I can only imagine his surprise to see this 43 years old guy at the time approaching him to pursue a PhD. I thank him very much for not having laughed at me and for his trust. His pragmatism, scientific rigour, attention to detail and dedication to the open source software community will continue to inspire me for years to come.

To Etienne I am also particularly grateful, for his faith in the project and so kindly and so humbly welcoming me in his team of medical doctors. His double training in medicine and statistics made him the perfect guide and bridge to the clinical world. Denis, Alex and Etienne have been the mentors any PhD student could dream of. None of this could have been possible without their competence, their benevolence, kindness, and availability.

I feel also very indebted to my collaborators from which I learnt so much. To Pierre Ablin at Inria on the mathematical aspects of this work: I will always remember the intensity of our sprint against the NeurIPS deadline. To Fabrice Vallée, Jona Joachim, Jerome Cartailler and Cyril Touchard at AP-HP: it is rare to meet a team of medical doctors so open to external collaborations and enthusiastic about what mathematics and compute science can bring to the practice of medicine. They made me realize how little I know on what ultimately matters the most.

I also want to thank my colleagues in the Parietal team at Inria. Their comradeship and support made my time so much pleasant. On account of his leader, Bertrand Thirion, Parietal is more than a team, it's a culture in which I grew naturally. In particular I feel very lucky for meeting Valentin, Hicham, Maëliss and Hubert, who became my friends. Writing this acknowledgment makes me realize that this

scientific endeavour was ultimately a human adventure. To all these people: you are what I will remember the most from these 3 years.

On a more personal note, I am incredibly grateful beyond words to my family and my friends for their unconditional affection since the beginning and their unfailing support. To my love, Virginie, who supports me in every moment: it is her daily love and intelligence that gives me the courage and the strength to face the exciting challenges of my life. And finally, to my two little girls Ava and Lisa. You're too little to even understand these words but big in your ability to wonder, laugh and love. Let this thesis be a testimony of the power of will. Never ever let anyone tell you what you can or can't do. This thesis is for you.

David Sabbagh

Paris, December 2021

Contents

Introduction

Contents

What reveals the most about us? Is it the colour of our eyes, our heart beats, our blood pressure, our clicks on a webpage? Even though all of this reveals some part of us, there is one single organ that contains all the information about our thoughts, our feelings, our memory, our actions: this is our brain. The billions of interconnected cells that compose our brain, firing at hectic pace, are literally our perception, our thinking, our emotions, our memories and ultimately define who we are. If our brain activity is expressing cognition then changes to cognition should be revealed in the brain signals extracted from this activity. These signals are therefore the source of promising predictive biological markers of our functioning and perhaps more importantly dysfunctioning.

Brain diseases have a dramatic impact on life, ranging from neurodegenerative diseases to loss of brain functions. Besides these devastating consequences, they also are among top causes of death in the European Union. It accounted for 18M deaths between 2011 and 2017 [Eura] among which 13M due to cerebrovascular and 5M due to neurodegenerative diseases like Parkinson's, Alzheimer's, dementia, stroke, multiple sclerosis or epilepsy. This is even more acute for patients over the age of 60. For example an American woman aged 65 today has almost $25\,\%$ chance of contracting Alzheimer's disease during her lifetime [Aa2]. This age group represents about one quarter of the global population and will continue to grow at a fast pace in the coming decades [Eurc; Eurb; Vol+20]: we expect twice more 65y+ humans in 2050 than today globally [Un2]. Pathologies of the brain are therefore one of

the biggest challenges for medicine today and brain health a top priority in public health.

We are not equal when facing neurocognitive disorders. But even today it is difficult to know if a patient will develop a particular brain disease or if his brain will age normally. As a consequence, these pathologies are too often detected at late stages, rendering treatment significantly less efficient. For the moment, no biological marker is able to early identify high-risk patients. Having an early test of cognitive dysfunction, directly built from brain signals and easily available for millions of persons, would allow for better detection and treatment of brain diseases. This is the subject of this thesis.

More precisely this theoretical and experimental work will investigate a **general method to build predictive biomarkers from brain signals, directly usable in the clinic, with an application to predict neurocognitive disorders**. The objective of this chapter is to provide a general overview of the sequence of challenges standing in the road to this endeavour. We need to understand:

How the brain operates: even if the precise working of the human brain is still largely unknown, we first need to have a rough picture of how it is structured and understand the basis of its activity.
How to extract signals from the brain: we then need to find a way to capture this activity and extract a measurable signal.
How to predict from brain signal: with brain signal as input data we should seek the best algorithm to predict from it, simple enough to be usable in the clinical settings.
What to predict from brain signal: once we have the input data and the algorithm we should determine the target of prediction, a target that is both easily available and linked to the clinical outcome of interest. This leads to the concept of brain age as a promising biomarker of neurocognitive disorders.
How to estimate brain age in the lab: we then need to put this all together and run experiments to see how to estimate this biomarker in the comfortable conditions of a research laboratory.
How to translate brain age to the clinic: then we'll investigate how to translate the brain age biomarker in the more challenging conditions of the clinic.

Each of these challenges are investigated in further details in subsequent sections.

How the brain operates

The human brain roughly weights as little as $1.5\,\mathrm{kg}$ and operates on the same power than a simple electric bulb ($\sim 20\,\mathrm{W}$, to be compared with the $8000\,\mathrm{W}$ consumed by

IBM Watson that outperformed the best human in Jeopardy in 2011). This sobriety hides a formidable complexity [Fac06].

With $\sim$100 billions of excitable nerve cells, the neurons, each connected to 7000 other neurons on average, the human brain is arguably the most complex organ of the human body, and the most complex known object in the universe. It consumes one fifth of the body total energy expenditure, a huge consumption compared to its relative weight. How it performs a wide range of cognitive functions, from visual recognition to language understanding, speech, social interaction, and executive control is, for the most part, still a mystery. Understanding the human brain is therefore one of the most significant challenges of the 21st century. Fortunately, some of its inner working is understood today [DA01; Ger+14].

Located under our skull, the brain is composed of two hemispheres [Figure 1 (Left)]. When removing one hemisphere to have a view on the inside [Figure 1 (Right)] we discover three main components: the hindbrain, the midbrain and the forebrain. Following the evolution of species, the most ancient structures contained in the hind and midbrains handle elementary behaviours: the *vegetative* functions (basic body functions common to all living entities like heat regulation, sweating, sleeping, eating, drinking and reproducing) and the *affective* functions (emotions and feelings), whereas the most recent parts located in the forebrain manages the higher-level *cognitive* functions (perceptions, thinking and action) [NVVH07; HP17].

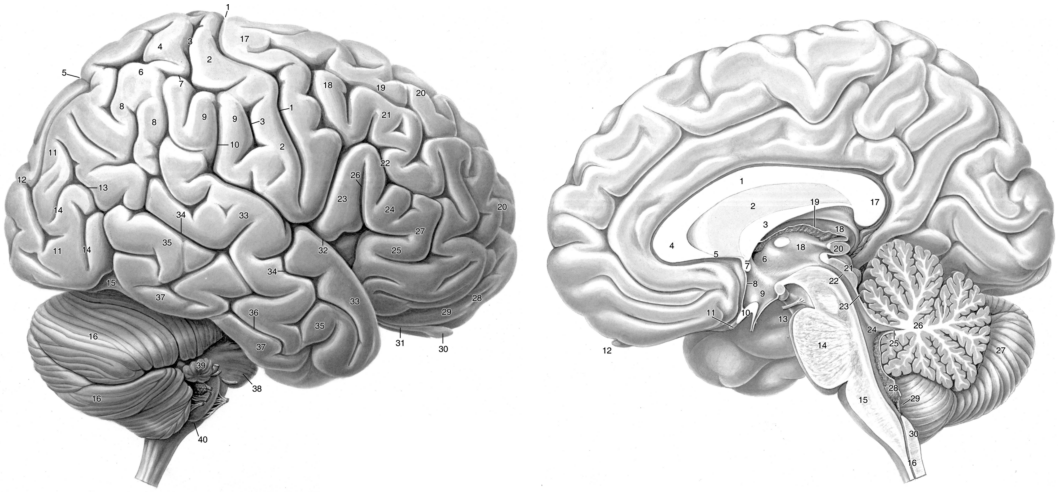


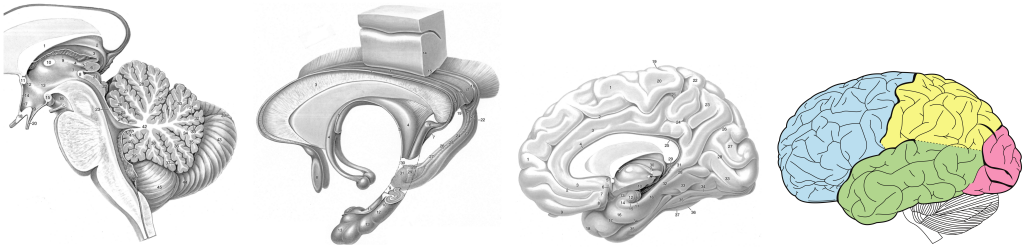
Fig. 1: The human brain under a lateral view **(Left)** and sagittal view **(Right)**, which unveils the three major internal structures (the forebrain, the midbrain and the hindbrain) detailed below. *Source:* [NVVH07]

The **hindbrain (the brain stem and cerebellum)** [Figure 2a] is the most ancient part of our brain [NVVH07]. It is composed of the medulla oblongata (which controls involuntary movements like heart rate, breathing, blood pressure), the pons (which deals with swallowing, bladder control, facial expressions, chewing, saliva, tears,

and posture) and the cerebellum (which controls the balance and coordination of our movements).

The **midbrain (the limbic system)** [Figure 2b] is a primitive survival system, fulfilling our animal needs and handling our emotions [NVVH07]. It is mainly composed of the amygdala (which deals with anxiety, sadness, and our responses to fear), the hippocampus (a scratch board for memory, first targeted by Alzheimer's disease), the thalamus (a sensory middleman that receives information from our sensory organs and sends them to the forebrain for processing) and the hypothalamus (the heart of vegetative functions).

The **forebrain (the cortex)** [Figure 2c] is the most recent part in the evolution of species. It handles our perceptions, our actions and our thinking [NVVH07]. It's divided into four lobes [Fac06]. The *frontal lobe* is mainly in charge of our thinking (reasoning, planning, decision-making and executive function, in particular in its front part called the prefrontal cortex), and our body's movement. The *parietal lobe* integrates information from our senses. The *temporal lobe* is associated with language, memory and emotions, and houses our auditory cortex. The *occipital lobe*, at the back of our head, is where our visual cortex resides and is almost entirely dedicated to vision.



a b c d

Fig. 2: Inside the three main components of the human brain: the hindbrain **(a)**, the midbrain **(b)** and the forebrain **(c)**. The forebrain, also known as the cortex, houses our higher cognitive abilities and is itself divided into four principal lobes **(d)**: the frontal lobe (blue), the temporal lobe (green), the parietal lobe (yellow) and the occipital lobe (red). *Source:* [NVVH07]

The cortex hosts the major part of our neurons [Fac06], and houses most of the brain abilities and higher-level functions. Some of them are detailed in [Figure 3]. For instance, our body movement is handled by a top strip in the prefrontal cortex called the primary motor cortex. The strip right next to it in the parietal lobe called the primary somatosensory cortex houses our sense of touch. Neural activity from the frontal lobe is recorded during general anaesthesia to monitor the depth of anaesthesia.

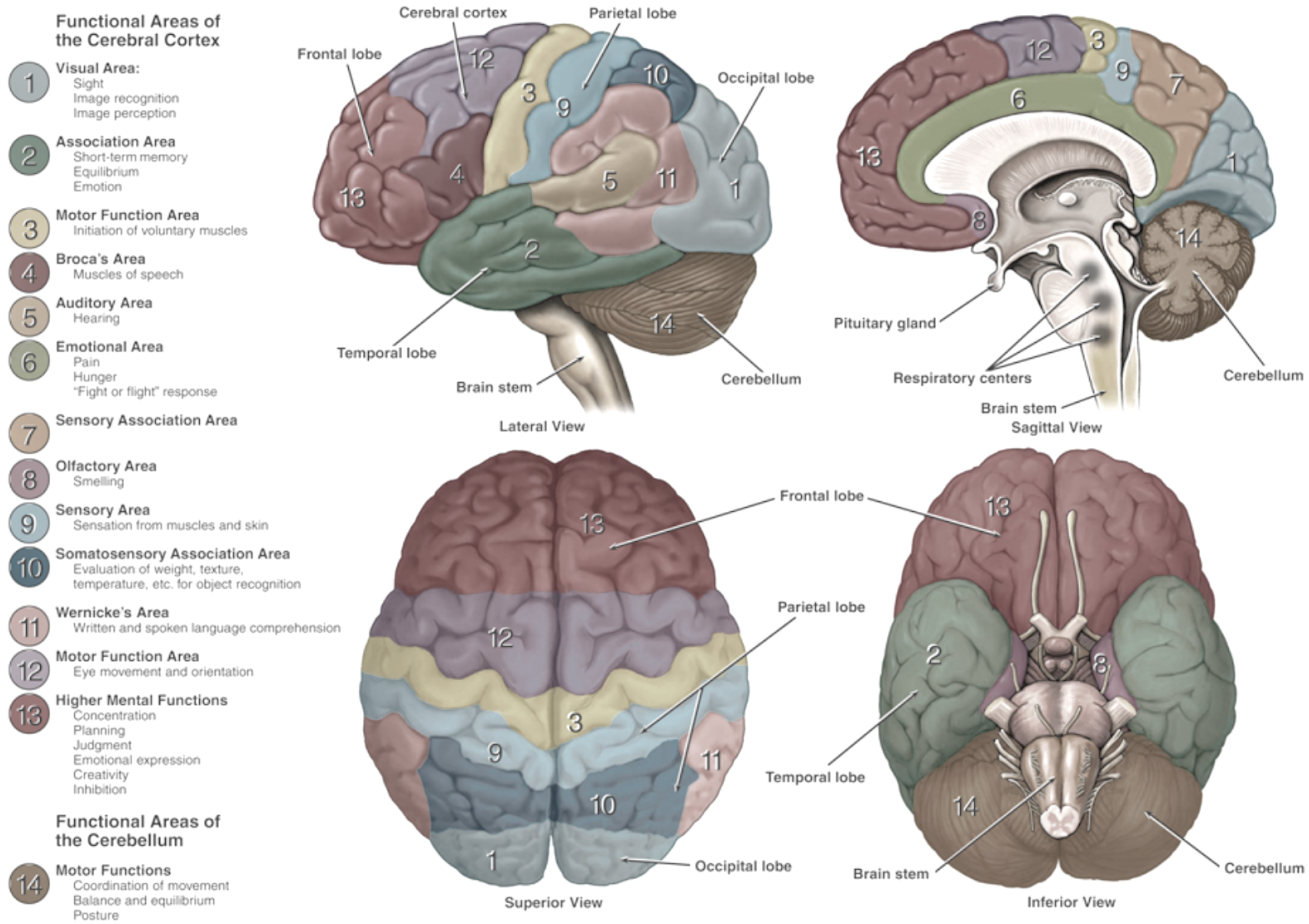


Fig. 3: Functional areas of the human cortex. *Source: dana.org.*

When looking at the cortex from the outside we can have the impression of a volume entirely filled with circonvolutions. This is wrong. If we could take the cortex off the brain and unfold it, we would end up with a very thin $2\,\mathrm{mm}$-thick sheet about the size of a $48\,\mathrm{cm}$ x $48\,\mathrm{cm}$ square. Basically a coin-thick dinner napkin. The explanation for the many folds and creases of the cortex when in our skull comes from evolution. Evolution built our brain outwards, adding newer and higher-level functions on top of pre-existing structures. Constrained by the maximum volume of our head compatible with a natural childbirth, the thin cortex grew by increasing its surface area by folding, ultimately creating the highly convoluted form we know today [NVVH07].

The brain is the organ of thought and mental abilities are localized in specific cortical areas. We now shift our focus from the where to the how is the brain governing our body. Let's zoom in and enter more deeply in the microscopic structure of the brain. Even though we figured out that the brain was the seat of our intelligence a long time ago, it wasn't until pretty recently (late 19th century with Golgi) that

science understood what the brain was made of: specialized cells called neurons. Neurons, like other cells, have a cell body (called the soma, where the nucleus is) but this body is extended by many short branching strands known as dendrites, and a separate one that is typically longer than the dendrites, known as the axon, with multiple terminals [Figure 4]. The axon terminals of a neuron connect with dentrites of other neurons at junctions called synapses [Ger+14]. Scientists realized that the neuron was the core unit in the vast communication network that makes up the brains and nervous systems of nearly all animals. But it wasn't until the 1950s that scientists worked out how neurons communicate with each other: with electricity.

The brain works on electricity [HP17] – all of our thoughts are generated through a network of neurons, that send signals to each other with the help of electrical currents. The more electrical signals, the more neuronal communication, which corresponds to more brain activity. Here's how it works. At all times, neurons send messages to other neurons at their synapses, mostly using chemical messengers called neurotransmitters. These chemicals, stored in vesicles, are released by the sending neuron's axon terminal in the narrow synaptic cleft - the tiny gap between neurons - and attach to specific receptors of the receiving neuron. This message pass through the neuron soma and, depending on the chemical, raise or lower its charge a little bit. But if enough chemicals are released to raise his charge over a certain threshold, then it triggers a pulse of electricity called an action potential: a brief reversal of the body's normal charge from negative to positive and then rapidly back down to his normal negative. An action potential lasts a few ms and moves at a few meters per second (very slow compared to the $300\,\mathrm{km/s}$ of light) without any variation of amplitude. We informally say that the neuron is 'firing'. This potential zips down the axon into the axon terminals which themselves touch several other neuron's dentrites at synapses. When the action potential reaches the axon's terminal, it causes them to release chemicals onto the other neuron's dentrites they're touching, which may or may not trigger an action potential in them. This is usually how information moves through the nervous system: the synapse converts a presynaptic electrical signal (the action potential) into a chemical signal release in the synaptic cleft, which itself is eventually transformed into a postsynaptic electrical signal. Sometimes, in situations

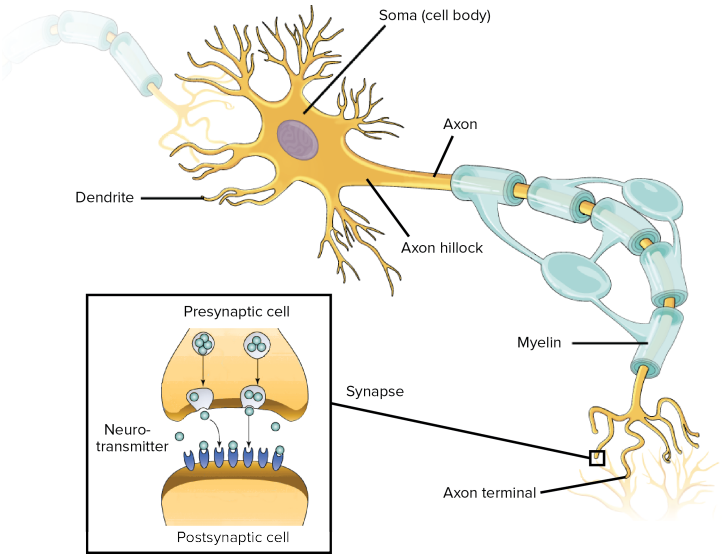


Fig. 4: Anatomy of a neuron.
Source: image modified from 'Neurons and glial cells: Figure 2' and 'Synapse' by OpenStax.

when the body needs to move a signal more quickly, neuron-to-neuron connections can themselves be electric, passing not through chemical but electrical synapses in which ions flow directly between cells.

The density of this network in the cortex is almost unthinkable: each $1\,\mathrm{mm}^3$ of cortex gray matter contains $50\,000$ neurons, each of them giving rise to 6000 synapses so totaling roughly $300\,\mathrm{M}$ synapses [NVVH07]. The thin convoluted cortex, constituting the bark of the brain, is called the grey matter in contrast with the space underneath, mostly occupied by wiring, the axons of cortical neurons, sheathed with a fatty white matter called myelin. We can think of the cortex as a command center that send many of its orders through the mass of axons making up the white matter beneath it. Now let's zoom out again to see the biggest picture.

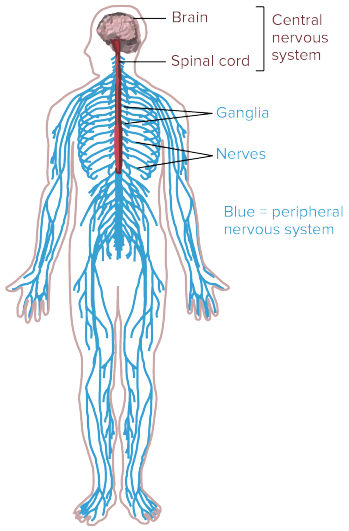


Fig. 5: Human central and peripheral nervous systems.
Source: modified from Wikipedia 'Nervous system diagram'.

The cortical axons of neurons in the brain might be taking information to either another part within the cortex, to the lower parts of the brain (brain stem or limbic system), or through the spinal cord (a massive bundle of axons composing the nervous system's superhighway) into the rest of the body. Indeed, the whole nervous system is divided into two parts: the central nervous system (the brain and spinal cord) and the peripheral nervous system (different types of neurons that radiate outwards from the spinal cord into the rest of the body). Bundles of axons of these neurons are wrapped together in a little cord called a nerve. Sensory nerves bring signals into the central nervous system, motor nerves carry signals out of it [Fac06].

Let's take an example to illustrate how these parts interact: when a fly touches our skin it stimulates many sensory nerves The axon terminals of the sensory neurons in the nerves start firing, send the signal into the spinal cord and up to the brain, more precisely in this case the somas in the somatosensory cortex. To trigger an action (chasing away the fly), the somasensory cortex then send action potentials to particular somas in the motor cortex that connect to the muscles in our arm that start firing, sending the signals back into the spinal cord and then out to the muscles of the arm. The axon terminals at the end of those neurons stimulate the arm muscles.

The scientific study of the brain has lead to remarkable advances since the middle of the twentieth century, both at the macroscopic level [NVVH07] (the major brain anatomical functions and structures) and at the microscopic level [DA01] (how a

	MRI	fMRI	MEG	EEG
Temporal resolution	Low	Low	High	High
Spatial resolution	High	High	Low	Low
Measures brain activity?	Only structure	Indirect	Direct	Direct
Level of expertise	Extensive training	Extensive training	Extensive training	Moderate training
Cost	Expensive	Expensive	Expensive	Accessible
Portability	Not portable	Not portable	Not portable	Fully portable

Tab. 1: Different non-invasive brain imaging techniques.

neuron fires). With hindsight, effort into localizing brain functions into distinct brain anatomical regions at a macroscopic level has not been very successful: only the most basic functions have been localized. So we had to look closer. At a microscopic level, we now know that our brain is composed of billions of neurons that communicate with each other mainly via synapses. This communication is based on the exchange of chemical substances between the neurons at the synapse and has the effect of producing electrical activity at their membranes. When neurons at a certain region activate together for some particular reason, their electric activity tend to synchronize and become measurable at a macroscopic scale. Neuroimaging, which captures this activity, helped us to shift from a view where every function is localized somewhere to a view of the brain as a network with patterns of communication between these regions. However, many questions remain unsolved, In particular we still have difficulties to grasp the middle level: how the brain builds representations of the reality and codes its sophisticated computation, like language, memory or mathematics with patterns of electrical activity. The fundamental question is: how the conscious subjective experience emerge from a neuronal network activity. How the cerebral states generate mental states that produce behaviour. We don't have the ambition to address those almost philosophical questions but one thing is sure: understanding the brain constitutes a major scientific challenge of our time. This challenge mainly relies on advanced techniques used to record the brain activity and extract a measurable signal, and then on signal processing tools used to interpret these recordings and hopefully deduce some useful information.

How to extract signals from the brain

To measure the brain activity, different brain imaging techniques are used. They are mainly characterized by the time scale of the measure (their temporal resolution), the accuracy of localizing the source of the activity (their spatial resolution) and their degree of invasiveness.

The main non-invasive measurement modalities are summarized in Table 1 and developed below:

MRI (Magnetic Resonance Imaging) [Haw+80; McR+17] uses a machine inducing a strong magnetic field in order to make the protons in the hydrogen atoms of the

water in our body to point in the same direction. Then it measures the energy emitted from the relaxation of protons to this aligned state when briefly disrupted by a radio pulse. This allows the computer to determine what the tissue looked like, depending on this energy that is released, and show an image of the tissue. MRI thus excels at isolating anatomical details, revealing the brain's structure and the different types of tissue present, like white and grey matter. This is the modality mostly used in literature to estimate the brain age. Yet, MRI only shows us a static anatomical image of the brain, not the brain's actual activity.

fMRI (Functional MRI) [Kwo+92; Log+01] uses the same mechanism than MRI to also measure the energy emitted from the relaxation of protons but this time aimed at determining the oxygenated blood flow changes in response to neural activity. The neuronal activation is therefore indirectly measured via local changes in the level of blood oxygenation, known as the BOLD (Blood-Oxygenation Level Dependent) response, with a limited temporal resolution (typically around $1\,\mathrm{s}$) due to slow changes of the blood flow. Nevertheless fMRI has a better spatial resolution now below $1\,\mathrm{mm}$, allowing to finely measure activity across different brain regions, enabling precise functional brain mappings.

EEG (electroencephalography) [Ber29; HP17] uses an array of electrodes on a cap placed on the scalp on a subject to directly measure the electrical activity of the brain. To facilitate comparisons between experiments, it is common practice to put the electrodes on standard positions. See Figure 2.4 for an example. The EEG amplitude mainly depends on the size of the active area as the voltage under each electrode is not the result of the electrical activity of a single neuron but instead a summed potential of populations of thousands of neurons. It also depends on the distance between the sources in the brain and the electrodes, taking into account the signal attenuation induced by the scalp. EEG signals typically are $50\,\mathrm{to}\,100\,\mu\mathrm{V}$ in amplitude, about $1\,\mathrm{M}$ times lower than voltages used to power home equipments, thus need to be amplified. Recordings of sufficient quality can nevertheless be performed in regular rooms and even in real-life settings using mobile EEG devices, with controlling head and body movements as they may cause artifacts. Thanks to its portability, EEG is operated in a wide array of peculiar situations, such as surgery [Bak+75], flying an aircraft [SS65] or sleeping [AJWW66]. For example EEG is used to diagnose pathologies for which the cerebral bioelectrical activity is susceptible to be perturbed, and especially to precise the location of cerebral tumors or different types of epilepsy and epileptic sources.

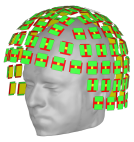
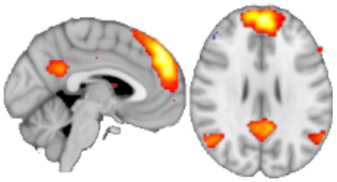
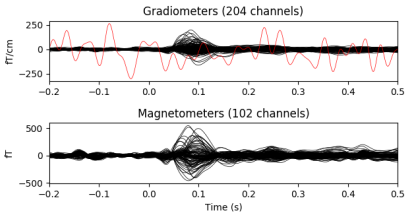
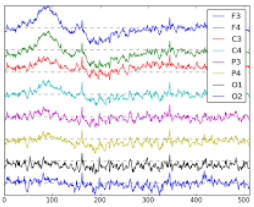
MEG (magnetoencephalography) [Häm+93; HP17] uses sensors to measure the magnetic field produced by the brain. Indeed, any electric current is associated with magnetic fields as a consequence of Maxwell's theory. Therefore, the brain generates tiny magnetic fields outside the head ($\sim 100\,\mathrm{fT}$) $10^{-8}\,\mathrm{t}$imes the strength of the earth's steady magnetic field, requiring very sensitive sensors and heavy noise cancellation. Their extreme sensitivity is challenged by many electromagnetic nuisance sources (any moving metal objects like cars or elevators) or electrically powered instruments,

generating magnetic induction that is orders of magnitude stronger than the brain's. The measurement itself is therefore done inside a special magnetically shielded room to dampen external ambient magnetic disturbances. Their influence can be further reduced by combining magnetometers coils (that directly measure the absolute magnitude of the magnetic field) with gradiometers coils (that record the gradient of the magnetic field in certain directions). Those gradiometers, arranged either in a radial or tangential (planar) way, record the gradient of the magnetic field towards 2 perpendicular directions hence inherently greatly emphasize brain signals with respect to environmental noise. Unlike EEG, MEG is not portable but captures a more selective set of brain sources with greater spectral and spatial definition [Ahl+10; HLR00], as the skull smears electrical but not magnetical signal.

More invasive techniques, using electrodes placed closer to the brain, are required to obtain both a good temporal and spatial resolution. Such techniques include **ECoG (electrocorticography)** [Pal06] which uses electrodes placed on the cortical surface below the skull and **LFP (local field potential)** [DCS99] which uses micro-electrodes directly placed inside the brain to record the electric potential in the extracellular space of the brain tissue. Small intracerebral electrodes are typically used to measure these potentials as opposed to large surface electrodes used in EEG, enabling measurement of more localized populations of neurons. These techniques provides extremely valuable recordings with excellent resolution and Signal-to-Noise Ratio (SNR) but are really invasive and offer a limited coverage of the brain.

Each of these neuroimaging modalities measure different aspects of brain function, hence provide unique windows into the brain, none of them being optimal on their own. The choice of the technique depends on the research question. As we want to extract biomarkers directly in the clinic we focus on non-invasive measurement modalities. If structural and functional details are necessary, then MRI or fMRI is a good choice if one is able to make the considerable investment required. For quick, affordable, and accessible insights about brain function, with a tight temporal resolution, EEG is the method of choice. For instance, a 4-channels EEG device is used in Lariboisière hospital in Paris to more easily monitor the depth of anaesthesia within the constraints of operating rooms. Typical signals extracted from these modalities are pictured in Table 2: MRI produce images, fMRI produce multivariate time series which are often visualized over an MRI image. EEG and MEG produce multivariate time series.

In this thesis, we will focus our attention on MEG and EEG modalities, which we will denote by M/EEG. Both methods rely on **electrophysiology**, the study of electrical properties of the biological cells and tissues, as they record the product of the electrical activity naturally occurring in the brain within the neurons (which gives rise to the magnetic fields outside the head recorded by MEG, and the electric

MRI	fMRI	MEG	EEG
			
			

Tab. 2: Illustration of different non-invasive measurements of brain activity. For each non-invasive brain imaging modalities (MRI, fMRI, MEG and EEG): typical devices used to record brain activity **(top row)** and corresponding extracted signal **(bottom row)**.
Sources (top row): MRI[image bank 123rf.com], fMRI[image bank 123rf.com], MEG [HP17], EEG[wikipedia] - Sources (bottom row): MRI[nicepng.com], fMRI [Var+10], MEG[MNE Python], EEG[Public BCI data Colorado State University]

currents on the scalp recorded by EEG). This activity is so small that only the synchronous activity of vast assemblies of neurons can be recorded.

Compared to the much younger techniques of MRI and fMRI, MEG and EEG have the advantage of directly measuring the neuronal activity. Using an array of very sensitive sensors positioned over the scalp, MEG and EEG deliver insight into the brain activity with high temporal but limited spatial resolution. Spatially, a fundamental assumption is that the activity recorded by M/EEG sensors at a given position may serve as a sign of brain activity at that given location. We can then try to infer which cognitive task a subject is performing just from the information coming from the EEG signals. For instance, it is known that when a human closes his eyes, the EEG signal in the occipital region oscillates at approximately $12\,\mathrm{Hz}$. Unfortunately, measuring electric activity in a given electrode does not necessarily mean that the region of the brain just underneath is active. This is because cortical current must go through several layers of brain tissue with different conductivity before attaining the scalp. As a consequence, at every spatial scalp position, the recorded activity is a mixture of all the underlying brain sources. This phenomenon is called volume conduction [NS05] and is mainly responsible for the poor spatial resolution of these techniques (around $2\,\mathrm{cm}$). Many works in the literature have investigated ways of inverting the volume conduction effect and recovering the activity at the brain level with spatial precision [DPm99]. We will later see in this thesis that this effect can even be bypassed without the need to invert it. On the other hand, M/EEG has a tight temporal resolution, allowing the detection of changes in brain activity in the order of milliseconds, with sampling rates between 250 and 2000 Hz in clinical

and research settings, making them extremely useful for extracting the temporal dynamics of brain activity.

To extract biomarkers from such heterogeneous multimodal brain data, the Machine Learning approach has recently received significant interest in clinical neuroscience [Woo+17].

How to predict from brain signals

Brain activity, when recorded on P sensors, produce signals that can be mathematically modelled as a multivariate time-series $\boldsymbol{x}(t) \in \mathbb{R}^P,\ t = 1 \dots T$. This signal contains both spatial information (at a particular time t_0, one record P values around the head, forming a random vector $\boldsymbol{x}(t_0) \in \mathbb{R}^P$), and temporal information (at each sensor located on a particular location k on the scalp, one record the variation of the signal across time, forming the univariate time-series $x_k(t) \in \mathbb{R}$). Typical number of time-samples is in the order of $T = 100\,000\,\mathrm{c}$orresponding to a few minutes of signal sampled at $1000\,\mathrm{Hz}$ and typical number of sensors ranges from $P = 10$ for clinical-grade EEG to 300 for research-grade MEG. This signal is therefore very high-dimensional: we need a few million data points to represent one M/EEG signal.

Once the M/EEG recordings are stored, and before analyzing the data, some signal preprocessing steps are carried out. A first important step is to filter artefacts, to avoid making conclusions about the brain activity based on elements that are not physiologically relevant. Artefacts commonly removed include: the spectral peak at 50 Hz due to the power line frequency, environmental artefacts and physiological artefacts (cardiac and ocular). A second common processing step is to bandpass filter the signal to some frequency interval carrying physiological information relevant for the analysis being done.

Let's suppose we want to predict a variable of interest y *e.g.,* a biomedical outcome, related to the brain activity $\boldsymbol{x}(t)$ through an unknown statistical relationship. It could be the health status (how sick one is), a physiological variable (the age) or a biomarker for any cognitive process. Due to the high dimensionality of the M/EEG signal, it is difficult for a human eye to quantify patterns in those brain data, especially for large quantity of data. One recent solution is to teach a computer to help automate the prediction, finding the most useful quantitative summaries in this wealth of data: this is the field of statistical learning, or Machine Learning (ML) [SSBD14]. ML algorithms when used for such a prediction task are designed to approximate the general relationship between y and $\boldsymbol{x}(t)$ using a dataset of examples: a series of recorded brain data $\boldsymbol{x}_i(t)$ and its corresponding target variable y_i for a lot of subjects $i = 1 \dots N$. After incorporating those examples (in the so-called training phase) the algorithm is able to predict the variable y from the brain data $\boldsymbol{x}(t)$ of *any* person (the generalization phase), not just the one it has seen during training.

When the prediction task aims at predicting a continuous variable ($y \in \mathbb{R}$) it is called a regression task, when it aims at predicting a categorical variable ($y \in$ finite set) it is called a classification task.

To map brain-behaviour the historical approach for clinical work was to use voxel by voxel classical statistical analysis: this is the realm of hypothesis testing, and multiple comparisons [Woo+17]. When we want to predict clinical endpoints from multiple brain signals (regression modeling) this is more efficiently done with a ML approach, that conveniently combine multiple inputs into a single prediction [Figure 6].

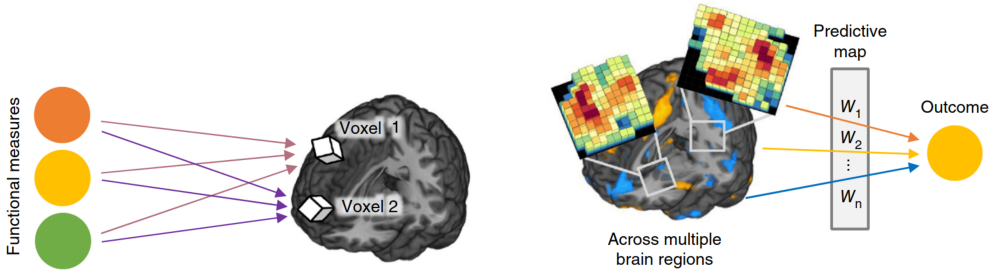


Fig. 6: Two different approaches to map brain signals to behaviour [Woo+17] **(Left)** Traditional brain mapping: Mass-univariate statistics. **(Right)** Predictive modeling: Combine multiple inputs into single prediction.

This approach has been successfully used to tackle both types of prediction tasks - classification [CR95; Nää75; PK95] and regression [Fru+17] - and increasingly easy to implement today thanks to readily available software packages. One of the most used packages world-wide is scikit-learn [Ped+11], developed in the Inria team "Parietal" in which I developed this thesis. It unfortunately comes with multiple caveats/challenges when used on clinical neuroscience data.

First, ML methods are designed to make good predictions, not to uncover the true probabilistic relationship between the target variable y and the predictor variables $\boldsymbol{x}$. They optimize an algorithm, fitting it to the data to minimize the expected prediction error on the population, not to uncover the true data generating mechanism. In other words ML is focussed on prediction, not inference [BI19]. It outputs a predictive model which can succeed to predict but fail to discover the data generating mechanism hence cannot be interpreted as a causal model.

Second, mathematical analysis of these algorithms shows that to perform well, *i.e.,* to generalize successfully from the examples seen in training to the general population, they need two main ingredients: lots of training examples (lots of data) and a prior knowledge about the data generating mechanism (some information about the unknown statistical relationship between $\boldsymbol{x}(t)$ and y, to guide the search of the predictor). Therefore ML effectiveness in psychiatry and neurology is mainly

constrained by the lack of large high-quality datasets [Var+17; Woo+17; Eng+18; Bzd17]. and comparably limited understanding about the data generating mechanisms [JK17]. This, potentially, limits the advantage of complex learning strategies. In clinical neuroscience, prediction can therefore be pragmatically approached with low-complexity classical machine learning algorithms [Dad+19] implementing simple learning strategies, expert-based feature engineering and increasing emphasis on surrogate tasks, for which dataset of examples are more easily found.

Regarding the features, many studies have shown the importance and predictive capabilities of the spectral content of M/EEG signal, *i.e.,* how it oscillates. This signal can indeed be decomposed in multiple simple waves or rhythms, characterized by their frequencies and amplitude [BD04; BL17].

In numerous experiments, where M/EEG activity of a subject was recorded while performing different cognitive tasks, it has been observed that the signal oscillate differently in different parts of the brain. M/EEG have an unparalleled capacity for capturing these brain rhythms without penetrating the skull [HP17].

Among them we can distinguish five types of rhythms or frequency bands:

- ▷ *delta* rhythm (frequency between $1\,\mathrm{Hz}$ to $4\,\mathrm{Hz}$; amplitude between $1\,\mu\mathrm{V}$ to $200\,\mu\mathrm{V}$): present in infants and in the deep state of sleep of adults, but that can convey serious cerebral suffering when present in the awaken adult [AYH18].
- ▷ *theta* rhythm (frequency between $4\,\mathrm{Hz}$ to $8\,\mathrm{Hz}$; amplitude between $150\,\mu\mathrm{V}$ to $200\,\mu\mathrm{V}$): rhythm of temporal and parietal regions, for example arising in children and adults in emotional conditions [AG01].
- ▷ *alpha* rhythm (frequency between $8\,\mathrm{Hz}$ to $12\,\mathrm{Hz}$; amplitude between $50\,\mu\mathrm{V}$ to $100\,\mu\mathrm{V}$): rhythm of the occipital region recorded on healthy awaken subjects, usually associated to a relaxed state of mind, *e.g.,* eyes closed in resting state. This rhythm mostly disappears when the subject opens his eyes or focus his attention on a mental activity and make way for the faster beta rhythm [Gol+02].
- ▷ *beta* rhythm (frequency between $12\,\mathrm{Hz}$ to $30\,\mathrm{Hz}$; amplitude between $10\,\mu\mathrm{V}$ to $50\,\mu\mathrm{V}$): rhythm originating in parietal and frontal regions associated to a normal state of consciousness [Pfu92].
- ▷ *gamma* rhythm (frequency between $30\,\mathrm{Hz}$ to $120\,\mathrm{Hz}$; amplitude between $2\,\mu\mathrm{V}$ to $10\,\mu\mathrm{V}$): associated with large scale brain network activity and cognitive phenomena such as working memory and attention. Altered gamma activity has been observed in many cognitive disorders such as Alzheimer's disease [VD+08].

Regarding the learning strategy, the gold standard method when predicting from M/EEG signals is source modeling, whereby a specific algorithm is used to find the

most probable sources in the brain that account for the recorded signal. This method, however, requires precise anatomical information provided by MRI scans. In the clinic, MRI recordings are rarely routinely available to do source reconstruction. Even when present in the hospital the machine is overloaded by patients that need it the most (not strictly necessary for knee surgery). An important question then is: when source localization is not available, and when we have some prior knowledge about the data generating mechanism, is there an optimal ML regression algorithm to predict from M/EEG signals, *i.e.,* an algorithm with perfect prediction in the limiting case of infinite number of samples? This important question is addressed by our first (methodological) contribution [Sab+19a] and will be investigated in Chapter 1.

Armed with this theoretically optimal algorithm to predict from our input data (the M/EEG brain signal), simple enough to be usable in the clinic, we will then focus on designing our surrogate task, the target of prediction (the y), a target that should be both easily available and a promising biomarker of neurocognitive disorders.

What to predict from brain signals: the brain age

Now that we have a clearer view on our input data (the M/EEG signal representing the brain activity) and that we found an optimal algorithm to predict from it, we focus our attention on the target of prediction.

In medicine, a biomarker is a measurable indicator of some disease state. For example, body temperature is a well-known biomarker for fever. Blood pressure is used to determine the risk of stroke. It is also widely known that cholesterol values are a biomarker and risk indicator for coronary and vascular disease. It can be discovered using genomics, proteomics technologies or imaging technologies. Our goal is to develop a biomarker of neurocognitive disorders through brain electrophysiology. Biomarkers are useful in a number of ways: they can help in early diagnosis, measuring the progress of a disease, evaluate most effective therapeutic regimes, prevent diseases, or identify drug target or drug response. A biomarker of a particular endpoint can be obtained by training a ML algorithm to accurately predict the endpoint. This training phase uses a dataset of patients for which we have both a brain signal and the corresponding endpoint [Par+15].

The gold-standard method to uncover risk factors and biomarkers in particular is large-scale population studies, generally based on meta-analyses or large biobanks [Cox+19]. When one can't afford the effort and the cost associated with them, we have to resort to experimental studies in clinical subgroups where ML can help in clinical diagnosis [Gau+19; Eng+18]. These studies that focus on clinical population are inevitably based on a limited number of patients, leading to small samples. Besides, as clinical data is rarely made public, meta-analyses are not

always possible. Those studies can therefore be statistically underpowered, and as a consequence often show optimistic biases in accuracy [PHV20; Woo+17].

To counter the scarcity of data samples of the precious clinical outcome (*e.g.,* in our case, cognitive decline), we adopted the alternative approach of designing a surrogate task: predict an endpoint that's widely available and then exploit its correlation with the actual endpoint of interest. As a surrogate variable we focused on age.

Our chronological age is determined by the number of years since our birth. But our body, our organs, our brain also have a biological age. Biological age could for instance be measured by looking at the integrity of the DNA in cells or by measuring the levels of proteins in the blood. Both chronological or biological ages are simple indicators of general health. Crucially, people with the same chronological age may have different biological ages. Individual-specific differences in their organs age reflect deviations from what is statistically expected and can be used to communicate risks [Spi16]. For example the bones age allows to identify growth pathologies between two children of the same age. Similarly, we can hope to be able to read out the age in the brain and that the age extracted from brain signals, the brain age, captures individual cerebral fragility.

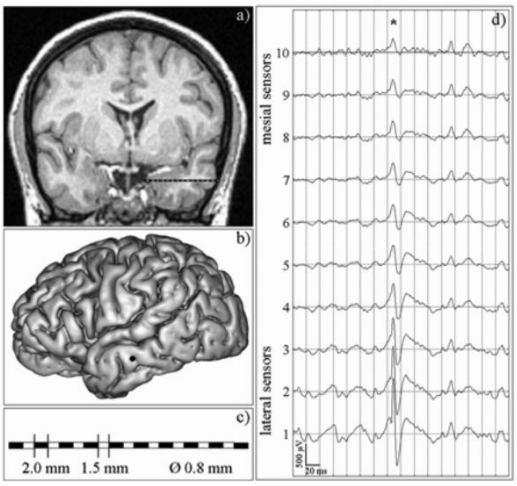


Fig. 7: How old is this brain? *Source: [CR+07].*

As a 70y old liver in a 50y person could provide hint of a chronic over-consumption of alcohol, an older brain could point to an undetected pathological brain aging. By definition, an healthy person should have a biological brain age similar to his chronological age. Our optimal ML model, previously designed, should then be first trained to accurately predict the chronological age of an healthy population. It would approximate age as a function of brain images. Given a new data point - a brain image - the function tells the expected age. This "prediction" expresses where the brain is positioned in the population, *e.g.,* whether that brain "looks" older or younger. The resulting measure brought by ML gives rise to brain predicted age as the solution to a regression problem from brain imaging, with more than 10 years of established literature [Dos+10]. It may seem irrelevant at first to predict the age as endpoint as there are very seldom situations where age is unknown. But we can hope that this brain predicted age contain information not present in chronological age. For instance, when computed

from fMRI data, it could capture volume reduction that comes with normal aging but could also reveal less volume than expected so captures pathological atrophy.

The simplest way to contrast the brain predicted age and chronological age is then to define the brain age delta ("BAD") as the arithmetic difference between predicted/biological age and passport/chronological age [Col+18] [Figure 8 (Upper left)]. Thus, the higher the BAD the 'older' the brain: a positive BAD denotes an 'older' brain, and negative BAD a younger brain. A key insight is that precocious brain aging (positive BAD, *i.e.,* high prediction error) reflects cognitive impairment, poor physical health and even risk of mortality [Lie+17] [Figure 8(Upper left and Lower)].

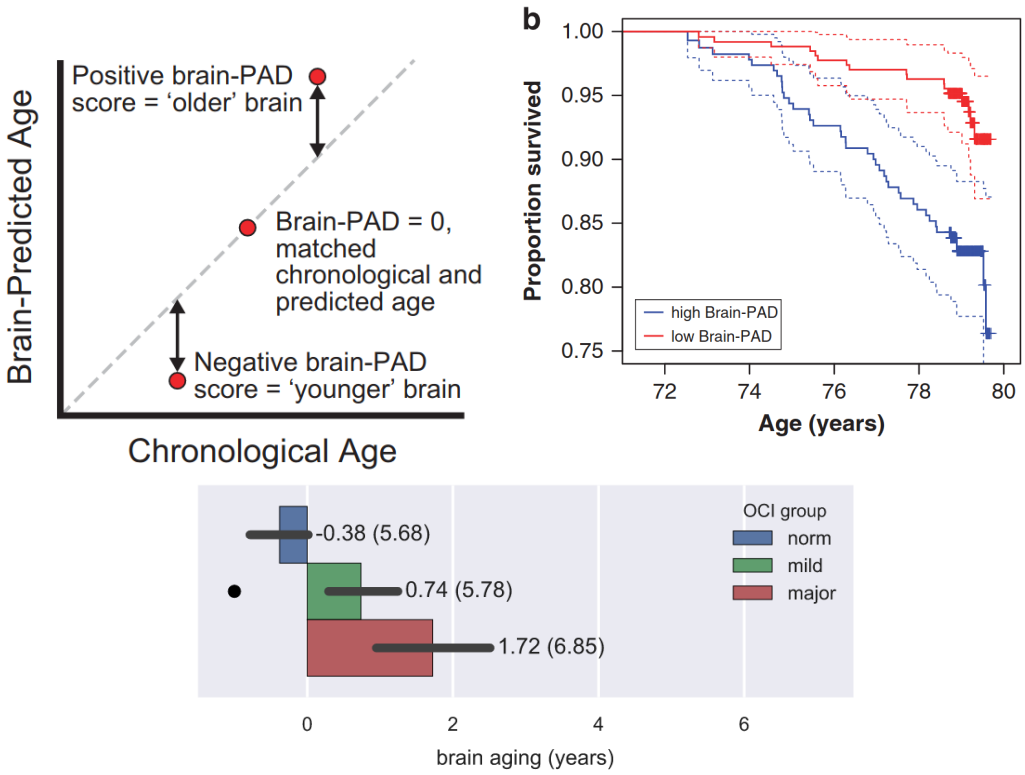


Fig. 8: Brain age delta as a surrogate biomarker of cognitive disorders. (Upper left) Defining the Brain-PAD, *i.e.,* the Brain age delta [Col+18] **(Upper right)** For the same chronological age, the survival rate is lower for high BAD, *i.e.,* older brain (in blue) compared to low BAD, *i.e.,* younger brain (in red) [Col+18] **(Lower)** Brain aging differences between objective cognitive impairment groups [Lie+17]

Thus, brain age delta does seem to contain information about pathological aging of the brain, hence can be interpreted as a (surrogate or proxy) biomarker of cognitive disorders. As counter-intuitive as it may sound we can obtain a measure of abnormal brain aging by choosing the age of healthy persons as our target of prediction from brain signals.

So we know what to predict (the brain age), how to predict (with our optimal ML model), but from which brain data? The BAD has been historically measured through MRI. Yet, we saw that other brain imaging modalities provide unique information about the brain. This raises the question of which brain imaging modality should our ML model use to compute brain age and which features are most informative about age. Let's imagine we are not in the clinic yet but in the comfortable conditions of a research laboratory where we have them all: MRI, fMRI and MEG.

How to estimate brain age in the lab

Brain biological age is typically estimated with MRI but can M/EEG be useful ? Until recently, most studies were dedicated to establish that M/EEG and MRI capture some similar information, for instance Brookes et al. [Bro+11] showed that fMRI resting state networks can be reconstructed from MEG, and Hipp and Siegel [HS15] that BOLD and MEG show similar spatial correlations across many frequency bands. We now have independent evidence that they also do carry independent information: Kumral et al. [Kum+19] showed that BOLD and EEG signal variability at rest differently relate to aging, Nentwich et al. [Nen+20] demonstrated that fMRI and EEG connectivity is different, Gaubert et al. [Gau+19] showed EEG-signatures in preclinical Alzheimer's disease.

Distinct features measured by all three techniques – MRI, fMRI and electrophysiology – have been associated with aging. For example, differences between younger and older people have been observed in the proportion of grey to white matter (through MRI), the communication between certain brain regions (through fMRI), and the intensity of neural activity in alpha band (through M/EEG). Literature on brain aging has historically focus on MRIs which, with their anatomical details, remain the go-to for predicting the biological age of the brain. But patterns of neuronal activity captured by electrophysiology also provide information about how well the brain is working. However, it remains unclear how electrophysiology could be combined with other brain imaging methods, like MRI and fMRI. Can data from these three techniques be combined to better predict brain age? We investigated this question in an article I co-authored [Eng+20].

We first trained our computer model with a subset of data from the Cam-CAN database, which holds MRI, fMRI, MEG and neuropsychological data for 650 healthy people aged between 17 and 90 years old. To handle the different modalities we used a computer model based on stacking: we first summarize the data in each modality with linear models (for which sample error grows only linearly with sample size) and then correct for bias of linear models with a non-linear Random Forest model [Eng+20].

We chose as baseline the model with standard anatomical MRI scans and compared different versions of the model with additional information MRI+fMRI+ MEG, MRI+MEG, MRI+fMRI. The Figure 9 depicts the mean-absolute-error (MAE) of these models relative to MRI only model (in blue, showing a relative difference of 0). We found that adding either MEG or fMRI to anatomical MRI led to a more accurate prediction of brain age. When both were added, the model was enhanced even further, with an absolute MAE of 4.7y. So we demonstrated that MEG contains unique, non-redundant information on age and cognitive aging vs fMRI.

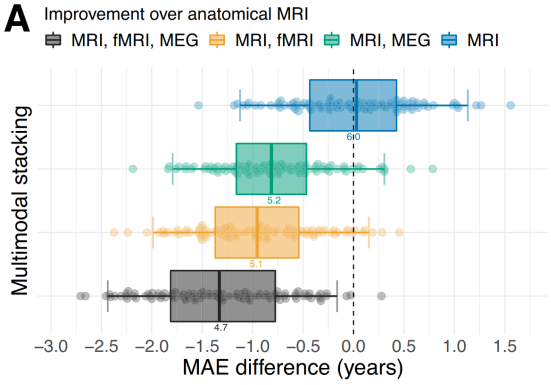


Fig. 9: Combining brain imaging modalities enhances brain age prediction. Mean Absolute Errors differences of models with different combinations of MRI, FMRI and MEG modalities compared to MRI only [Eng+20]

If combining multimodal brain data (MRI, fMRI, MEG) markedly improve brain age prediction performance, acquiring multiple modalities can be difficult in clinical practice, especially due to missing values. We showed that our tree-based algorithm would hold up if some data were missing. And we found that combining MEG, fMRI and MEG, even when some modalities were missing in some cases, was always better than using single modalities. Our tree-based methods bring flexible missing value handling.

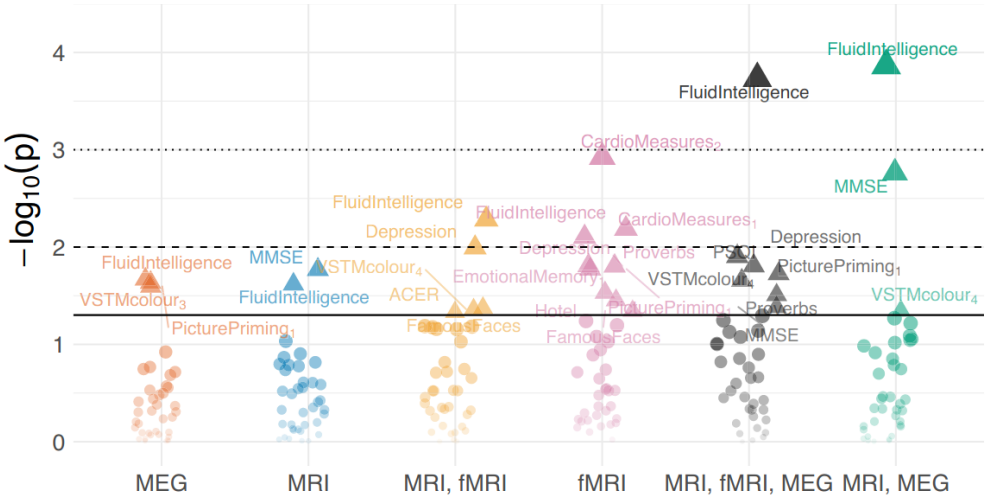


Fig. 10: Combining modalities improves the characterization of neurocognitive outcomes. Residual correlation between BAD and 38 neuropsychological assessments [Eng+20]

This flexible algorithm learnt better model of aging but is it relevant for neuropsychological score? We demonstrated that this combination also lead to enhanced characterization of neurocognitive phenotypes [Figure 10]: with unique discoveries or increased effect sizes associating the out-of-sample BAD with neurocognitive outcomes. The predictions correlated with the cognitive fitness of individuals. People with older brains tended to complain about the quality of their sleep and scored worse on memory and speed-thinking tasks. This suggests that BAD can be a good surrogate biomarker of cognitive aging and contains useful clinical information. Not only adding MEG boosts performance, but it also improves brain-behaviour correlation.

Moreover, when we focussed on models with only MEG as modality, we showed that MEG most influential features for brain age predicting modeling are source powers across frequency bands: In [Figure 11], light blue is the model taking all sorts of sensor-space features and dark blue/red/orange are models based on source-space features that were extracted by source localization. Linear combinations of source-power across frequency bands explains most of the MEG-specific performance. As stand-alone models, we found that source-level alpha and beta band power were most informative. Aligned with numerous findings in literature, we observed that MEG source power is a potent feature for predictive modeling. In the literature, MEG classical features comprises sensor alpha power [Eng+18], alpha peak [Bab+06], $1/f$ slopes [Voy+15], power topography [Gau+19; Fru+17], evoked latency [Pri+17], source powers [Sab+19a] and power envelope correlations [Kha+18].

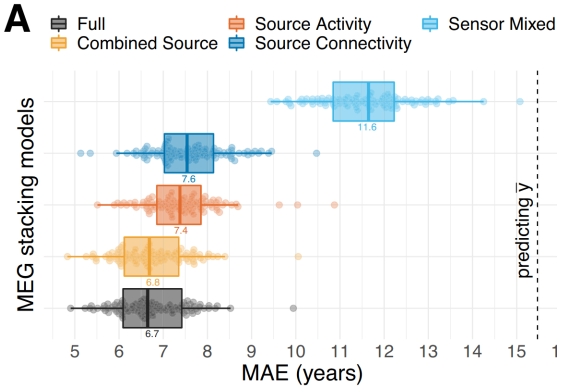


Fig. 11: Investigating most influential features to predict brain age from MEG. MEG performance is predominantly driven by source power [Eng+20]

This study gives us hope in our endeavour. It demonstrates that the BAD, used as an index of cognitive aging, contain useful clinical information and has the potential to be used as biomarker of neurocognitive disorders. BA is best predicted combining all modalities but predicting from MEG only leads to an acceptable performance (when used with source reconstruction). MEG seems to bring unique information with contribution best explained by cortical source power spectra, a feature that can also be accurately measured by EEG. This hints that EEG (used by most hospitals rather than MEG tests) could potentially be substituted for MEG without an impact on the predictive power of the model.

Unfortunately, it also suppose conditions not easily available in clinical practice, and certainly not compatible with a usage in the operating room. First, it requires to have MRI data: even the MEG only experiment relied on MEG features that require MRI acquisitions and tedious data processing to do source reconstruction Second, it uses research-grade high-fidelity MEG devices. Finally it requires highly-preprocessed MEG data. We will show in our first contribution in Chapter 1 that our proposed method can accommodate the absence of MRI data under certain ideal conditions but does it hold on real M/EEG data when those conditions are challenged? Is our method performant and robust enough to accommodate low-fidelity and low-preprocessed EEG measures (clinical-grade EEG vs research-grade MEG)? This will be investigated in our second (empirical) contribution [Sab+20] and described in Chapter 2.

We will see that our regression model has indeed the potential to be used in the clinic: it operates in sensor-space (avoiding costly source localization), it is robust to environmental and physiological artefacts and it accommodates cheap EEG recordings. This optimal, robust and light model is then a good candidate to develop our BAD biomarker. Now that we have a method to robustly determine brain age in the lab, the critical question is: does it translate to the clinic, is it really usable in the operating room?

How to translate brain age to the clinic

In the clinic, virtually all patients undergoing surgery go through a general anesthesia (GA). This procedure concerns millions of people every year: more than 300 million worldwide in 2020 [Csj]. If we include rachianesthesia and loco-regional procedures this number is even far greater. In France, $9.5\,\mathrm{M}$ of general anesthesia procedures have been performed in 2010 (excluding childbirth) with an average yearly increase rate of $1.89\,\%$ between 1991 and 2010 [Dad+15]. Since a precise physiological monitoring is required during GA, often including brain monitoring, this means that there exists a population-wide dataset of neural signals, today largely untapped.

Besides, the period of GA is a particularly favorable period to extract signals from patients with minimal artefact due to muscle-inhibitor drugs, hence a particularly adapted moment to build biomarkers. Moreover, EEG is already routinely used in the operating room during general anesthesia to monitor the depth of anesthesia as recommended by scholar societies of anaesthesiologists. Yet, despite its ubiquity, the wealth of physiological signals recordings including EEG and potential good signal quality, GA has never been used to estimate brain age.

Estimating brain measures from EEG during general anesthesia as a biomarker of neurocognitive disorders is worth investigating. The preliminary work I coauthored [Tou+20] showed a first evidence that EEG-based brain measures could be developed in the clinic under GA to reveal cognitive disorders. This prospective study demonstrated that EEG response to a GA based on propofol, a widely used drug to induce loss of consciousness, can reveal preoperative cognitive decline (CD). We collected EEG and propofol target concentration infusion (TCI) on 42 patients and used a preoperative cognitive assessment test called MoCA to detect CD. We focussed on two features: the propofol requirement to induce unconsciousness (the TCI needed to maintain the patient in a stable anesthesia state) and the frontal alpha band power during this period (AP). We assessed CD prediction based on 3 logistic based classification models including the age (AGE), TCI and AP (HELP1), and the three variables altogether (HELP2) [Figure 12]. The model HELP, including TCI and AP, better predicted CD than age and was more parsimonious than HELP2. Hence, TCI and AP contributes additively to reveal patient with preoperative cognitive decline.

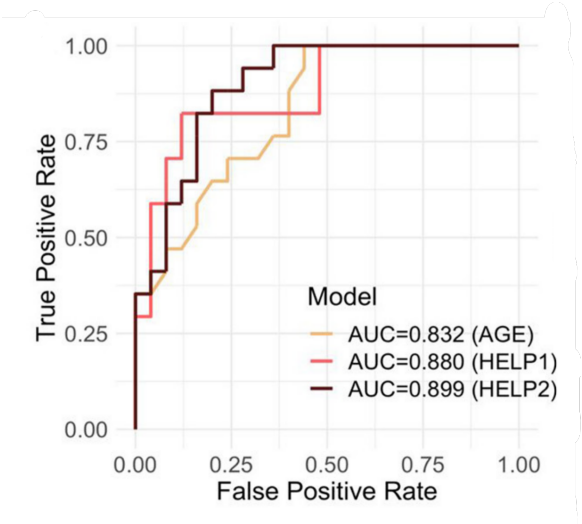


Fig. 12: Generalization performance of predicting occurrence of CD. Predicting CD from TCI and AP with 3 models including the age (AGE), TCI and AP (HELP1), and three variables altogether (HELP2), compared by their ROC curves and associated AUC [Tou+20].

To further investigate the brain age during GA we collaborated with anesthetists from the Lariboisière hospital in Paris and collected EEG of roughly 345 patients, together with associated clinical and demographical information. We first observed clear age-related patterns in the EEG recordings of those patients [Figure 13 (Left)]. Those EEG brain signatures nevertheless seem to be modified by the use of different maintenance drugs [Figure 13 (Right)]:

there seem to be systematic differences in alpha power under propofol and Sevoflurane drugs, especially in older men.

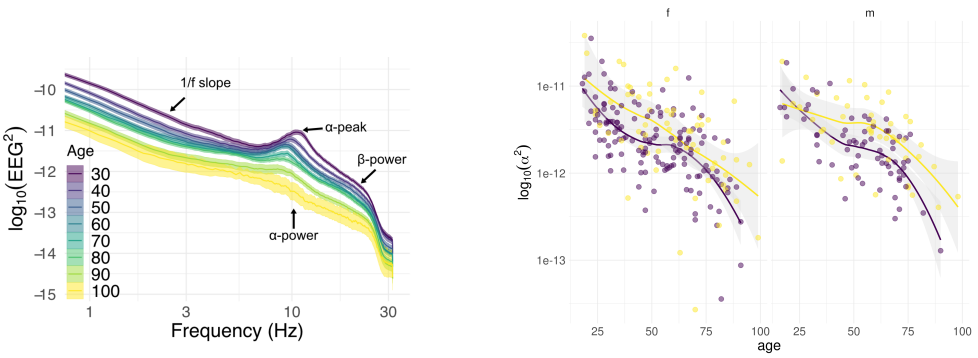


Fig. 13: Preliminary data exploration of a cohort of 345 patients from Lariboisière hospital, in stable GA state. (Left) Power Spectral Density of 345 patients during stable GA averaged and color-coded by age decade. **(Right)** Alpha band log-powers of the same patients vs age, when maintained in the stable anesthesia state using either propofol (purple) or Sevoflurane (yellow).

This rises the following questions. Can we predict the brain age in the clinic from EEG during anesthesia, *i.e.,* is the translation of lab-developed brain age valid in GA settings and does the drug impact brain age prediction under GA? How to perform EEG-based brain age prediction during anesthesia, taking the drug into account? Does Brain Age Delta (BAD) have a clinical meaning, *i.e.,* does this biomarker actually indexes cognitive disorders? These three questions will be investigated in third (clinical) contribution [Sab+21, in prep.] detailed in Chapter 3.

Thesis outline

Each of my three main contributions is designed to overcome a particular obstacle standing in the road to translation of brain age biomarker to the clinic:

- ▷ Absence of MRI / source localization: gold-standard regression models on brain signals rely on features that require source reconstruction, hence MRI acquisitions and tedious data processing. Overcoming this challenge and showing which regression model to use when source localization is not available constitutes my first (methodological) contribution [Sab+19a] detailed in Chapter 1
- ▷ Absence of research-grade brain imaging devices: adapting this model to real-world clinical M/EEG signals, low-fidelity devices and analyze the preprocessing impact on performance constitutes my second (empirical) contribution [Sab+20] in Chapter 2
- ▷ Specific conditions of GA: building out of this model a clinical biomarker of neurocognitive disorders usable in the operating room is covered in my third (clinical) contribution [Sab+21, in prep.] in Chapter 3.

Proposition

My own mathematical contributions, in the form of propositions, are denoted in boxes of this kind.

These results were presented at various national and international conferences (JDSE 2019 for which I received the best paper award, NeurIPS 2019, OHBM 2020) and have been accepted at two symposiums (VPH 2020, CompAge 2020) and two summer schools (AI4Health, DS3). I also co-authored three additional publications: the two papers detailed in this introduction [Tou+20; Eng+20] along with a recent benchmark paper on brain age [Eng+21] for which I contributed data analysis tools.

All numerical illustrations have been carried out on publicly available datasets: Cam-CAN [Tay+17], TUH [Har+14] and FieldTrip [Oos+11] with the exception of the unique GA dataset collected in Lariboisière hospital in Paris, exclusively for this thesis.

Finally, in order to foster reproducible research, Python and R code for all methods discussed in this thesis are available online on public repositories:

- ▷ `https://github.com/DavidSabbagh/NeurIPS19_manifold-regression-meeg/`
  Python code for the NeurIPS 2019 article [Sab+19a]- tools to preprocess raw MEG data from Cam-CAN dataset, vectorize covariance matrices, launch simulations and real-data analysis.
- ▷ `https://github.com/DavidSabbagh/meeg_power_regression`
  Python and R code for the NeuroImage 2020 article [Sab+20] - tools to preprocess raw MEG & EEG data from Cam-CAN, TUH and provided by the FieldTrip website, to analyze regression performance, inspect model by error-decomposition and assess pre-processing impact on performance.
- ▷ `https://github.com/DavidSabbagh/larib-EEG`
  Python and R code for the clinical article [Sab+21, in prep.] - tools to collect data and metadata from Lariboisière hospital, extract the features, explore the data, compare and inspect the regression models, perform data and statistical analysis.

We used the R-programming language and its ecosystem for visualizing the results [R C19; AUT19; Wic16; CSM17] and run part of the statistical analysis. Data analysis has been performed with Python 3.7 and only relies on open-source libraries: the Scikit-Learn software [Ped+11], the MNE software for processing M/EEG data [Gra+14], the PyRiemann package [CBA13] for manipulating Riemannian objects, and 'Coffeine' (`https://github.com/coffeine-labs/coffeine`) whom I developed the core features during my PhD and that provides a high-level interface to all predictive modeling techniques we present in this thesis.

1 Theory of power regression on sensor-space M/EEG with Riemannian Geometry

Contents

Mathematical notations used in the chapter

$\mathbb{Z}$	set of integer numbers
$x \in \mathbb{R}$	Scalar (lower case)
$\boldsymbol{x} \in \mathbb{R}^P$	Vector of size P (bold lower case)
$\|\boldsymbol{x}\|_2$	ℓ_2 norm of vector $\boldsymbol{x}$: $\sqrt{\sum_i x_i^2}$
$\boldsymbol{M}$	Matrix (bold uppercase)
$\boldsymbol{I}_N$	Identity matrix of size N
$[\cdot]^\top$	Transposition of a vector or a matrix
$\text{tr}(\boldsymbol{M})$	Trace of matrix $\boldsymbol{M}$
$\text{diag}(\boldsymbol{M})$	Diagonal of matrix $\boldsymbol{M}$
$\|\|\boldsymbol{M}\|\|_F$	Frobenius norm of matrix $\boldsymbol{M}$: $\sqrt{\text{Tr}(\boldsymbol{M}\boldsymbol{M}^\top)} = \sqrt{\sum \|M_{ij}\|^2}$
$\text{rank}(\boldsymbol{M})$	Rank of matrix $\boldsymbol{M}$
$\text{Upper}(\boldsymbol{M})$	upper triangular coefficients of $\boldsymbol{M}$, with unit weights on the diagonal and $\sqrt{2}$ weights on the off-diagonal
$\boldsymbol{S}$	cross-spectral density matrix
$\boldsymbol{C}$	spatial covariance matrix
$\mathcal{M}_P$	Space of $P \times P$ square real-valued matrices
$\mathcal{S}_P$	Subspace of $P \times P$ symmetric matrices: $\{\boldsymbol{M} \in \mathcal{M}_P, \boldsymbol{M}^\top = \boldsymbol{M}\}$
$\mathcal{S}_P^{++}$	Subspace of $P \times P$ symmetric positive definite matrices: $\{\boldsymbol{M} \in \mathcal{S}_P, \boldsymbol{x}^\top \boldsymbol{M} \boldsymbol{x} > 0, \forall \boldsymbol{x} \in \mathbb{R}^P, \boldsymbol{x} \neq 0\}$ $\boldsymbol{M}$ is full rank, invertible (with $\boldsymbol{M}^{-1} \in \mathcal{S}_P^{++}$) $\boldsymbol{M}$ is diagonalizable with real positive eigenvalues: $\boldsymbol{M} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^\top$ with $\boldsymbol{U}$ orthogonal matrix of eigenvectors ($\boldsymbol{U}\boldsymbol{U}^\top = \boldsymbol{I}_P$) and $\boldsymbol{\Lambda} = \text{diag}(\boldsymbol{\lambda})$ matrix of positive eigenvalues ($\lambda_1 \geq \cdots \geq \lambda_P > 0$)
$\mathcal{S}_P^{+}$	Subspace of $P \times P$ symmetric semi-definite positive (SPD) matrices: $\{\boldsymbol{S} \in \mathcal{S}_P, \boldsymbol{x}^\top S \boldsymbol{x} \geq 0, \forall \boldsymbol{x} \in \mathbb{R}^P\}$
$\mathcal{S}_{P,R}^{+}$	Subspace of SPD matrices of fixed rank R: $\{\boldsymbol{S} \in \mathcal{S}_P^{+}, \text{rank}(\boldsymbol{S}) = R\}$
$\log(\boldsymbol{M})$	Logarithm of matrix $\boldsymbol{M} \in \mathcal{S}_P^{++} : \boldsymbol{U}\ \text{diag}(\log(\lambda_1), \ldots, \log(\lambda_n))\ \boldsymbol{U}^\top \in \mathcal{S}_P$
$\exp(\boldsymbol{M})$	Exponential of matrix $\boldsymbol{M} \in \mathcal{S}_P : \boldsymbol{U}\ \text{diag}(\exp(\lambda_1), \ldots, \exp(\lambda_n))\ \boldsymbol{U}^\top \in \mathcal{S}_P^{++}$
$\mathcal{N}(\mu, \sigma^2)$	Normal (Gaussian) distribution of mean μ and variance σ^2
$\mathbb{E}_s[\boldsymbol{x}]$	Expectation of any random variable $\boldsymbol{x}$ *w.r.t.* its subscript s when needed
$(\mathcal{Z}, \mathcal{F}, l)$	task components: set of task objects, set of potential solutions, objective function
S	sample $(z_1, \ldots, z_n) \in \mathcal{Z}^n$
$\mathcal{D}^n$	distribution of sample S
$\mathcal{H}$	hypothesis class
$\mathcal{T}_{\boldsymbol{M}}$	tangent space at point $\boldsymbol{M}$

Acronyms used in the chapter

BCI	brain-computer interface
DTFT	discrete-time Fourier transform
ERM	empirical risk minimization
ERP	event-related potential
EOG	electro-oculogram
ECG	electro-cardiogram
fMRI	functional magnetic resonance imaging
M/EEG	magneto- and electroencephalography
ML	machine learning
MAE	mean absolute error
MSE	mean squared error
MNE	mnimium norm estimate
MRI	magnetic resonance imaging
PAC	probably approximately correct
PSD	power spectral density
PCA	principal component analysis
SPD	symmetric positive definite
SPoC	source power comodulation
SSS	Signal Space Separation
SSP	Signal Space Projection
WSS	wide-sense stationary

In the Introduction, we investigated how the brain is structured, how it operates and how to capture its activity by extracting a measurable signal $\boldsymbol{x}(t)$. In this chapter, we now seek a regression algorithm to predict, from this brain signal as input data, any outcome y physiologically related to neural dynamics. This is a key step to develop a predictive clinical biomarker. This algorithm should be both performant, hence adapted to the particular class of signals it works on, and simple enough to be usable in the clinic, which discard most of the classical algorithms. This chapter presents such an algorithm as my first contribution [Sab+19a], along with the theoretical foundations on which it rely.

Section 1.1 summarizes the Machine Learning (ML) theoretical framework to perform prediction. We will investigate the different sources of errors of any prediction algorithm and show that a performing predictive biomarker should be based on a statistically consistent algorithm. Such algorithm has to be adapted to the probability distribution of its input hence require some form of prior knowledge about the data generating mechanism (No-Free-Lunch theorem). Our data signal is nevertheless not well adapted to classical regression algorithms that require low dimensional inputs (to fight curse of dimensionality) that live in a vector space. Armed with these theoretical insights, the following sections investigate how to meet these requirements.

Section 1.2 will allow us to find such a low dimensional compact representation of the M/EEG signal: its spatial covariance matrix. It introduces statistical tools for the analysis of multivariate time series and discuss two fundamental assumptions that are typically done regarding their statistics: wide-sense stationarity and ergodicity. These assumptions allow to estimate a set of parameters that describe the statistical behaviour of a real multivariate time series: its mean vector and its cross-spectral density matrices, a continuous set of positive definite matrices. This set can be further summarized by using one of such matrices: the band-specific covariance matrix. One may then approximately compare two time series by comparing the covariance matrices used to parametrize them. These matrices are most often rank-deficient due to practical reasons investigated at the end of this section.

Armed with this compact matrix representation, we'll still need to vectorize it. The set of covariance matrices, either full rank or not, is known to have a particular intrinsic geometry - they live on a Riemmanian manifold - and in Section 1.3 we give an overview of its properties. Understanding the intrinsic geometry of a set of data points very often lead to more efficient algorithms, *e.g.,* deep learning methods adapted to data defined in a manifold [Bro+17], or text classification using Riemannian geometry [Leb05]. This is particularly true in the field of brain-computer interfaces (BCI) where classification methods have been largely improved using such geometry-aware algorithms [Lot+18; CBB17; YBL17]. Here, using concepts from differential geometry, we will present a distance between positive definite matrices

that is invariant to affine-invariant transformations (e.g. the action of a matrix), a very useful property when parametrizing multivariate time series. This will lead to an appropriate vector representation of a covariance matrix.

Section 1.4 introduces a prior knowledge about our task in a form of a generative model of M/EEG signals and outcome and will detail state-of-the-art approaches to predict from such M/EEG signals. We will delve into three classical family of methods, well-suited in certain contexts but that are unsatisfying for the present task. Biophysical source modeling using anatomically constrained inverse methods needs anatomical prior knowledge so requires specialized workforce and equipment, which may reduce wider applicability in clinical practice. Statistical sources modeling using unsupervised spatial filtering (*e.g.,* ICA) are blind to the prediction target hence requires additional modeling efforts for subsequent regression tasks. Sensor-space linear modeling leverages the power of linear models but is not optimal when predicting from brain rhythms, hence not adapted to our generative models assumptions. We will show that none of them is suitable to our task, hence the need for a new algorithm.

Finally, Section 1.5 introduces statistically consistent regression algorithms to predict from M/EEG signals, adapted to our generative model and usable in the clinic. We provide mathematical proof of consistency under certain conditions. We also present simulations to illustrate these mathematical guarantees and investigate the robustness of our algorithms to model violations. One of these model relies on concepts from Riemannian geometry and can easily be adapted to handle rank-deficient covariances.

Simulations in this chapter and real-world experiments in the next two were carried out using Python libraries including: `scikit-learn` [Ped+11], `MNE-python` [Gra+14] and `pyRiemann`[1]. We also ported to Python some part of the Matlab code of Manopt toolbox [Bou+14] for computations involving Wasserstein distance. The scripts generating associated figures are available in the `GitHub` repositories:

- ▷ https://github.com/DavidSabbagh/NeurIPS19_manifold-regression-meeg/
- ▷ https://github.com/DavidSabbagh/meeg_power_regression

Every code runs on a standard laptop.

[1] `http://pyriemann.readthedocs.io/`

1.1 Statistical Learning theory

In this section, we will present the mathematical framework of Statistical Learning, or Machine Learning (ML). For this introduction we adapted concepts and partially borrowed notations from [SSBD14].

1.1.1 Learning a task

Real-world task. Let's assume we face with a real-world task we'd like to solve. This task could be: compute the area of circle of radius r, add two big integer numbers, decide which animal portrays an image, predict breast cancer from mammography, predict cognitive function from M/EEG brain activity, or clustering active EEG electrodes that record task-relevant neurophysiological activity. We will see that these tasks, very different in nature, can share a common formalism but are tackled using different approaches. The very first step to solve the task is to model it mathematically.

Model the task. We represent the task by a triplet:

$$\mathcal{T} = (\mathcal{Z}, \mathcal{F}, l) \ , \tag{1.1}$$

where $\mathcal{Z}$ is the set of task's objects, called the instance set, formally a measurable space; $\mathcal{F}$ is the set of all potential solutions, formally a set of functions; l is the error made by a solution on an object, called a loss or a cost, formally a function $\mathcal{F} \times \mathcal{Z} \mapsto \mathbb{R}$.

For example *prediction* is the task of predicting an outcome y from a statistically related input $\boldsymbol{x} \in \mathbb{R}^P$. When y is a continuous (*resp.* discrete) variable, the prediction task is called regression (*resp.* classification). Hence, a regression task operates on pairs of variables $(\boldsymbol{x}, y) \in (\mathcal{X}, \mathcal{Y})$ as objects, with $\boldsymbol{x}$ observed and y unobserved, and looks for the function $f : \mathcal{X} \mapsto \mathcal{Y}$ that best predicts y given $\boldsymbol{x}$, where 'best' may mean 'with minimal quadratic error'. Therefore, this task can be modeled by the triplet $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ a bounded subset of $\mathbb{R}^P \times \mathbb{R}$, $\mathcal{F}$ the set of all bounded functions from $\mathcal{X}$ to $\mathcal{Y}$ and $l(f, (x, y)) = (f(x) - y)^2$, the squared loss. Many tasks can be formalized this way. Some examples are detailed in table 1.1:

Task	$\mathcal{Z}$	$\mathcal{F}$	l	
Prediction:Regression	$\mathcal{X} \times \mathcal{Y}$ bounded subset of $\mathbb{R}^P \times \mathbb{R}$	$\{f : \mathcal{X} \mapsto \mathcal{Y}$ bounded$\}$	$l(f, (x, y)) = (f(x) - y)^2$	squared loss
Prediction:Classification	$\mathcal{X} \times \{0, 1\}$	$\{f : \mathcal{X} \mapsto \{0, 1\}\}$	$l(f, (x, y)) = \mathbb{1}_{\{f(x) \neq y\}}$	0-1 loss
Large margin classification	$\mathcal{X} \times \{0, 1\}$ bounded subset of RKHS	bounded subset of RKHS	$\max(0, 1 - y < x, h >)$	hinge loss
K-means clustering	$\mathbb{R}^P$	all subsets of $\mathbb{R}^P$ of size K	$l(f, z) = \min_{c \in f} \|c - z\|^2$	dist.to nearest centroid
Density estimation	$\subset \mathbb{R}^P$	{bounded pdf on $\mathcal{Z}$}	$l(f, z) = -\log(f(z))$	negative log-likelihood

Tab. 1.1: Example of tasks

The task we want to solve in this thesis is a particular regression task: predict a continuous neuro-outcome $y \in \mathbb{R}$ from a multivariate M/EEG signal $\boldsymbol{x}(t) \in \mathbb{R}^P$. The content of this section is nevertheless very general and apply to any kind of task that can be formalized this way.

Solving the task. The first approach is to try to solve it *exactly*, *i.e.,* find a solution that never make any mistake for all objects $z \in \mathcal{Z}$. This exact solution can either be analytical or algorithmic. For example the task of computing the area of a circle of radius r can be solved exactly analytically using the mathematical tool of integration and yield $f^*(r) = \pi r^2$. The task of adding two big integer numbers cannot be solved analytically but has an algorithmic solution, actually the first algorithm most kids learn in first grade. Yet, some tasks don't have any analytic nor algorithmic solutions. This is generally the case either because they must adapt to fluctuating environments (changes over time or over users) or because they are too complex to program, being specific to human capabilities (no good known algorithmic solution) or beyond human capabilities (analysis of very large and complex dataset). Examples of such tasks are: deciding which animal portrays an image [LNH09] (multi-class classification task), predict breast cancer from mammography [Bar+06] (binary classification task), predict age from M/EEG brain activityi [Sab+20] (regression task), or clustering active EEG electrodes that record task-relevant neurophysiological activity [Sab+19b] (clustering task).

When a task cannot be solved exactly, either with an analytic formula or a predefined algorithm, we have to look for *approximate* solutions, allowing ourselves to make errors from time to time while trying to be maximally right on average. In this approach, we represent the class of objects by a random variable $\boldsymbol{Z} \in \mathcal{Z}$ following an unknown probability distribution $\mathcal{D}$ that reflect the common properties of those objects and then look for the solution with smallest error on average:

$$\boldsymbol{Z} \in \mathcal{Z} \sim \mathcal{D} \tag{1.2}$$

$$f^* = \underset{f \in \mathcal{F}}{\arg\min}\ L(f) \quad \text{with} \quad L(f) = \mathbb{E}_{\boldsymbol{Z} \simeq \mathcal{D}}[l(f, \boldsymbol{Z})] \tag{1.3}$$

For a regression problem, $\mathcal{D}$ is the joint probability distribution of $z = (\boldsymbol{x}, y)$ that describes, in its most general form, how both variables are statistically related. The distribution $\mathcal{D}$ is unique to the particular task we want to solve. For instance the joint probability distribution of (M/EEG signal, age) is of course very different from the one describing (mammography, breast cancer indicator).

Even if we know $\mathcal{D}$, finding f^* in general, for any task $\mathcal{T}$, is NP-hard, hence a hopeless endeavour. Yet, in this case, certain tasks have an explicit solution. For instance, the regression task with squared loss $l(f, (x, y)) = (f(x) - y)^2$ has the solution $f^*(x) = \mathbb{E}[Y|X = x]$, the regression tasks with absolute loss $l(f, (x, y)) = |f(x) - y|$

is solved by $f^*(x) = \mathbb{M}\text{edian}[Y|X = x]$, the binary classification task with 0-1 loss $l(f,(x,y)) = \mathbb{1}_{\{f(x)\neq y\}}$ by $f^*(x) = \mathbb{1}_{\mathbb{P}[Y=1|X=x]>1/2}$. This optimal solution f^*, called the Bayes solution, achieves the minimal possible error that can be achieved on this task: for this it needs to access the oracle, *i.e.,* know $\mathcal{D}$, to compute the expectation. Note that most often $L(f^*) \neq 0$ so even this optimal solution makes an error, called the irreducible error. In the literature, $L(f)$ is called the true risk or the true error or the generalization error of solution f and is a measure of its performance.

Since $\mathcal{D}$ is unknown, we can't compute L, nor the optimal solution f^* hence can't solve $\mathcal{T}$. Yet, if we have a sample from $\mathcal{D}$, we can still hope to learn it.

(PAC-)Learning a task. If we can perceive the world $\mathcal{D}$ through a random and finite sample

$$S = (z_1, \ldots, z_n) \sim \mathcal{D}^n \quad , \tag{1.4}$$

i.e., access a realization of n i.i.d. random variables drawn from $\mathcal{D}$, then we can hope to benefit from that limited experience of the unknown world $\mathcal{D}$ to learn the task: improve our performance at task $\mathcal{T}$ (lower L) with more experience (larger n). Learning the task $(\mathcal{T},\mathcal{D})$ therefore amounts to find a function $A : S \mapsto \mathcal{F}$, called a learning rule, that uses the sample S to output an hypothesis $h_S \in \mathcal{F}$ that is, with high probability, arbitrarily close to the optimal (Bayes) solution f^* with enough samples:

$$h_S = A(S) \in \mathcal{F} \quad s.t. \quad L(h_S) \simeq L(f^*) \tag{1.5}$$

One example of learning rule is the so-called Empirical Risk Minimization (ERM) that finds the hypothesis that minimizes the (computable) empirical risk: $h_S = \text{ERM}_{\mathcal{F}}(S) = \arg\min_{h\in\mathcal{F}} L_S(h)$ with $L_S(h) = \mathbb{E}_{\boldsymbol{Z}\sim\text{data}}[l(h,\boldsymbol{Z})] = \frac{1}{n}\sum_i l(h,z_i)$. Other learning rules are widely used: SRM (Structural Risk Minimization) that specifies weights over subsets of hypotheses of $\mathcal{H}$ reflecting preferences over some subclass of hypotheses, RLM (Regularized Loss Minimization) that jointly minimizes the empirical risk and a regularization function, MDL (Minimum Description Length) where hypotheses with shorter descriptions are preferred or SGD (Stochastic Gradient Descent) that directly minimizes, under certain conditions, the risk function $L(h)$ and not an intermediate objective function $L_S(h)$ without the need to know $\mathcal{D}$.

Since we face a task $\mathcal{T}$, with an unknown distribution $\mathcal{D}$, we would like to learn $\mathcal{T}$ for any distribution $\mathcal{D}$. Is it possible? More precisely, given a task $\mathcal{T}$, can we find a learning rule A and a sample size n such that for every distribution $\mathcal{D}$, if A receives n i.i.d. examples from $\mathcal{D}$, it outputs a solution h_S that has close-to-minimal error with high probability. Unfortunately, the No-Free-Lunch theorem states that no such universal learner exists [SSBD14].

No-Free-Lunch theorem. For every learning rule A there exists a distribution $\mathcal{D}$ on which it fails to learn $\mathcal{T}$ (it outputs an hypothesis h_S likely to have a large error),

whereas for the same distribution, there exists another learner that will succeed (it outputs a hypothesis with close to minimal error). In other words, the theorem states that no learner can succeed on all learnable tasks (on all distribution $\mathcal{D}$). Every learner has tasks on which it fails while other learners succeed: the learner has to be specified to the task at hand. Therefore, when approaching a particular learning task $(\mathcal{T}, \mathcal{D})$, the learner should have some prior knowledge on the distribution $\mathcal{D}$ in order to succeed. One type of such prior knowledge is that $\mathcal{D}$ comes from some specific parametric family of distributions. This is the realm of Maximum Likelihood parameter estimation, and is not the approach taken in this thesis. Indeed, our objective is to make distribution-agnostic predictions, not to uncover the true data generating mechanism that requires to make an appropriate assumption on a particular probabilistic model for the data.

Another approach is to adapt the learning rule to the task at hand to avoid the distributions that will cause us to fail when learning that task. The fundamental problem of any learning rule presented above is that its search space, the class of all functions $\mathcal{F}$, is 'too big': every possible function is considered a good candidate. This represents lack of prior knowledge. According to the No-Free-Lunch theorem it will fail on some learning task. For instance the ERM learning rule can find a function that has no error on the sample $L_S(h_S) = 0$ (because it can output arbitrary complex functions) but high error on the population $L(h_S) \simeq 1/2$ (high generalization error). This phenomenon is called overfitting: the learning rule has been misled by the training data. The idea is then to restrict the search space $\mathcal{F}$ of our learner to a predefined hypothesis class $\mathcal{H} \subset \mathcal{F}$ and introduce the best hypothesis within this class:

$$h^* = \underset{h \in \mathcal{H}}{\arg\min}\ L(h)\ . \tag{1.6}$$

We say that we learn the hypothesis class $\mathcal{H}$ if we find a learning rule that outputs an hypothesis $h_S \in \mathcal{H}$ that can be, with high probability, arbitrarily close to the best hypothesis in $\mathcal{H}$, with enough samples, and for any distribution $\mathcal{D}$:

$$h_S = A(S) \in \mathcal{H} \quad s.t. \quad L(h_S) \simeq L(h^*) \quad \forall \mathcal{D} \tag{1.7}$$

In this case, we have learnt the task, *i.e.,* found $h_S \quad s.t.\ L(h_S) \simeq L(f^*)$, only if $L(h^*) \simeq L(f^*)$. In a sense, this approach is another type of prior knowledge: the belief that one of the members of the hypothesis class $\mathcal{H}$ is a low-error model for our task, *i.e.,* the best hypothesis h^* in $\mathcal{H}$ is close to the best solution f^* in $\mathcal{F}$, the optimal Bayes solution.

Note that we can never be sure to exactly find the best solution in $\mathcal{H}$ $((L(h_S) = L(h^*))$ because we only access the world $\mathcal{D}$ through a sample S that is 1/ random (so they will always be a chance that S is not representative of $\mathcal{D}$ *e.g.,* if a domain point is sampled over and over again) and 2/ finite (even if S is representative there may

always be some details of $\mathcal{D}$ it will fail to reflect). We can therefore only hope that there exists a sample size above which we can find a Probably (with confidence $1-\delta$) Approximately (up to an error of ϵ) Correct solution h_S, a PAC solution, *i.e.,* an hypothesis that probably approximately has the minimal possible error in $\mathcal{H}$:

$$\forall \mathcal{D} \quad \forall \epsilon, \delta \quad \mathbb{P}\left[|L(h_S) - L(h^*)| < \epsilon\right] \geq 1 - \delta \tag{1.8}$$

with ϵ being the accuracy parameter that determines how far we are from optimum (we forgive the learner to make small mistakes) and $1-\delta$ being the confidence parameter that determines how likely we meet the accuracy requirement. If this holds, *i.e.,* if we can, with enough samples, come arbitrarily close to the best solution in $\mathcal{H}$ with high probability for any distribution $\mathcal{D}$, we say that the learning rule A learns $\mathcal{H}$ or that $\mathcal{H}$ is PAC-learnable with learner A. The amount of samples necessary to reach a given accuracy and confidence is called the *sample complexity* of learning. The No-Free-Lunch theorem precisely states that the class of all functions $\mathcal{F}$ is not PAC-learnable.

Learning algorithm. To successfully learn a task, we need a complete learning algorithm, composed of: an hypothesis class $\mathcal{H}$ on which we restrict the search space of the learning rule, a learning rule A that uses the sample S to choose a function in $\mathcal{H}$ (often the minimizer of some computable objective function L_S), and an optimization algorithm to actually compute this chosen function.

For example, the linear regression algorithm consists in using the ERM learning rule over the hypothesis class of linear functions, using no optimization algorithm since we have an analytic formula for the minimum. The LASSO (resp. Ridge) regression algorithm consists in using the RLM learning rule with a l_1 (resp. l_2)-regularization function to learn linear functions. Support Vector Machine algorithm uses the RLM scheme to learn linear functions using an iterative optimization algorithm. Decision Tree algorithms use the MDL learning rule to learn the class of decision trees. Neural Networks uses the SGD learning rule to learn a class of function defined by multiple layers of linear functions composed over non-linear activation functions. Note that, in general, this class of function $\mathcal{H}$ can be defined as the composition of functions that pre-process the data, functions that transform the data and functions that actually compute the solution. Most of classical regression algorithms operates on objects in $\mathcal{Z} = (\mathcal{X}, \mathcal{Y}) \subset \mathbb{R}^n \times \mathbb{R}$, *i.e.,* on Euclidean objects.

We have seen that a learning rule is not enough to learn the task, we must also restrict its search space to an hypothesis class. But how should we choose a good hypothesis class? On the one hand, we want to believe that the smallest error achievable by a hypothesis from this class is close to the smallest error achievable on the task. On the other hand, we have just seen that we cannot simply choose

the richest class – the class of all functions over the given domain. This trade-off is discussed in the following section.

1.1.2 Performance of a learning algorithm

The generalization error of a learning algorithm can be decomposed into three errors of very different nature:

$$L(h_S) = \underbrace{L(f^*)}_{\text{bayes error}} + \underbrace{L(h^*) - L(f^*)}_{\text{approximation error}} + \underbrace{L(h_S) - L(h^*)}_{\text{estimation error} \geq 0} \tag{1.9}$$

Bayes error. This component is the error made by the Bayes optimal solution. For a prediction task, this error is zero only when y is a deterministic function of x. For instance, if the binary variable y is a deterministic function of x corrupted with additive random noise, we can never hope to perfectly separate the two classes so we'll always have an error. The Bayes error is the minimal, yet inevitable, error due to the possible non-determinism of the world in our model. Since it is the best performance we can hope for (reachable only if we know $\mathcal{D}$), this error is also called the irreducible error.

Approximation error. This component reflects the quality of our prior knowledge, measured by the minimal risk achievable by a hypothesis in our hypothesis class, $L(h^*) = \min_{h \in \mathcal{H}} L(h)$. It measures how much error we have due to the restriction to a specific class, *i.e.,* how much *inductive bias* the algorithm has towards choosing a hypothesis from $\mathcal{H}$. This error is small only if the best function in $\mathcal{H}$ is close to the best function in $\mathcal{F}$, ideally with the true function linking y and $\boldsymbol{x}$ belonging to $\mathcal{H}$. We then say that $\mathcal{H}$ is adapted to $\mathcal{D}$, or has a small bias. The approximation error depends on the size, or complexity, of $\mathcal{H}$ (it can decrease with 'larger' $\mathcal{H}$), on the distribution $\mathcal{D}$, but does not depend on the learning rule A, nor the sample size n.

Estimation error. This component arises because the learning rule looks for an hypothesis that is not a minimizer of the (unknown) generalization error L. For example, the ERM learning rule minimizes the empirical risk that is only an estimate of the true risk. By the definition of Eq. (1.8), this error can be made arbitrarily small with enough samples, for any $\mathcal{D}$, if $\mathcal{H}$ is learnable by A. The quality of the estimation, hence this error, depends on the size of $\mathcal{H}$ (it decreases with 'smaller' $\mathcal{H}$) and the sample size n (it decreases with larger n).

Statistical Learning theory gives us crisp characterization of hypothesis classes that are learnable. For example if $\mathcal{H}$ has finite VC-dimension [2] then it is learnable, with the ERM learning rule. If $\mathcal{H}$ is a countable union of finite VC-dimensional spaces

[2] The VC-dimension of a class of functions is a combinatorial property that denotes the maximal sample size that can be shattered by the class.

then it is learnable with SRM. If the learning task is convex-smooth or convex-Lipschitz, then it is learnable with RLM or SGD [SSBD14]. An hypothesis class of finite dimension is a particular example of a class of finite VC-dimension. For such $\mathcal{H}$, it can be shown that, with probability $1-\delta$,

$$L(h_s) \leq L(h^*) + 2\sqrt{\frac{\log\left(2|\mathcal{H}|/\delta\right)}{2n}} \ . \tag{1.10}$$

This equation gives us a bound on the estimation error that is in theory never accessible. A consequence of (1.10) is that the larger the set of hypothesis $\mathcal{H}$, the looser the upper bound for $L(h_S)$, and, therefore, it is harder to know whether it is close to $L(h^*)$ or not. In other words, when the class of hypothesis is too large it is harder to control the estimation error of the solution. For infinite dimensional $\mathcal{H}$ (but finite VC-dimension) it is still possible to bound the estimation error in a similar way to (1.10). More details can be found in [SS+10; SSBD14]. This error also encompasses an optimization error, representing how far the actual hypothesis resulting from the optimization algorithm is from the target hypothesis. The runtime necessary to reach a given precision is called the *computational complexity* of learning.

So how should we choose a good hypothesis class $\mathcal{H}$? Since our goal is to minimize the excess error $L(h_S) - L(f^*)$, we face a trade-off regarding the last two terms called the *bias-complexity trade-off*. On one hand, a very rich class $\mathcal{H}$ is more likely to have a small approximation error (small bias) but on the other hand might have a large estimation error (a higher risk of overfitting). To learn the task $\mathcal{T}$ (find h_S with lowest generalization error *i.e.,* s.t. $L(h_S) \simeq L(f^*)$) we have to choose $\mathcal{H}$ that is small enough to be learnable by A ($L(h_S) \simeq L(h^*)$) and large enough to be well adapted to $\mathcal{D}$ ($L(h^*) \simeq L(f^*)$). Of course, an ideal choice for $\mathcal{H}$ is the class that contains only one solution: the Bayes optimal solution f^*. Unfortunately it depends on the underlying distribution $\mathcal{D}$, which we do not know (in fact, learning would have been unnecessary had we known $\mathcal{D}$). The goal of Statistical Learning theory is to study how rich we can make $\mathcal{H}$ while still being learnable, *i.e.,* maintaining reasonable estimation error.

In summary, to successfully learn a task $(\mathcal{T}, \mathcal{D})$ we need three resources:

- ▷ **Domain knowledge resources** [information on $\mathcal{D}$] to determine an appropriate set of possible solutions, *i.e.,* to choose an hypothesis class $\mathcal{H}$ adapted to $\mathcal{D}$. This allows to have a low approximation error. For example the prior knowledge that the hypothesis class of a convolutional neural network is adapted to the statistics of natural images.
- ▷ **Statistical resources** [learning theory, enough sample size n] to determine which solution to pick in this set and reach it with a given accuracy and precision, *i.e.,* to choose a learning rule that learns $\mathcal{H}$ with adapted sample

complexity, This allows to have a low estimation error. For example we'll choose the ERM learning rule if $\mathcal{H}$ has finite VC dimension and learning theory gives us the number of samples necessary the reach the best solution in $\mathcal{H}$ with a given accuracy and precision.

▷ **Computational resources** [powerful computer] to determine precisely how to pick this solution, *i.e.,* to design an algorithm that implements the learning rule with reasonable computational complexity. This allows to have a low optimisation error.

Cross-validation. Once we have designed or chosen a learning algorithm, how can we compute its true error. A practical way to get an estimate of the generalization error $L(h_S)$ is by evaluating its empirical error on a set of data points that were not considered during the minimization procedure leading to its estimation. Based on this idea, one may assess how good the solution proposed by a learning algorithm is for a certain dataset S using a cross-validation procedure [Bis07]:

▷ Partition S into K subsets containing (approximately) the same number of elements: $S = S_1 \cup \cdots \cup S_K$.

▷ For $k = 1 \dots K$, define the train and test folds: $S_k^{\text{train}} = S \setminus S_k$ and $S_k^{\text{test}} = S_k$ and compute the empirical error calculated on each test fold, $L_k = L_{S_k^{\text{test}}}\left(h_{S_k^{\text{train}}}\right)$.

▷ Define the average performance of the learning algorithm on dataset S by $L = \frac{1}{K}\sum_{k=1}^{K} L_k$, which is the average empirical error of the solutions proposed by the learning algorithm on each test fold; the expected value of L is the generalization error of h_S [SSBD14].

Even if this method often works very well in practice, the exact behaviour of cross-validation is not yet fully understood theoretically [BHT21].

1.1.3 Lessons for regression on M/EEG signals

In many cases, empirical research focuses on designing good hypothesis classes for a certain domain, *i.e.,* classes for which the approximation error would not be excessively high, classes adapted to our task at hand. The idea is that although we can't access the oracle $\mathcal{D}$ and do not know how to construct the optimal classifier f^*, we still have some prior knowledge on our specific task, which enables us to design hypothesis classes for which both the approximation error and the estimation error are not too large. Indeed, our first contribution, derived in this chapter, has been to design a class $\mathcal{H}$ perfectly adapted to M/EEG signals and source-power outcome, in the sense that it shows no approximation error. Such algorithms are said to be statistically consistent. When learning from a sufficiently large number of samples (no estimation error) and without noise in y (no irreducible error), a regression algorithm with no approximation error will have no generalization error

(see (1.9)). It has then learnt a function that perfectly approximates the true function asymptotically. The focus of this theoretical section is therefore to find a statistically consistent power-regression algorithm on M/EEG signals. The class $\mathcal{H}$ we found (composed by linear functions) is also learnable with the ERM learning rule and shows no optimization error.

Our prediction problem has a much larger number of features than number of observations. Indeed, the number of observations (*e.g.,* subjects) N is a few hundreds in most studies whereas the dimension of the input is a few millions ($100\,000\,\mathrm{time}$ samples corresponding to few minutes of M/EEG signal sampled at $1000\,\mathrm{Hz}$, for each of the few hundreds electrodes P). Such problems have become of increasing importance, especially in genomics, computational biology and neuroscience. In this setting high variance and overfitting are a major concern [Has+05]. As a result, simple, highly regularized approaches like the Ridge regression algorithm, using the SRM learning rule, often become the methods of choice.

To tackle our task of predicting a continuous outcome from M/EEG signals, the learning theory just described teaches us four important lessons. First, the generalization bounds (1.10) (together with the curse of dimensionality, see [Has+05]) calls for low-dimensional inputs, requiring to derive a compact yet complete summary of the high-dimensional M/EEG signal $\boldsymbol{x}(t)$: this will be investigated in Section 1.2 and lead to the spatial covariance matrix. Second, most commonly used ML algorithms operate on Euclidean objects living in a vector space, requiring to vectorize the above representation: this is the subject of Section 1.3. Third, the No-Free-Lunch theorem imposes a form of prior knowledge about our task to successfully learn it: in Section 1.4 we will derive a generative model $\mathcal{D}$ of both the M/EEG signal and the outcome, backed by our understanding of the physiological mechanism generating brain electrical activity and of the physics of M/EEG acquisition. Finally, to be performant, our predictive biomarker should be based on a statistically consistent regression algorithm: we presented such an algorithm, derived in Section 1.5, leveraging our data generative model to design a perfectly adapted function class $\mathcal{H}$.

1.2 Multivariate time series analysis

Most phenomena are visible to us via signals - measurable temporal variations of different sources of activity. The scientific study of these phenomena therefore relies on the analysis of these signals. For example the yearly land-ocean surface temperatures recorded at different locations on Earth may help to study global warming, the quarterly earnings per share of biggest companies in a particular industry may reflect its global financial health, brain electrical activity recorded from various locations on the scalp via M/EEG could allow testing how different brain areas are responding to a particular stimulus. To study these phenomena, scientists need mathematical tools to quantify, understand, model, and predict the time evolution of such signals. Multivariate time series are the standard mathematical tool for describing and analysing signals coming from measurements from multiple sensors during a physical or biological experiments. In this section, we define multivariate time series and present probabilistic tools to analyze them. We will discuss some common assumptions regarding their statistics and then define a compact representation in the form of the covariance matrix. Finally we'll explain why these matrices are often rank-deficient in the case of M/EEG signals.

1.2.1 Statistical and temporal moments

When an analog signal (valued at continuous time $t \in \mathbb{R}$) is measured by an electronic device, it is generally sampled with a uniform interval of length T_s, the sampling period, by recording its values at times $\{tT_s\}_{t\in\mathbb{Z}}$. This gives rise to a digital signal (valued at discrete-time $t \in \mathbb{Z}$). Discrete signal processing has replaced analog signal processing in most applications with more sophisticated and precise algorithms. Then, reconstructing the initial continuous-time signal from its samples is, in some conditions, possible using interpolation algorithms.

Most real-world digital **deterministic signals** can be modeled by a **multivariate time series**. A multivariate time series is a collection of P-dimensional vectors, indexed by $t \in \mathbb{Z}$:

$$\boldsymbol{x}(t) = \begin{bmatrix} x_1(t) \\ \vdots \\ x_P(t) \end{bmatrix} \tag{1.11}$$

Each dimension $x_k(t),\ k = 1 \dots P$ in $\boldsymbol{x}(t)$ represents a different quantity: it can describe the time evolution of a certain stock composing a portfolio of P stocks in Finance [Lut07], the voice recorded at one of the P microphone in an acoustic experiment [OS94], or the neural activity registered by one electrode placed on a subject's scalp [SC07] in a P-channel EEG recording.

To analyze properties of a class of deterministic signals, such as general resting-state EEG signals of different subjects, the standard approach is to use a probabilistic framework and represent this class by a random signal, whose probability distribution reflects the common properties of those signals. This **random signal** is modeled by a **stochastic process**. In this approach, each time sample $\boldsymbol{x}(t)$ is considered as a random vector $\boldsymbol{x}(t,\omega)$ in $\mathbb{R}^P$, generated by some statistical law whose probability density function is $p_{\boldsymbol{x}(t)}$. Each possible realization $\{\boldsymbol{x}(t,\omega_0)\}_{t\in\mathbb{Z}}$ of this stochastic process is a multivariate time series called a trajectory. All trajectories correspond to all signals of the class. This modeling allows to efficiently code signals of a same class and to separate the signal of interest from a noise whose stochastic features are different. To enhance readability, please note that we will use the overloaded notation $\boldsymbol{x}(t)$ to denote both the random vector $\boldsymbol{x}(t,\omega)$ and one of its realization $\boldsymbol{x}(t,\omega_0)$, where the context will make the definition unambiguous. We focus on signals taking values in $\mathbb{R}$.

The statistics of the signal $\boldsymbol{x}(t)$ is completely described by its **statistical moments**, among which its first order moment, a vector called the statistical mean at each time t,

$$\boldsymbol{\mu}(t) = \mathbb{E}\left[\boldsymbol{x}(t)\right] = \int_{\mathbb{R}^P} \boldsymbol{y}\; p_{\boldsymbol{x}(t)}(\boldsymbol{y})\mathrm{d}\boldsymbol{y} \;\in \mathbb{R}^P \;, \tag{1.12}$$

and its second order moment, a matrix called the statistical autocovariance between two times t and s,

$$\begin{aligned} \boldsymbol{R}(t,s) &= \mathbb{E}\Big[\big(\boldsymbol{x}(t)-\boldsymbol{\mu}(t)\big)\big(\boldsymbol{x}(s)-\boldsymbol{\mu}(s)\big)^\top\Big] \in \mathbb{R}^{P\times P} && (1.13)\\ &= \int_{\mathbb{R}^P\times\mathbb{R}^P} \big(\boldsymbol{y}-\boldsymbol{\mu}(t)\big)\big(\boldsymbol{z}-\boldsymbol{\mu}(s)\big)^\top\, p_{[\boldsymbol{x}(t),\boldsymbol{x}(s)]}(\boldsymbol{y},\boldsymbol{z})\mathrm{d}\boldsymbol{y}\mathrm{d}\boldsymbol{z}\;, && (1.14)\end{aligned}$$

where $p_{[\boldsymbol{x}(t),\boldsymbol{x}(s)]}$ is the joint probability density function for $\boldsymbol{x}(t)$ and $\boldsymbol{x}(s)$, and $\boldsymbol{x}^\top$ denotes the transpose of $\boldsymbol{x}$. Other statistical quantities may also be defined, such as higher-order moments (kurtosis, skewness, etc.) [NM93] or the entropy of the time series [BV00], but we will not consider them in this thesis.

Similarly, we can define their temporal counterparts, the **temporal moments**, whereby instead of fixing the time t and averaging across all possible realizations of the random vector, we fix a particular realization and average across all the time course. The resulting quantities are therefore random objects: the temporal mean,

$$\overline{\boldsymbol{\mu}} = \langle \boldsymbol{x}(t)\rangle = \lim_{T\to+\infty}\frac{1}{T}\sum_{t=0}^{T-1}\boldsymbol{x}(t)\;, \tag{1.15}$$

and the temporal autocovariance:

$$\overline{\boldsymbol{R}(\tau)} = \Big\langle \big(\boldsymbol{x}(t+\tau)-\overline{\boldsymbol{\mu}}\big)\big(\boldsymbol{x}(t)-\overline{\boldsymbol{\mu}}\big)^\top\Big\rangle \tag{1.16}$$

1.2.2 Statistical assumptions

Inferring statistically valid conclusions about populations from samples requires some background assumptions, statements that one makes to make the building of theoretical models easier. For instance we can use these assumptions to obtain better estimators for describing the statistical law of the samples (less bias and smaller variance), as well as clearer interpretations about the underlying stochastic process that generated them [Lut07]. However, they must be made carefully, since incorrect assumptions can yield to highly inaccurate conclusions.

Stationarity. One of the most common assumptions refers to how the statistics of $\boldsymbol{x}(t)$ evolves in time. Strict-sense stationarity assumes that any joint probability law is invariant by any temporal shift, which we can roughly think of as "the statistical moments do not depend on time". Most of the time in signal processing we observe only one realization, from which one want to estimate certain parameters of the underlying stochastic process. The poorness of this numerical information generally limits the investigation to the study to its mean and autocovariance. Since we only study the second-order moments of the process, we generally adopt the milder hypothesis of *wide-sense stationarity* (WSS), that assumes stationarity up the second order. This implies that the statistical mean of the multivariate time series do not depend on time,

$$\boldsymbol{\mu}(t) = \boldsymbol{\mu} \,, \tag{1.17}$$

and that the autocovariance matrix between two times t and s depends only on their difference $\tau = s - t$, and not on time t:

$$\boldsymbol{R}(t,s) = \boldsymbol{R}(t,t+\tau) = \boldsymbol{R}(\tau) = \mathbb{E}\left[\boldsymbol{x}(t+\tau) - \boldsymbol{\mu})(\boldsymbol{x}(t) - \boldsymbol{\mu})^\top\right] \,. \tag{1.18}$$

Under the WSS hypothesis, the **cross-power spectral density** of a multivariate time series is defined as the discrete-time Fourier transform (DTFT) of the sequence of auto-covariance matrices [Pri83; PW93]:

$$\boldsymbol{S}(f) = \sum_{k\in\mathbb{Z}} \boldsymbol{R}(k) e^{-j2\pi fk} \,. \tag{1.19}$$

The cross-power spectral density matrices are positive definite matrices whose diagonal values (also known as the **power spectral density or PSD**) describe how the power (or variance) of each time series in $\boldsymbol{x}(t)$ is distributed along the frequency domain; the out-of-diagonal values portray the statistical correlation between the time series in each pair of dimensions in the frequency domain. For example the spectral power of a univariate white noise (a WSS process whose values at different times are uncorrelated) is constant. These matrices completely

characterizes the autocovariance of the process which can be retrieved by the inverse Fourier transform.

Stationarity ensures interesting statistical properties on the time series, but is not always adequate to assume, in particular when the goal is to identify changes in the statistics of the samples, such as detecting changes in neural connectivity [Ast+08; RB15]. But even in this context, it is common to assume that the changes in the statistics are relatively smooth, so that samples close in time have approximately the same statistics. In this approach, one can consider that the samples in a small sliding window can be described by the same mean vector and autocovariance matrices. Then, the evolution of $\boldsymbol{x}(t)$'s statistics can be described by how its mean and auto-covariance matrices evolve from one window to the next. This window should not be too small (would then yield poor statistical estimators), nor too large (may blur the dynamics of study).

Ergodicity and parameter estimation. Another common (yet often implicit) assumptions refers to how the time-course of $\boldsymbol{x}(t)$ evolves in statistics. Strict-sense ergodicity assumes that trajectories of a stochastic process will eventually visit all parts of its space in a uniform sense, which we can roughly think of as "the temporal moments do not depend on ω". Similarly to stationarity, we generally adopt the milder hypothesis of *wide-sense ergodicity*, that assumes ergodicity up the second order.

In this thesis we will only consider wide-sense stationary and ergodic signals taking real values. A fundamental property of such signals is that their temporal moments (time averages) and the statistical moments (ensemble or population averages) are equal, with the very important experimental consequence that only one trajectory is enough to determine all the statistical moments. Assuming that $\boldsymbol{x}(t)$ is wide-sense stationary and ergodic over T samples, $\{\boldsymbol{x}(0), \dots, \boldsymbol{x}(T-1)\}$, we can write the estimators for (1.12), (1.13) and (1.19) as:

$$\hat{\boldsymbol{\mu}} = \frac{1}{T} \sum_{t=0}^{T-1} \boldsymbol{x}(t) \tag{1.20}$$

$$\hat{\boldsymbol{R}}(\tau) = \frac{1}{T-|\tau|} \sum_{t=0}^{T-1-|\tau|} \Big(\boldsymbol{x}(t+|\tau|) - \hat{\boldsymbol{\mu}}\Big)\Big(\boldsymbol{x}(t) - \hat{\boldsymbol{\mu}}\Big)^{\top} \tag{1.21}$$

$$\hat{\boldsymbol{S}}(f) = \mathrm{DTFT}(\hat{\boldsymbol{R}}(\tau)) \tag{1.22}$$

The cross-spectral density matrices can also be obtained via spectral estimation methods such as the periodogram or Welch's method [PW93].

Without loss of generality, we will consider that all time series are zero-mean, so that their parametrization, up to the second order, may be done using their cross-spectral density matrices only. Furthermore, we will assume that the time series have been bandpass filtered and so their spectral content is supported on a set of frequencies denoted by $\mathcal{F}$.

A special note on the Gaussianity assumption. Another usual assumption concerns the statistics of $\boldsymbol{x}(t)$ and assume that $p_{\boldsymbol{x}(t)}$ can be approximated by a multivariate Gaussian distribution. Under this hypothesis, the second-order moments, the mean vector and sequence of autocovariance matrices $\boldsymbol{R}(\tau)$ (or, equivalently, the cross-spectral density matrices $\boldsymbol{S}(f)$) exhaustively describe the full statistical behavior of $\boldsymbol{x}(t)$ [Pri83]. This hypothesis is appealing in numerous ways *e.g.,* a WSS Gaussian process has also the much stronger property of being strict-sense stationary. It can also be used for defining a notion of distance between two time series. Since the second-order moments encapsulates information about the whole probability distribution, we can use them to compare two time series, $\boldsymbol{x}_i(t)$ and $\boldsymbol{x}_j(t)$ by defining their cross-spectrum distance as

$$d_S(\boldsymbol{x}_i(t), \boldsymbol{x}_j(t))^2 = \int_{\mathcal{F}} d^2(\boldsymbol{S}_i(f), \boldsymbol{S}_j(f)) \mathrm{d}f \ , \tag{1.23}$$

where $\boldsymbol{S}_i(f)$ and $\boldsymbol{S}_j(f)$ are the cross-spectral density matrices of $\boldsymbol{x}_i(t)$ and $\boldsymbol{x}_j(t)$, respectively, and d is some distance between matrices. The choice of the distance d will be extensively discussed in Section 1.3. This approach of comparing two time series based on their cross-spectral densities is more appropriate than directly comparing their samples on a given realization because it is based on information defining the whole probability distribution. It also lead to stronger performance on classification [CRJ19], regression [Sab+19a] and clustering tasks [Cha+13].

The Gaussianity assumption is often justified in the literature by different arguments. If several independent factors play a role in the generation of the data, one may use the central limit theorem to argue that the sum of all their contributions yields a statistical behavior that can be well described by a Gaussian distribution [PP02]. As a result, Gaussian processes appear in numerous physical phenomena. It also gives a numerical advantage: parameter estimation under the Gaussian model yields convex optimization problems that have analytic solutions. Finally, without any knowledge about the statistics of the data, in particular with no information on the moments of order strictly higher than two, the Gaussian distribution is the most conservative assumption: among all probability distributions defined on an infinite domain and with a given mean and variance, it is the one with maximal differential entropy *i.e.,* the one that requires the maximum quantity of information to encode it.

However, the Gaussian assumption is very strong and is not always justified. For instance, the statistics of rare events are better described by Poisson distributions. In

general, the use of Gaussian distributions to model data is appropriate when one has no knowledge about the physical phenomena that generates its samples. As we will develop a generative model of M/EEG signals and related biomedical outcomes, we will not resort to this strong hypothesis. This will give a wider generality to our contribution, and will yield a more robust biomarker. Note that we can still consider that, up to the second order, we have:

$$d_S(\boldsymbol{x}_i(t), \boldsymbol{x}_j(t))^2 \simeq \int_{\mathcal{F}} d^2(\boldsymbol{S}_i(f), \boldsymbol{S}_j(f)) \mathrm{d}f \ . \tag{1.24}$$

1.2.3 The covariance matrix

A good estimate of the cross-spectral density matrix requires enough time samples. If not the case, it is wiser to condense the information contained in the spectrum into a single parameter. Comparing two time series then amounts to compare their corresponding parameters. One can do this by noticing that the inverse DTFT applied to the cross-spectral density matrices of a zero-mean $\mathcal{F}$-bandpass filtered time series $\boldsymbol{x}(t)$ gives

$$\int_{\mathcal{F}} \boldsymbol{S}(f) \mathrm{d}f = \boldsymbol{R}(0) = \mathbb{E}[\boldsymbol{x}(t)\boldsymbol{x}(t)^\top] = \boldsymbol{C} \ , \tag{1.25}$$

which is the covariance matrix of $\boldsymbol{x}(t)$ and can be calculated without having to estimate its spectrum. Denoting the data matrix $\boldsymbol{X} \in \mathbb{R}^{P \times T}$ with T the number of time samples, and using (1.21) with $\tau = 0$, the covariance matrix reads:

$$\boldsymbol{C} \simeq \frac{\boldsymbol{X}\boldsymbol{X}^\top}{T} \in \mathbb{R}^{P \times P} \quad \text{with} \quad \boldsymbol{X} \in \mathbb{R}^{P \times T} \ . \tag{1.26}$$

For the sake of simplicity, we assume that T is the same for each time series, although it is not required by the following method. The diagonal of this matrix represents the variance (power) of each sensor, while the off-diagonal terms contain the covariance between each pair of signals. Negative values in the off-diagonal express negative correlations.

We may then approximately compare two time series $\boldsymbol{x}_i(t)$ and $\boldsymbol{x}_j(t)$ using the distance between their respective covariance matrices $\boldsymbol{C}_i$ and $\boldsymbol{C}_j$ as

$$d_S(\boldsymbol{x}_i(t), \boldsymbol{x}_j(t))^2 \simeq d^2(\boldsymbol{C}_i, \boldsymbol{C}_j) \ . \tag{1.27}$$

Under our assumptions, comparing two time series boils down to estimating their covariance matrices. A common problem arising from covariance estimation in M/EEG signals is rank-deficiency, mainly occurring for two different reasons. First, high-dimensional statistics show that one need enough available samples to expect a good estimate of the covariance matrix that describes the statistics of the data. This may not be the case for some applications. The covariance $\boldsymbol{C}$ is a $P \times P$ matrix

which is a sum of T rank-one matrices (as per (1.21) with $\tau = 0$). Therefore, there should be at least P samples available for the estimate $\hat{\boldsymbol{C}}$ to have a chance of not being rank-deficient. In fact, the number of parameters to estimate in a covariance grows quadratically with dimension so many more samples are required than there are sensors to accurately estimate such matrices [EG15; Rod+17]. A common approach for alleviating such problem, called 'shrinkage', use a regularization term that adds a weighted Identity matrix to $\hat{\boldsymbol{C}}$, with the optimal weight being determined from the data. Many methods have been proposed for determining the weight of the regularization term [Che+10; EG15]. Second, even in the case where we have enough available samples to estimate the covariance matrix, the signal preprocessing used to enhance the SNR of noisy signals lead to inherently rank-deficient covariance matrices. This is the case of M/EEG signals.

1.2.4 M/EEG preprocessing induces rank-deficiency

Even though the magnetic shielded room and gradiometer coils can help to reduce the effects of external interferences on MEG signal (as seen in the Introduction chapter), the problem mainly remains and further reduction is needed. Also, additional artifact signals can be caused by movement of the subject during recording if the subject has small magnetic particles on his body or head. The Signal Space Separation (SSS) method can help mitigate those problems [TK05]. Besides, physiological artifacts (eye blinks and heart beats) can cause prominent artifacts in the recording. The Signal Space Projections (SSP) method is typically used to reduce them [UI97]. Both of these methods are described below and enable to increase the SNR, at a price of rendering covariance matrices rank-deficient.

Signal Space Separation (SSS)

The Signal Space Separation (SSS) method [TK05], also called Maxwell Filtering, is a biophysical spatial filtering method, purely MEG specific, that aim to produce MEG signals cleaned from external interferences and from movement distortions and artifacts.

A MEG device records the neuro-magnetic field distribution by sampling the field simultaneously at P distinct locations around the subject's head. At each moment of time, the measurement is a vector $\boldsymbol{x} \in \mathbb{R}^P$ where P is the total number of recording channels. In theory, any direction of this vector in the signal space represents a valid measurement of a magnetic field, however the knowledge of the location of possible sources of magnetic field, the geometry of the sensor array and electromagnetic theory (Maxwell's equations and the quasi static approximation) considerably constrain the relevant signal space and allow us to differentiate between signal space directions consistent with a brain's field and those that are not.

To be more precise, it has been shown that the recorded magnetic field is a gradient of a harmonic scalar potential. A harmonic potential $V(\boldsymbol{r})$ is a solution of the Laplacian differential equation $\nabla^2 V = 0$, where $\boldsymbol{r}$ is represented by its spherical coordinates (r, θ, ψ). It is known that a harmonic function in a three-dimensional space can be represented as a series expansion of spherical harmonic functions $Y_{lm}(\theta, \phi)$:

$$V(\boldsymbol{r}) = \sum_{l=1}^{\infty} \sum_{m=-l}^{l} \alpha_{lm} \frac{Y_{lm}(\theta, \phi)}{r^{l+1}} + \sum_{l=1}^{\infty} \sum_{m=-l}^{l} \beta_{lm} r^l Y_{lm}(\theta, \phi) \tag{1.28}$$

We can separate this expansion into two sets of functions: those proportional to inverse powers of r and those proportional to powers of r. From a given array of sensors and a coordinate system with its origin somewhere inside of the helmet, we can compute the signal vectors corresponding to each of the terms in 1.28.

Following notations of [TK05], let $\boldsymbol{a}_{lm}$ be the signal vector corresponding to term $\frac{Y_{lm}(\theta,\phi)}{r^{l+1}}$ and $\boldsymbol{b}_{lm}$ the signal vector corresponding to $r^l Y_{lm}(\theta, \phi)$, our measurement is given by:

$$\boldsymbol{x} = \sum_{l=1}^{\infty} \sum_{m=-l}^{l} \alpha_{lm} \boldsymbol{a}_{lm} + \sum_{l=1}^{\infty} \sum_{m=-l}^{l} \beta_{lm} \boldsymbol{b}_{lm} \tag{1.29}$$

A set of P such signal vectors forms a basis in the P dimensional signal space. This basis is not orthogonal, but still linearly independent so any measured signal vector has a unique representation in this basis:

$$\boldsymbol{x} = [\boldsymbol{S_{in}} \quad \boldsymbol{S_{out}}] \begin{bmatrix} \boldsymbol{x}_{in} \\ \boldsymbol{x}_{out} \end{bmatrix} , \tag{1.30}$$

where the sub-bases $\boldsymbol{S}_{in}$ and $\boldsymbol{S}_{out}$ contain the basis vectors $\boldsymbol{a}_{lm}$ and $\boldsymbol{b}_{lm}$ respectively, and vectors $\boldsymbol{x}_{in}$ and $\boldsymbol{x}_{out}$ contain the corresponding α_{lm} and β_{lm} values.

It can be shown that the basis vectors corresponding to the terms in the second sum in expansion (1.28) represent the perturbation sources external to the helmet. We can then separate the components of field arising from sources inside and outside of the helmet. We can therefore decompose the signal vector $\boldsymbol{x}$ into the sum of 2 components: $\boldsymbol{\phi}_{in} = \boldsymbol{S}_{in}\boldsymbol{x}_{in}$ in the brain signal space with basis $\boldsymbol{S}_{in}$ and $\boldsymbol{\phi}_{out} = \boldsymbol{S}_{out}\boldsymbol{x}_{out}$ in the interference space with basis $\boldsymbol{S}_{out}$ (not necessarily orthogonal to $\boldsymbol{S}_{in}$). By discarding $\boldsymbol{\phi}_{out}$, we are left with the part of the signal coming from inside of the helmet only. Hence, $\boldsymbol{\phi}_{in}$ reproduces in all the MEG channels the signals that would be seen if no interference from sources external to the helmet existed: we have performed signal space separation. To further reduce the noise, we can discard the high l, m end of the spectrum: indeed the spherical harmonic functions are

known to contain increasingly higher spatial frequencies when going to higher index values (l,m) so that the signals from real magnetic sources are mostly contained in the low l, m end of the spectrum.

After projection in the lower-dimensional SSS basis we project back the signal in its original space producing a signal $\boldsymbol{X}^{clean} = \boldsymbol{S}_{in}\boldsymbol{S}_{in}^{\top}\boldsymbol{X} \in \mathbb{R}^{P\times T}$ ($\boldsymbol{S}_{in} \in \mathbb{R}^{P\times R}$) with a much better SNR (reduced noise variance) but with a rank $R \leq P$. As a result each reconstructed sensor is then a linear combination of R synthetic source signals, which modifies the inter-channel correlation structure, rendering the covariance matrix rank-deficient.

Signal Space Projection (SSP)

If one knows, or can estimate, K linearly independent source patterns $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_K$ that span the space $S = \text{span}(\boldsymbol{a}_1, \ldots, \boldsymbol{a}_K) \subset \mathbb{R}^P$, one can estimate an orthonormal basis $U_K \in \mathbb{R}^{P\times K}$ of S by taking the first K left singular vectors in the singular value decomposition (SVD) of the matrix formed by the source patterns in columns. We can then separate the sensor-space signal $\boldsymbol{x}$ into a signal $\boldsymbol{s}_{\|}$ produced by those K sources (belonging to the subspace S) and a signal $\boldsymbol{s}_{\perp}$ that can't be produced by any linear combination of those sources (belonging to the corresponding orthogonal subspace):

$$\boldsymbol{x} = \boldsymbol{x}_{\|} + \boldsymbol{x}_{\perp} = U_K U_K^{\top}\boldsymbol{x} + (I - U_K U_K^{\top})\boldsymbol{x} \tag{1.31}$$

If S is formed from specific artifacts then $\boldsymbol{x}_{\perp}$ is mostly free of those.

This is the idea behind the Signal Space Projections (SSP) method [UI97]. In practice SSP is used to reduce physiological artifacts (eye blinks and heart beats) that cause prominent artifacts in the recording. SSP projections are computed from time segments contaminated by the artifacts and the first component (per artifact and sensor type) are projected out. To take a more concrete example, in the Cam-CAN MEG dataset [Sha+14], eye blinks are monitored by 2 electro-oculogram (EOG channels), and heart beats by an electro-cardiogram (ECG channel). The EOG and ECG channels are used to identify the artifact events (after a first band-pass filter to remove DC offset and an additional [1-10]Hz filter applied only to EOG channels to remove saccades vs blinks). As an illustration of this process, in our MEG experiment described in Section 2.1.2, we filtered the raw signal in [1-35]Hz band and created data segments (called epochs) around those events, rejecting those whose peak-to-peak amplitude exceeds a certain global threshold. For each artifact and sensor type these epochs are then averaged and the first component of maximum variance is extracted via PCA. Signal is then projected in the orthogonal space, again leading to rank-deficient covariance matrices. This follows the guidelines of the MNE software [Gra+14]. It is interesting to compare SSP with SSS geometrically. Like SSP, SSS is a form of projection. Whereas SSP

empirically determines a noise subspace based on data (empty-room recordings, EOG or ECG activity, etc.) and projects the measurements onto a subspace orthogonal to the noise, SSS mathematically constructs the external and internal subspaces from spherical harmonics and reconstructs the sensor signals using only the internal subspace (*i.e.,* does an oblique projection).

To summarize this section, we saw that our assumptions of stationarity and ergodicity allowed to represent a M/EEG time series $\boldsymbol{x}(t)$ by its covariance matrix $\boldsymbol{C}$ in a certain frequency band, either full rank or rank-deficient. Covariance matrices are positive matrices, so they have a particular structure. This structure implies that they don't live in a flat space (a vector space), they live in a curved space, called a manifold. Hence, a covariance matrix can't be readily vectorized. Even though some algorithms can manipulate covariance matrices directly as input [Bar+12], most commonly used ML regression algorithms (and also the most performant ones for M/EEG classification tasks) assume an Euclidean structure of their input requiring to vectorize the covariance matrix. To do that properly, we must first introduce some important concepts regarding Riemannian manifold.

1.3 Riemannian matrix manifolds

1.3.1 Riemannian manifolds

In this work, we consider **differentiable manifolds** $\mathcal{M}$ in $\mathbb{R}^P$ of dimension K. Intuitively differentiable manifolds are "curved" spaces that locally at each point resemble a flat vector space (see [AMS09], chap. 3 and [PFA06]). Examples of differentiable manifolds in $\mathbb{R}^3$ are curves (one-dimensional manifolds which locally look like a straight line) and surfaces (two-dimensional manifolds which locally look like a plane). More precisely, each point of the manifold $\boldsymbol{M} \in \mathcal{M}$ is associated to a vector space called **tangent space** at $\boldsymbol{M}$, denoted $\mathcal{T}_{\boldsymbol{M}}$. For any matrix $\boldsymbol{M}' \in \mathcal{M}$, as $\boldsymbol{M}' \to \boldsymbol{M}$, $\boldsymbol{\xi}_{\boldsymbol{M}} = \boldsymbol{M}' - \boldsymbol{M} \in \mathcal{T}_{\boldsymbol{M}}$. It is the set of derivatives of curves on the manifold passing through $\boldsymbol{M}$. The dimension of $\mathcal{T}_{\boldsymbol{M}}$ is K, the dimension of $\mathcal{M}$. The differentiable manifold becomes **Riemannian** when each tangent space $\mathcal{T}_{\boldsymbol{M}}$ is endowed with a **metric**, i.e. an inner product $\langle \cdot, \cdot \rangle_{\boldsymbol{M}} : \mathcal{T}_{\boldsymbol{M}} \times \mathcal{T}_{\boldsymbol{M}} \to \mathbb{R}$, that defines a local Euclidean structure. This metric is supposed smooth across points on the manifold. An Euclidean space is a particular Riemannian manifold with a constant metric. We can then define:

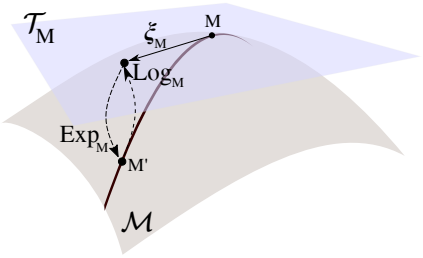


Fig. 1.1: Tangent Space, exponential and logarithm on Riemannian manifold illustration

A norm on the tangent space $\mathcal{T}_{\boldsymbol{M}}$: $\|\boldsymbol{\xi}_{\boldsymbol{M}}\|^2_{\boldsymbol{M}} = \langle \boldsymbol{\xi}_{\boldsymbol{M}}, \boldsymbol{\xi}_{\boldsymbol{M}} \rangle_{\boldsymbol{M}}$ for $\boldsymbol{\xi}_{\boldsymbol{M}} \in \mathcal{T}_{\boldsymbol{M}}$.

The length of a path between two points $\boldsymbol{M}, \boldsymbol{M}' \in \mathcal{M}$: for a path $\gamma : [0,1] \to \mathcal{M}$ such that $\gamma(0) = \boldsymbol{M}$ and $\gamma(1) = \boldsymbol{M}'$, the length of γ is $L(\gamma) = \int_0^1 \|\dot{\gamma}(t)\|_{\gamma(t)} dt$. This generalizes the usual notion of path length in Euclidean spaces.

A distance on the manifold $\mathcal{M}$, defined as the minimum length of paths: $d(\boldsymbol{M}, \boldsymbol{M}') = \min L(\gamma)$ such that $\gamma(0) = \boldsymbol{M}$, and $\gamma(1) = \boldsymbol{M}'$. This distance is called the *geodesic* distance. If $\mathcal{M}$ is an Euclidean space, this distance is simply the usual Euclidean distance: $d(\boldsymbol{M}, \boldsymbol{M}') = \|\boldsymbol{M} - \boldsymbol{M}'\|_2$, achieved when γ is a straight line between $\boldsymbol{M}$ and $\boldsymbol{M}'$.

The Frechet mean $\overline{\boldsymbol{M}}$ of a set of points $\boldsymbol{M}_i \in \mathcal{M}$ is defined as

$$\overline{\boldsymbol{M}} = \text{Mean}_d(\boldsymbol{M}_1, \ldots, \boldsymbol{M}_N) = \underset{\boldsymbol{M} \in \mathcal{M}}{\arg\min} \sum_{i=1}^{N} d(\boldsymbol{M}, \boldsymbol{M}_i)^2 \quad . \tag{1.32}$$

This is a generalization of averaging on manifolds. Indeed, in an Euclidean space, the average $\frac{1}{N}\sum_{i=1}^N \boldsymbol{M}_i$ is the Frechet mean of $\boldsymbol{M}_1, \ldots \boldsymbol{M}_N$ with respect to the Euclidean distance $d(\boldsymbol{M}, \boldsymbol{M}') = \|\boldsymbol{M} - \boldsymbol{M}'\|_2$. Another example in $\mathbb{R}_+^*$, the geometric mean

between positive numbers $a_1, \ldots, a_N > 0$, given by $\overline{a} = (a_1 \times \cdots \times a_N)^{1/N}$, is the Frechet mean of $(a_1, \ldots, a_N)$ with respect to the distance $d(a, a') = |\log(\frac{a}{a'})|$.

The exponential mapping $\text{Exp}_{\boldsymbol{M}} : \mathcal{T}_{\boldsymbol{M}} \to \mathcal{M}$ is the operation that maps the tangent space, which has a simple Euclidean structure, to the manifold which might have a much more complicated structure. It satisfies $d(\text{Exp}_{\boldsymbol{M}}(\boldsymbol{\xi}_{\boldsymbol{M}}), \boldsymbol{M}) = \|\boldsymbol{\xi}_{\boldsymbol{M}}\|_{\boldsymbol{M}}$ for $\boldsymbol{\xi}_{\boldsymbol{M}} \in \mathcal{T}_{\boldsymbol{M}}$ small enough.

The logarithm mapping $\text{Log}_{\boldsymbol{M}} : \mathcal{M} \to \mathcal{T}_{\boldsymbol{M}}$ is defined as the reciprocal of the exponential mapping which hence verifies $\|\text{Log}_{\boldsymbol{M}}(\boldsymbol{M}')\|_{\boldsymbol{M}} = d(\boldsymbol{M}, \boldsymbol{M}')$ for $\boldsymbol{M}' \in \mathcal{M}$ close enough from $\boldsymbol{M}$. It maps the manifold to the tangent space, while preserving the local properties of the manifold.

The vectorization operator. The logarithm mapping is of crucial importance in practical applications, since it allows to manipulate and store vectors (belonging to the tangent space) instead of points on the manifold. To be more concrete, since each tangent space is a K-dimensional Euclidean space, there exists a linear and invertible mapping $\phi_{\boldsymbol{M}} : \mathcal{T}_{\boldsymbol{M}} \to \mathbb{R}^K$ such that $\|\boldsymbol{\xi}_{\boldsymbol{M}}\|_{\boldsymbol{M}} = \|\phi_{\boldsymbol{M}}(\boldsymbol{\xi}_{\boldsymbol{M}})\|_2$ for all $\boldsymbol{\xi}_{\boldsymbol{M}} \in \mathcal{T}_{\boldsymbol{M}}$. Combining $\phi_{\boldsymbol{M}}$ and $\text{Log}_{\boldsymbol{M}}$ gives the vectorization operator at $\in \mathcal{M}$, $\mathcal{P}_{\boldsymbol{M}} = \phi_{\boldsymbol{M}} \circ \text{Log}_{\boldsymbol{M}}$ which maps $\mathcal{M}$ to $\mathbb{R}^K$, and verifies:

$$\|\mathcal{P}_{\boldsymbol{M}}(\boldsymbol{M}')\|_2 = d(\boldsymbol{M}, \boldsymbol{M}') \quad \text{for } \boldsymbol{M}' \in \mathcal{M} \ . \tag{1.33}$$

This operator explicitly captures the local Euclidean properties of the Riemannian manifold. Fig. 1.1 illustrates these concepts. Finally, if a set of matrices $\boldsymbol{M}_1, \ldots, \boldsymbol{M}_N$ is located in a small portion of the manifold, denoting $\overline{\boldsymbol{M}} = \text{Mean}_d(\boldsymbol{M}_1, \ldots, \boldsymbol{M}_N)$, it holds:

$$d(\boldsymbol{M}_i, \boldsymbol{M}_j) \simeq \|\mathcal{P}_{\overline{\boldsymbol{M}}}(\boldsymbol{M}_i) - \mathcal{P}_{\overline{\boldsymbol{M}}}(\boldsymbol{M}_j)\|_2 \tag{1.34}$$

The vectorization operator is key for machine learning applications: it projects points in $\mathcal{M}$ on $\mathbb{R}^K$, and the distance d on $\mathcal{M}$ is approximated by the distance ℓ_2 on $\mathbb{R}^K$. Therefore, those vectors can be used as input for any standard regression technique, which often assumes a Euclidean structure of the data.

As a final note, all the notions developed above are based on the metric $\langle \cdot, \cdot \rangle_{\boldsymbol{M}}$. Different metrics lead to different geodesic distances, Frechet means, exponential and logarithm mapping and vectorization operator. Choosing the right metric for a particular problem may lead to substantial benefits. For additional details on matrix manifolds, see [AMS09], chap. 3.

1.3.2 The positive definite manifold $\mathcal{S}_P^{++}$

In this thesis, we are interested in one Riemannian manifold in particular: the manifold of positive definite matrices $\mathcal{S}_P^{++}$ [FM03], to which most full rank covariance

matrices $\boldsymbol{C}$ belong. This is not a vector space, as for example the difference of two positive definite matrices may not be positive definite. It is a differentiable manifold of dimension $\frac{P(P+1)}{2}$, with fixed tangent spaces $\mathcal{T}_{\boldsymbol{C}} = \mathcal{S}_P^+$ for all $\boldsymbol{C} \in \mathcal{S}_P^{++}$.

We endow the manifold with the **geometric metric** given by: $\langle \boldsymbol{P}, \boldsymbol{Q} \rangle_{\boldsymbol{C}} = \mathrm{Tr}(\boldsymbol{P}\boldsymbol{C}^{-1}\boldsymbol{Q}\boldsymbol{C}^{-1})$. The associated **geometric norm** generalizes the Froebenius norm: $\|\boldsymbol{P}\|_{\boldsymbol{I}}$ (identity) $= \|\boldsymbol{P}\|_F$ (Frobenius) for $\boldsymbol{P} \in \mathcal{T}_{\boldsymbol{I}}$. This metric has two main advantages. First, it is affine-invariant as for any invertible matrix $\boldsymbol{W}$ it verifies $\langle \boldsymbol{W}\boldsymbol{P}\boldsymbol{W}^\top, \boldsymbol{W}\boldsymbol{Q}\boldsymbol{W}^\top \rangle_{\boldsymbol{W}\boldsymbol{C}\boldsymbol{W}^\top} = \langle \boldsymbol{P}, \boldsymbol{Q} \rangle_{\boldsymbol{C}}$. Second, it leads to closed-form formulas for most Riemannian notions seen above:

The geometric geodesic distance on the manifold $\mathcal{S}_P^{++}$ is:

$$d_G^2(\boldsymbol{C}, \boldsymbol{C}') = \|\log(\boldsymbol{C}^{-1/2}\boldsymbol{C}'\boldsymbol{C}^{-1/2})\|_F^2 = \sum_{k=1}^{P} \log^2(\lambda_k) \ , \tag{1.35}$$

where $\lambda_k, k = 1 \dots P$ are the real eigenvalues of $\boldsymbol{C}^{-1/2}\boldsymbol{C}'\boldsymbol{C}^{-1/2}$, or equivalently of $\boldsymbol{C}^{-1}\boldsymbol{C}'$. Please note that these 2 matrices have the same eigenvalues but not the same singular values (hence not the same Frobenius norm) since $\boldsymbol{C}^{-1}\boldsymbol{C}'$ is not symmetrical. Compared to the more naive Frobenius distance $d_F^2(\boldsymbol{C}, \boldsymbol{C}') = \|\boldsymbol{C} - \boldsymbol{C}'\|_F^2 = \sum_{k=1}^{P} \lambda_k^2$ where λ_k are the eigenvalues of $\boldsymbol{C} - \boldsymbol{C}'$, the geometric distance is a geodesic distance, which takes into account the intrinsic geometry of the positive definite manifold [Bha09]. We also see that the singular matrices act as a barrier for this distance: if $\boldsymbol{C}$ or $\boldsymbol{C}'$ is close from being singular, one eigenvalue λ_k goes either to 0 or $+\infty$, and $d_G(\boldsymbol{C}, \boldsymbol{C}')$ goes to infinity. Following the corresponding property of the geometric metric, this distance is **affine invariant**, *i.e.,*

$$\text{For } \boldsymbol{W} \text{ invertible, } d_G(\boldsymbol{W}\boldsymbol{C}\boldsymbol{W}^\top, \boldsymbol{W}\boldsymbol{C}'\boldsymbol{W}^\top) = d_G(\boldsymbol{C}, \boldsymbol{C}') \ . \tag{1.36}$$

This is an important property for our purpose: assume that $\boldsymbol{C}$ and $\boldsymbol{C}'$ are covariances of some signals $\boldsymbol{x}$ and $\boldsymbol{x}' \in \mathbb{R}^P$, the distance d_S (1.24) between $\boldsymbol{x}$ and $\boldsymbol{x}'$ is also invariant to linear transforms of the signals $\boldsymbol{W}\boldsymbol{x}$ and $\boldsymbol{W}\boldsymbol{x}'$: the distance is blind to global mixing effects. Indeed, when using the geometric distance in (1.24), we have for any invertible matrix $\boldsymbol{W} \in \mathbb{R}^{P \times P}$,

$$d_S(\boldsymbol{W}\boldsymbol{x}_i(t), \boldsymbol{W}\boldsymbol{x}_j(t))^2 \simeq \int_{\mathcal{F}} d^2(\boldsymbol{W}\boldsymbol{S}_i(f)\boldsymbol{W}^\top, \boldsymbol{W}\boldsymbol{S}_j(f)\boldsymbol{W}^\top)\mathrm{d}f \ , \tag{1.37}$$

$$= \int_{\mathcal{F}} \sum_{k=1}^{d} \log^2(\mu_k(f))\mathrm{d}f \ , \tag{1.38}$$

where $\mu_k(f)$ are the eigenvalues of matrix $(\boldsymbol{W}\boldsymbol{S}_i(f)\boldsymbol{W}^\top)^{-1}(\boldsymbol{W}\boldsymbol{S}_j(f)\boldsymbol{W}^\top)$
$= \boldsymbol{W}^{-T}\boldsymbol{S}_i^{-1}(f)\boldsymbol{S}_j(f)\boldsymbol{W}^T$ which, by similarity, has the same eigenvalues of $\boldsymbol{S}_i^{-1}(f)\boldsymbol{S}_j(f)$. Therefore,

$$d_S(\boldsymbol{W}\boldsymbol{x}_i(t), \boldsymbol{W}\boldsymbol{x}_j(t))^2 \simeq \int_{\mathcal{F}} d^2(\boldsymbol{S}_i(f), \boldsymbol{S}_j(f))\mathrm{d}f \simeq d_S(\boldsymbol{x}_i(t), \boldsymbol{x}_j(t))^2 \,. \tag{1.39}$$

This shows that distance (1.24) is invariant to affine transformations of time series themselves, a property that is very useful in practice. Indeed, we often observe mixing effects when working with time series related to physical phenomena. These effects can be often be approximated by the action of a linear operator, as we will see in Section 1.4.2 for M/EEG signals. In such cases, the distance (1.24) is invariant to these effects. For instance, very conveniently, the geometric distance between two time series recorded with different measurement scales (*e.g.,* V or μV) is the same. The effect of slightly moving the positions of electrodes on a subject's scalp or the effects caused by the mixture of different sources of activity in a person's brain are also examples of such mixing effects approximated by a linear operator, hence that leave the distance between time series unchanged. This distance has achieved state-of-art performance in EEG-based Brain-Computer Interfaces (BCI) classification tasks [Bar+12] using only covariance matrices as features of EEG signals.

The Frechet mean, for $P = 1$, is the geometric mean between positive scalars. In higher dimension, no closed-form formula for the Frechet mean has been discovered, but iterative algorithms to compute it are available [JVV12; CPB16]. The mean is also affine invariant, in the sense that $\boldsymbol{W}\,\overline{\boldsymbol{C}}\,\boldsymbol{W}^\top = \overline{\boldsymbol{W}\boldsymbol{C}_i\boldsymbol{W}^\top}$ (applying linear mixing on the mean of the covariances is the same as computing the mean over all individually mixed covariances). As a consequence, if the matrices $\boldsymbol{C}_i$ were jointly diagonalizable, i.e. $\boldsymbol{C}_i = \boldsymbol{A}\Lambda_i\boldsymbol{A}^\top$ with $\boldsymbol{A}$ invertible and Λ_i diagonal, we would have $\overline{\boldsymbol{C}} = \boldsymbol{A}\left(\prod_{i=1}^N \Lambda_i\right)^{1/N}\boldsymbol{A}^\top$. This property is used in the proof of consistency of the Riemann regression algorithm in Section 1.5.1.

The logarithm mapping at $\boldsymbol{C}$ is given by $\mathrm{Log}_{\boldsymbol{C}}(\boldsymbol{C}') = \boldsymbol{C}^{1/2}\log(\boldsymbol{C}^{-1/2}\boldsymbol{C}'\boldsymbol{C}^{-1/2})\boldsymbol{C}^{1/2} \in \mathcal{T}_{\boldsymbol{C}}$, and **the vectorization operator** *w.r.t.* $\boldsymbol{C}$ is

$$\mathcal{P}_{\boldsymbol{C}}(\boldsymbol{C}') = \mathrm{Upper}(\boldsymbol{C}^{-\frac{1}{2}}\mathrm{Log}_{\boldsymbol{C}}(\boldsymbol{C}')\boldsymbol{C}^{-\frac{1}{2}}) = \mathrm{Upper}(\log(\boldsymbol{C}^{-1/2}\boldsymbol{C}'\boldsymbol{C}^{-1/2})) \,, \tag{1.40}$$

where $\mathrm{Upper}(\boldsymbol{M})$ is the vector of size $P(P+1)2$ containing the upper triangular coefficients of $\boldsymbol{M}$, with unit weights on the diagonal and $\sqrt{2}$ weights on the off-diagonal. This weighting ensures that the vector and the matrix have same norms ($\|\mathrm{Upper}(\boldsymbol{M})\|_2 = \|\boldsymbol{M}\|_F$). Once again, if $\boldsymbol{C}$ and $\boldsymbol{C}'$ are covariances of $\boldsymbol{x}$ and $\boldsymbol{x}'$, it amounts to whitening $\boldsymbol{x}'$ with $\boldsymbol{C}$, and then applying a "spectral" non-linear transform on the resulting covariance, where the transform only changes the eigenvalues and not the eigenvectors.

1.3.3 The fixed rank SDP manifold $\mathcal{S}_{P,R}^{+}$

When a covariance matrix is -rank-deficient, it does not belong to $\mathcal{S}_P^{++}$ but to $\mathcal{S}_{P,R}^{+}$, the subspace of SPD matrices of fixed rank R. Unlike $\mathcal{S}_P^{++}$, it is hard to endow the $\mathcal{S}_{P,R}^{+}$ manifold with a distance that yields tractable or cheap-to-compute logarithms [VAV09]. This manifold is classically viewed as $\mathcal{S}_{P,R}^{+} = \{\mathbf{Y}\mathbf{Y}^\top | \mathbf{Y} \in \mathbb{R}_*^{P\times R}\}$, where $\mathbb{R}_*^{P\times R}$ is the set $P \times R$ matrices of rank R [Jou+10]. This view allows to write $\mathcal{S}_{P,R}^{+}$ as a quotient manifold $\mathbb{R}_*^{P\times R}/\mathcal{O}_R$, where $\mathcal{O}_R$ is the orthogonal group of size R. This means that each matrix $\mathbf{Y}\mathbf{Y}^\top \in \mathcal{S}_{P,R}^{+}$ is identified with the set $\{\mathbf{Y}\mathbf{Q} | \mathbf{Q} \in \mathcal{O}_R\}$.

It has recently been proposed [MA18] to use the standard Frobenius metric on the total space $\mathbb{R}_*^{P\times R}$. This metric in the total space is equivalent to the **Wasserstein distance** [BJL18] on $\mathcal{S}_{P,R}^{+}$:

$$d_W(\boldsymbol{S}, \boldsymbol{S}') = \left[\mathrm{Tr}(\boldsymbol{S}) + \mathrm{Tr}(\boldsymbol{S}') - 2\mathrm{Tr}((\boldsymbol{S}^{\frac{1}{2}}\boldsymbol{S}'\boldsymbol{S}^{\frac{1}{2}})^{\frac{1}{2}})\right]^{\frac{1}{2}} \tag{1.41}$$

This provides cheap-to-compute **logarithm mapping**:

$$\mathrm{Log}_{\boldsymbol{Y}\boldsymbol{Y}^\top}(\boldsymbol{Y}'\boldsymbol{Y}'^\top) = \boldsymbol{Y}'\boldsymbol{Q}^* - \boldsymbol{Y} \in \mathbb{R}_*^{P\times R} \ , \tag{1.42}$$

where $\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top = \boldsymbol{Y}^\top\boldsymbol{Y}'$ is a singular value decomposition and $\boldsymbol{Q}^* = \boldsymbol{V}\boldsymbol{U}^\top$. The **vectorization operator** is then given by $\mathcal{P}_{\boldsymbol{Y}\boldsymbol{Y}^\top}(\boldsymbol{Y}'\boldsymbol{Y}'^\top) = \mathrm{vect}(\boldsymbol{Y}'\boldsymbol{Q}^* - \boldsymbol{Y}) \in \mathbb{R}^{PR}$, where the vect of a matrix is the vector containing all its coefficients.

This framework offers closed-form projections in the tangent space for the Wasserstein distance, which can be used to perform regression. Importantly, since $S_P^{++} = S_{P,P}^{+}$, we can also use this distance on the positive definite matrices. This distance possesses the *orthogonal invariance* property:

$$\text{For } \boldsymbol{W} \text{ orthogonal, } d_W(\boldsymbol{W}^\top\boldsymbol{S}\boldsymbol{W}, \boldsymbol{W}^\top\boldsymbol{S}'\boldsymbol{W}) = d_W(\boldsymbol{S}, \boldsymbol{S}') \ . \tag{1.43}$$

This property is weaker than the affine invariance of the geometric distance. A natural question is whether such an affine invariant distance also exists on this manifold. Unfortunately, it is shown in [BS09] that the answer is negative for $R < P$. The proof is sufficiently simple to be derived here.

Theorem. *There is no continuous affine invariant distance on $S_{P,R}^{+}$ if $R < P$*

Proof. We show the result for $P = 2$ and $R = 1$; the demonstration can straightforwardly be extended to the other cases. The proof, from [BS09], is by contradiction. Assume that d is a continuous invariant distance on $S_{2,1}^{+}$. Consider $\boldsymbol{A} = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}$ and $\boldsymbol{B} = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}$, both in $S_{2,1}^{+}$. For $\varepsilon > 0$, consider the invert-

ible matrix $\boldsymbol{W}_\varepsilon = \begin{pmatrix} 1 & 0 \\ 0 & \varepsilon \end{pmatrix}$. We have: $\boldsymbol{W}_\varepsilon \boldsymbol{A} \boldsymbol{W}_\varepsilon^\top = \boldsymbol{A}$, and $\boldsymbol{W}_\varepsilon \boldsymbol{B} \boldsymbol{W}_\varepsilon^\top = \begin{pmatrix} 1 & \varepsilon \\ \varepsilon & \varepsilon^2 \end{pmatrix}$. Hence, as ε goes to 0, we have $\boldsymbol{W}_\varepsilon \boldsymbol{B} \boldsymbol{W}_\varepsilon^\top \to \boldsymbol{A}$ Using affine invariance, we have: $d(\boldsymbol{A}, \boldsymbol{B}) = d(\boldsymbol{W}_\varepsilon \boldsymbol{A} \boldsymbol{W}_\varepsilon^\top, \boldsymbol{W}_\varepsilon \boldsymbol{B} \boldsymbol{W}_\varepsilon^\top)$. Letting $\varepsilon \to 0$ and using continuity of d yields $d(\boldsymbol{A}, \boldsymbol{B}) = d(\boldsymbol{A}, \boldsymbol{A}) = 0$, which is absurd since $\boldsymbol{A} \neq \boldsymbol{B}$. □

To close this section, we have now compactly represented a M/EEG time series $\boldsymbol{x}(t)$ by its covariance matrix $\boldsymbol{C}$ in a certain frequency band. This matrix can either be full rank $\boldsymbol{C} \in \mathcal{S}_P^{++}$ or most often rank-deficient $\boldsymbol{C} \in \mathcal{S}_{P,R}^{+}$. In any case, the only way to vectorize a covariance matrix while enjoying the affine invariance property is by projecting it into a tangent space. Only once a matrix has been vectorized this way, can we use it as features to a standard regression algorithm. We will now present our prior knowledge on the M/EEG data generating mechanism and detail classical approaches to predict from such signals.

1.4 Generative models of M/EEG signals and outcome

In this thesis we are interested in predicting a continuous bio-medical neuro-outcome $y \in \mathbb{R}$ from brain activity, measured by MEG/EEG and represented by the multivariate signal $\boldsymbol{x}(t) \in \mathbb{R}^P$, where P corresponds to the number of sensors. As prior knowledge about this task we will derive a neurophysiological generative model of brain activity, and its approximation by a statistical generative model. Note that, in this thesis, we use the term *generative model* in the statistical sense of a probabilistic model of the M/EEG observations and the biomedical outcomes.

1.4.1 Prior knowledge

Physiological generative model. We assume the existence of $M \gg P$ electrical *physiological sources* in the brain that emerge from the synchronous activity of cortical layer IV pyramidal neurons [Häm+93]. The activity of these neural current generators form the time series $\boldsymbol{z}(t) \in \mathbb{R}^M$, where t represents time. These sources can be thought of as localized current sources, such as a patch of cortex with synchronously firing neurons, or a large set of patches forming a network. The underlying assumption is that these unobservable physiological sources are at the origin of the M/EEG signals $\boldsymbol{x}(t)$, and that they are statistically related to y. Often they are even the actual generators of y, *e.g.,* , the neurons producing the finger movement of a person. Here, we embrace the statistical machine learning paradigm where one aims to learn a predictive model from a set of N labeled training samples, $(\boldsymbol{x}_i(t), y_i), i = 1, \ldots, N$, which we see, fundamentally, as a function approximation problem. We will consider predicted outcomes that do not depend on time. The physics of the problem and the linearity of the quasi-static approximation of Maxwell's equations guarantee that MEG/EEG acquisition is linear too: the signals measured are obtained by linear combination of the underlying physiological sources. This leads to:

$$\boldsymbol{x}_i(t) = \boldsymbol{G}_i \, \boldsymbol{z}_i(t) \ , \tag{1.44}$$

where $\boldsymbol{G}_i \in \mathbb{R}^{P \times M}$ is the leadfield, also commonly referred to as gain matrix. Therefore, the observed M/EEG signal $\boldsymbol{x}_i(t) \in \mathbb{R}^P$ recorded by the external sensors contains information on unobserved brain internal sources $\boldsymbol{z}_i(t) \in \mathbb{R}^M$, distorted by individual brain anatomy represented by $\boldsymbol{G}_i$. Note that here the j-th column of $\boldsymbol{G}_i$ is not necessarily constrained to be the forward model of a focal electrical current dipole in the brain. It can also correspond to large distributed sources. Besides, the neuro-outcome is also related to the sources through an unknown function:

$$y_i = \phi(\boldsymbol{z}_i(t)) \ . \tag{1.45}$$

This reality is illustrated as the area outside the cloud in Fig. 1.2.

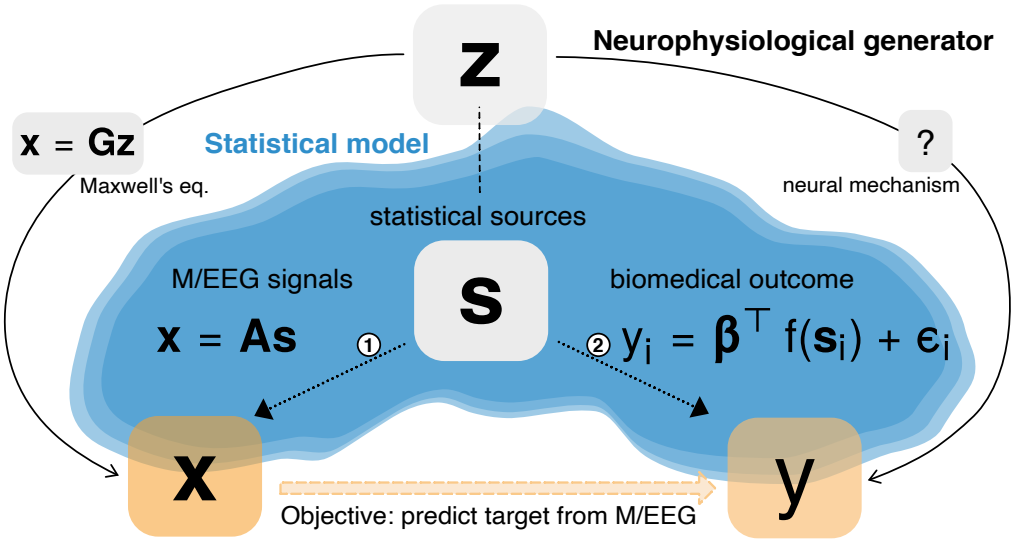


Fig. 1.2: Generative model for regression with M/EEG. Unobservable neuronal activity z gives rise to observed M/EEG data $\boldsymbol{X}$ and an observed biomedical outcome y. The M/EEG data $\boldsymbol{X}$ is obtained by linear mixing of z through the leadfield $\boldsymbol{G}$. The outcome y is derived from z through often unknown neural mechanisms. The statistical model (blue cloud) approximates the neurophysiological data-generating mechanisms with two sub-models, one for the M/EEG signals $\boldsymbol{X}$ (path 1), one for the biomedical outcome y (path 2). Both models are based on a vector s of uncorrelated statistical sources that, may refer to localized cortical activity or synchronous brain networks. The ensuing model generates y from a linear combination of the statistical sources s. The generative model of $\boldsymbol{X}$ follows the ICA model [HO00] and assumes linear mixing of the source signals by $\boldsymbol{A}$, interpreted as a linear combination of the columns of the leadfield $\boldsymbol{G}$. The generative model of y assumes a linear model in the parameters β but allows for non-linear functions in the data, such as the power or the log-power. The mechanisms governing path 1 implies that the sources s appear geometrically distorted in $\boldsymbol{X}$. This makes it impossible for a linear model to accurately capture this distortion if y, in path 2, is generated by a non-linear function of s. This study focuses on how to mitigate this distortion without biophysical source modeling when performing regression on M/EEG source power.

Statistical generative model of the M/EEG signals $\boldsymbol{x}_i(t)$. We simplify this physiological model by a statistical model inspired by Independent Component Analysis [HO00, ICA], a popular approach to model M/EEG signals [Mak+96]. The M physiological sources are modelled by $Q \leq P$ *statistical sources* $\boldsymbol{s}(t) \in \mathbb{R}^Q$, that correspond to unknown latent variables. Again, these variables are assumed to be linearly related to measured signal $\boldsymbol{x}(t)$ (a valid generative model for M/EEG data called the *linear instantaneous mixing model* [Häm+93]) and to be statistically related to the outcome variable y. The area inside the cloud depicted in Fig. 1.2 illustrates the statistical generative models. We consider an extension of noise-free Blind Source Separation [Bel+97] and assume the measured signal arises from the activity of Q statistical sources, contaminated by an additive noise:

$$\boldsymbol{x}(t) = \sum_{j=1}^{Q} s_j(t)\boldsymbol{a}_j^s + \boldsymbol{n}(t) \ . \tag{1.46}$$

The sensor signal $\boldsymbol{x}(t) \in \mathbb{R}^P$ is a linear combination of Q unit vectors in $\mathbb{R}^P$ called the *source patterns*. Each of the Q source pattern $\boldsymbol{a}_j^s$ is the sensor measure of a unit-amplitude source and is weighted by the corresponding source amplitude $s_j(t) \in \mathbb{R}$. This model is conveniently written in matrix form, for each sample $i = 1 \dots N$:

$$\boldsymbol{x}_i(t) = \boldsymbol{A}_i^s \, \boldsymbol{s}_i(t) + \boldsymbol{n}_i(t) \ , \tag{1.47}$$

where $\boldsymbol{s}_i(t) \in \mathbb{R}^Q$ is the *source vector* formed by the time series of the Q sources amplitude of sample i and $\boldsymbol{n}_i(t) \in \mathbb{R}^P$ is the contamination due to noise. The columns of the time-independent *mixing matrix* $\boldsymbol{A}^s \in \mathbb{R}^{P \times Q}$ are the Q linearly independent *source patterns* [Hau+14], which correspond to topographies on the sensor array: $\boldsymbol{A}^s = [\boldsymbol{a}_1^s, \dots, \boldsymbol{a}_Q^s] \in \mathbb{R}^{P \times Q}$. Each quantity in the right-hand side of Eq. (1.47), $\boldsymbol{A}$, $\boldsymbol{s}_i(t)$ and $\boldsymbol{n}_i(t)$, is unknown and should be inferred from $\boldsymbol{x}_i(t)$. This setting encompasses both event-level regression, where the samples $\boldsymbol{x}_i(t)$ are multiple epochs of signal from a unique subject (i stands for a particular time window), and subject-level regression where the samples represent the full signal of multiple subjects (i then stands for a particular subject).

The following proposition shows that, under certain assumptions, our generative model (1.47) has a full-rank formulation.

Proposition : Full-rank formulation of M/EEG signal generative model

Under the following statistical model assumptions:

- ▷ The signal $\boldsymbol{x}(t) \in \mathbb{R}^P$ arises from $Q < P$ statistical sources contaminated by additive noise
- ▷ The source space is the same for all samples and of dimension Q,
- ▷ The noise space is the same for all samples,
- ▷ The source and noise spaces are not mixed (*i.e.,* in direct sum in $\mathbb{R}^P$),

the statistical model (1.47) can be compactly rewritten as :

$$\boldsymbol{x}_i(t) = \boldsymbol{A}\boldsymbol{\eta}_i(t) \ , \tag{1.48}$$

where $\boldsymbol{A} \in \mathbb{R}^{P \times P}$ is an invertible matrix (which includes source and noise patterns) and $\boldsymbol{\eta}_i(t) \in \mathbb{R}^P$ is the concatenation of source and noise signals.

Proof

We assume that the measured signal arises from the activity of sources, contaminated by additive noise: $\boldsymbol{x}_i(t) = \boldsymbol{x}_i^s(t) + \boldsymbol{x}_i^n(t) \in \mathbb{R}^P$.
Since the sources span a space of dimension Q, shared across all samples, there exists Q linearly independent vectors of $\mathbb{R}^P$ forming the columns of a sample-independent matrix $\boldsymbol{A}^s = [\boldsymbol{a}_1^s, \dots, \boldsymbol{a}_Q^s]$ such that $\boldsymbol{x}_i^s(t) = \boldsymbol{A}^s \boldsymbol{s}_i(t)$ with $\boldsymbol{s}_i(t) \in \mathbb{R}^Q$.
We say that two subspaces are in direct sum in $\mathbb{R}^P$ (or that they are supplementary in $\mathbb{R}^P$, or that $\mathbb{R}^P$ is the co-product of the two subspaces) if any vector in $\mathbb{R}^P$ can be uniquely decomposed as a sum of vectors from these subspaces. This is equivalent to saying that the juxtaposition of their basis forms as basis of $\mathbb{R}^P$, or, as $\mathbb{R}^P$ has a finite dimension, that the sum of their dimension is P and their intersection is reduced to the null vector. We then informally say that the subspaces are not mixed.
Since noise space is in direct sum with the source space in $\mathbb{R}^P$, and is shared across samples, there exists $P - Q$ linearly independent vectors of $\mathbb{R}^P$ forming the columns of a sample-independent matrix $\boldsymbol{A}^n = [\boldsymbol{a}_1^n, \dots, \boldsymbol{a}_{P-Q}^n] \in \mathbb{R}^{P \times (P-Q)}$ such that $\boldsymbol{x}_i^n(t) = \boldsymbol{A}^n \boldsymbol{n}_i(t)$ with $\boldsymbol{n}_i(t) \in \mathbb{R}^{P-Q}$.
So we have $\boldsymbol{x}_i(t) = \boldsymbol{A}^s \boldsymbol{s}_i(t) + \boldsymbol{A}^n \boldsymbol{n}_i(t)$. Denoting $\boldsymbol{A} = [\boldsymbol{A}^s, \boldsymbol{A}^n] = [\boldsymbol{a}_1^s, \dots, \boldsymbol{a}_Q^s, \boldsymbol{a}_1^n, \dots, \boldsymbol{a}_{P-Q}^n] \in \mathbb{R}^{P \times P}$ and $\boldsymbol{\eta}_i(t) = [s_{i,1}(t), \dots s_{i,Q}(t), n_{i,1}(t), \dots, n_{i,P-Q}(t)] \in \mathbb{R}^P$, the generative model can be rewritten as:

$$\boldsymbol{x}_i(t) = \boldsymbol{A}\boldsymbol{\eta}_i(t) \ ,$$

The matrix $\boldsymbol{A}$ is invertible since the source and noise subspaces span all $\mathbb{R}^P$.

By making additional statistical assumptions, the following proposition shows that the covariance matrix of the M/EEG signal has a very particular structure:

Proposition : Structure of covariance matrix of M/EEG signal

If we further assume that:

- ▷ the sources are zero-mean and uncorrelated,
- ▷ the sources are uncorrelated from the noise,

the covariances are full rank and given by:

$$\boldsymbol{C}_i = \boldsymbol{A}\boldsymbol{E}_i\boldsymbol{A}^\top \ , \tag{1.49}$$

where $\boldsymbol{E}_i$ is a block diagonal matrix, whose upper $Q \times Q$ block is $\mathrm{diag}(\boldsymbol{p}_i)$ with $\boldsymbol{p}_i = \mathbb{E}[\boldsymbol{s}_i^2(t)] \in \mathbb{R}^Q$, the power of the sources of sample i.

Proof

Recalling from (1.48) that $\boldsymbol{x}_i(t) = \boldsymbol{A}\boldsymbol{\eta}_i(t)$ we have that $\boldsymbol{C}_i = \mathbb{E}[\boldsymbol{x}_i(t)\boldsymbol{x}_i(t)^\top = \boldsymbol{A}\boldsymbol{E}_i\boldsymbol{A}^\top$ where $\boldsymbol{E}_i = \mathbb{E}[\boldsymbol{\eta}_i(t)\boldsymbol{\eta}_i(t)^\top]$. If we assume that the components of the sources of a given sample i are zero-mean and uncorrelated, the covariance matrix of sources is diagonal: $\mathbb{E}[\boldsymbol{s}_i(t)\boldsymbol{s}_i^\top(t)] = \mathrm{diag}(\boldsymbol{p}_i)$, where $p_{i,j} = \mathbb{E}[s_{i,j}^2(t)]$ is the power, *i.e.,* the variance over time of the j-source of sample i. If they are also uncorrelated from the noise, we have $\mathbb{E}[\boldsymbol{s}_i(t)\boldsymbol{n}_i(t)^\top] = 0$. As a result $\boldsymbol{E}_i$ is a block diagonal matrix, whose upper $Q \times Q$ block is the covariance of sources $\mathrm{diag}(\boldsymbol{p}_i)$ and the $(P-Q) \times (P-Q)$ lower block is the covariance of the noise. $\boldsymbol{E}_i = \begin{bmatrix} \ddots & & 0 & \\ & (p_{i,j})_j & & \mathbf{0} \\ 0 & & \ddots & \\ & \mathbf{0} & & \boldsymbol{M} \end{bmatrix}$ In particular, these covariances $\boldsymbol{C}_i$ are full rank.

Note that this statistical generative model is a simplification of the biophysical generative mechanism: the number of true sources may exceed the number of sensors, $M \gg P$ (whereas we assume $Q < P$), the source and noise spaces may not be the same for all samples, as for instance the gain $\boldsymbol{G}_i$ is sample-dependent in subject-level regression (whereas $\boldsymbol{A}$ is sample-independent), the real sources $\boldsymbol{z}_i$ may not be uncorrelated [Nol+06].

The assumption that the noise subspace is not mixed with the source subspace is motivated by the fact that environmental perturbations (by definition independent from brain activity) generate the strongest noise in M/EEG recordings. On the other hand, physiological noise, due to cardiac or ocular activity, systematically interacts with brain signals and is necessarily captured by the statistical sources $\boldsymbol{s}$. Overall, these assumptions may not be realistic but are useful for modeling purposes. Model violations will be addressed in section 1.5.2.

Statistical generative model of the biomedical outcome y. As we know that powers reveals cognition [Eng+20], the proposed framework models y_i as a function of the sources powers:

$$y_i = \boldsymbol{\beta}^\top f(\boldsymbol{p}_i) \ , \tag{1.50}$$

where $\boldsymbol{p}_i = \mathbb{E}_t[\boldsymbol{s}_i^2(t)] \in \mathbb{R}^Q$ is the power of sources of sample i, $f : \mathbb{R}^+ \to \mathbb{R}$ is a known increasing function (applied component-wise to a vector) and $\boldsymbol{\beta} \in \mathbb{R}^Q$ are regression coefficients.

Linear models in the sources powers (f = identity) or log-powers ($f = \log$) are commonly used in the neuroscience literature and support numerous statistical learning models on M/EEG [Bla+08; Däh+14a; GB08]. In particular Buzsáki and

Mizuseki [BM14] discusses a wide body of evidence arguing in favor of log-linear relationships between brain dynamics and cognition. Both possibilities for f will be considered in below sections.

According to (1.50), y_i is related to the power of the sources hence to $\mathrm{diag}(\boldsymbol{p}_i)$. According to (1.49) the sensor signal $\boldsymbol{x}_i$, through its covariance matrix $\boldsymbol{C}_i$, also contain information about powers of the sources $\boldsymbol{p}_i$ in $\boldsymbol{E}_i$ but this information is noisy and distorted through unknown linear field spread $\boldsymbol{A}$. As our task is to uncover the relationship between y_i and $\boldsymbol{x}_i$, this unknown mixing makes it challenging to find optimal regression algorithms with no approximation error.

The broadband covariance (computed on the raw signal without temporal filtering) largely reflects low-frequency power as consequence of the predominant 1/f power spectrum, hence, is rarely of interest for predicting. In practice, one prefers frequency specific models, where the previous relationships are obtained after $\boldsymbol{s}_i(t)$ has been bandpass filtered in a specific frequency range. In frequency-specific models, the powers are replaced by band-powers: power of the source in the chosen frequency band. Note that source power in a given frequency band is simply the variance of the signal in that frequency band.

1.4.2 The classical approaches to predict from M/EEG observations

We will now present three family of approaches classically used to predict from M/EEG observations. We will see that, as each of those methods are well-suited in certain contexts, they all fall short for our specific task of prediction in a clinical setting.

Biophysical sources modeling. Since both our input $\boldsymbol{x}_i(t)$ and output y_i are related to physiological sources $\boldsymbol{z}_i(t)$, the most natural approach is to get back to our physiological generative model and try to estimate the sources before fitting a regression model. One important family of approaches for predictive modeling with M/EEG is therefore relying on explicit biophysical source modeling. It consists in estimating the locations, amplitudes and extents of the sources from the MEG/EEG data. This estimation is known as the inverse problem [Bai17]. To solve it, anatomically constrained inverse methods are used to infer the most likely electromagnetic source configuration given the observations [Häm+93]. Common techniques rely on fitting electrical-current dipoles [MLL92] or involve penalized linear inverse models to estimate the current distribution over a pre-specified dipole grid [HI94; Lin+06; VVB88; HS14]. Anatomical prior knowledge is injected through the well-defined forward model: Maxwell equations enable computing leadfields from the geometry and composition of the head, which predict propagation from a known source to the sensors [Häm+93; MLL99]. Let us denote $\boldsymbol{G} \in \mathbb{R}^{P \times Q}$ the

instantaneous mixing matrix that relates the sources in the brain to the MEG/EEG measurements. Here $Q \geq P$ corresponds to the number of candidate sources in the brain. This forward operator matrix is obtained by solving numerically Maxwell's equations after specifying a geometrical model of the head, typically obtained using an anatomical MRI image [HP17]. Using $\boldsymbol{G}$ the Minimum Norm Estimate (MNE) source imaging technique [HI84] offers a way to solve the inverse problem. MNE can be seen as standard Tikhonov regularized estimation, also similar to a ridge regression in statistics, and is therefore linear. Using such problem formulation, the sources are obtained from the measurements with a linear operator which is given by:

$$\boldsymbol{W}_{\text{MNE}} = \boldsymbol{G}^\top (\boldsymbol{G}\boldsymbol{G}^\top + \lambda \boldsymbol{I}_P)^{-1} \in \mathbb{R}^{Q \times P} \ . \tag{1.51}$$

From a signal-processing standpoint, when these steps lead to a linear estimation of the sources, the rows of this linear operator $\boldsymbol{W}_{\text{MNE}}$ can be seen also as spatial filters that are mapped to specific locations in the brain. MNE approach can then be thought of as biophysical spatial filtering, informed by the individual anatomy of each subject. From the estimated sources, one can then learn to predict y as the distortions induced by individual head geometry are mitigated, see for example [Wes+18; Kie+19; Kha+18]. While approaching the problem from this perspective has important benefits, such as the ability to exploit the geometry and the physical properties of the head tissues of each subject, there are certain drawbacks. First, the inverse problem is ill-posed and notoriously hard to solve. Second, computing $\boldsymbol{G}_i$ requires costly T1-weighted MRI acquisitions and time-consuming manual labor by experienced MEG/EEG practitioners [Bai17]: precise measure of the head in the MEG device coordinate system, anatomical coregistration and tedious data-cleaning to mitigate electromagnetic artefacts caused by environmental or physiological sources of non-interest outside of the brain. Using a MRI template, *e.g.,* MNI brain, would alleviate this issue but amounts to consider a common average brain, ignoring inter-individuals anatomical variability. Hence, this method is likely to yield suboptimal performances, as hinted by a recent work [Eng+21] benchmarking different models for brain age prediction. Besides, it is still costly in terms of computation and rest on the relevance of the MNI brain model. This gold-standard approach of biophysical sources modeling is therefore hard to automate and poses challenges to clinical practice. This justifies the statistical generative model approximation used in this thesis enabling to learn a regression model without biophysical source modeling.

Statistical sources modeling. A second family is motivated by unsupervised decomposition techniques such as Independent Component Analysis [HO00; Mak+97], which yield estimates of maximally independent statistical sources that can be used for prediction and corresponding spatial filters [SNS14; WM09; SG10]. In general, spatial filtering consists in computing linear combinations of the original P sensors signals to produce so-called 'spatially filtered' signals, or 'source' signals,

$\boldsymbol{W}^\top \boldsymbol{x}_i$. The weights of the combination form a spatial filter. Considering $R \leq P$ filters, it corresponds to the columns of the matrix $\boldsymbol{W} \in \mathbb{R}^{P \times R}$ of rank R which is common to all samples (*e.g.,* subjects). If $R < P$, then spatial filtering reduces the dimension of the data. The covariance matrices of 'spatially filtered' signals $\boldsymbol{W}^\top \boldsymbol{x}_i \in \mathbb{R}^R$ is readily obtained as: $\boldsymbol{\Sigma}_i = \boldsymbol{W}^\top \boldsymbol{C}_i \boldsymbol{W} \in \mathbb{R}^{R \times R}$. With probability one, $\operatorname{rank}(\boldsymbol{\Sigma}_i) = \min(\operatorname{rank}(\boldsymbol{W}), \operatorname{rank}(\boldsymbol{C}_i)) = R$, hence $\boldsymbol{\Sigma}_i \in S_R^{++}$. Since the $\boldsymbol{C}_i$'s do not span the same image, applying $\boldsymbol{W}$ destroys some information. Recently, geometry-aware dimensionality reduction techniques, both supervised and unsupervised, have been developed on covariance manifolds [HYS16; HSH17]. Many different spatial filters have been designed to produce virtual signals that help the prediction task. Such methods model the data as an independent set of statistical sources that are entangled by a so-called mixing matrix, often interpreted as the leadfields. Here, the sources are purely statistical objects and no anatomical notion applies directly. In practice, unsupervised spatial filters are often combined with source modeling and capture a wide array of situations ranging from single dipole-sources to entire brain-networks [HIN09; Bro+11; Del+12]. Being unsupervised, hence blind to the target y, these methods are not optimal for regression. Also, each ICA filters is fitted independently for each sample (subject) making it difficult to compare the resulting filters.

Linear models in sensor-space. Finally, a third family directly applies general-purpose machine learning directly on sensor space signals $\boldsymbol{x}_i(t)$ without explicitly considering the data generating mechanism. Following a common trend in other areas of neuroimaging research [Dad+19; Sch+19; He+19], linear prediction methods have turned out extraordinarily well-suited for this task, *i.e.,* , logistic regression [And+15], linear discriminant analysis [War+16], linear support vector machines [Kin+13]. The success of linear models deserves separate attention as these methods enable remarkable predictive performance with simplified fast computation [Par+05]. While interpretation and incorporation of prior knowledge remain challenging, significant advances have been made in the past years. This has led to novel methods for specifying and interpreting linear models [Hau+14; VS19]. Recent work has even suggested that for the case of learning from evoked responses, linear methods are compatible with the statistical models implied by source localization and unsupervised spatial filtering [Kin+18; KD14; SWS15]. Indeed, if the outcome is linear in the source signal, *i.e.,* , due to the linear superposition principle, the mixing in the input amounts to a linear transform of the sources that can therefore be captured by a linear model with sufficient data. Additional source localization or spatial filtering should therefore be unnecessary in this case.

On the other hand, the situation is more complex when predicting outcomes from brain rhythms, *e.g.,* , induced responses [TBB99] or spontaneous oscillations. As brain-rhythms are not strictly time-locked to external events, they cannot be accessed

by averaging. Instead, they are commonly represented by the signal power in shorter or longer time windows and often give rise to log-linear models [BM14; RBB15]. A consequence of such non-linearities is that it cannot be readily captured by a linear model. Moreover, simple strategies such as log-transforming the power estimates only address the issue when applied at the source-level: the leadfields have already spatially smeared the signal presented on the sensors.

Alternative approaches. This leads back to spatial filtering approaches. Beyond source localization and unsupervised filtering, supervised spatial filtering methods have recently become more popular beyond the context of BCIs. These methods solve generalized eigenvalue problems to estimate coordinate systems constructed with regard to criteria relevant for prediction. For example, spatio-spectral-decomposition (SSD) is an unsupervised technique that enhances SNR with regard to power in surrounding frequencies [NNC11]. On the other hand, Common Spatial Patterns [Kol91], Joint Decorrelation [CP14] and Source Power Comodulation (SPoC) focus on correlation with the outcome [Bla+08; Däh+14a; Däh+13], MultiView ICA [Ric+20] extends the ICA model to group studies, whereas [Dmo+12] have proposed variants of Canonical Correlation Analysis (CCA) [Hot92; Däh+14b] without orthogonality constraint to focus on shared directions of variation between related datasets or by proposing shared envelope correlations as optimization target [Däh+14b]. This yields a two-step procedure: 1) spatial filters model the correlation induced by the leadfields and provide unmixed time series 2) some non-linear transforms such as logarithms are applied to these time series as the validity of linear equations is now secured.

A more recent single-step approach consists in learning directly from spatially correlated power-spectra with linear models and Riemannian geometry [Bar+11; Bar+13; YBL17; RJC19; Fru+17]. This mathematical framework, introduced in Section 1.3, provides principles to correct for the geometric distortions arising from linear mixing of non-linear sources. These models are blind to the linear mixing $\boldsymbol{A}^s$ and working with the signals $\boldsymbol{x}$ is similar to working directly with the sources $\boldsymbol{s}$. Riemannian geometry is a natural setting where such affine invariance properties are found [FM03]. It allows to represent the covariance matrices used for representing the M/EEG signal as Euclidean objects for which linear models apply. This approach has turned out to be promising for enhancing classification of event-level data and has been the important ingredient of several winning solutions in recent data analysis competitions, *e.g.,* , the seizure prediction challenge organized by the University of Melbourne in 2016. Classification based on tangent vectors has also been used in [Bar+12] for BCI classification. Recently, this approach has been explored for prediction of subject-level brain volume from clinical EEG in Alzheimer's disease in about 100 patients [Fru+17].

We have presented classical methods for regressing an outcome from M/EEG signals. Biophysical sources modeling using anatomically constrained inverse methods corrects for distortions induced by individual head anatomy but is not scalable as it requires MRI scans and MEG manual expertise: it is therefore not well adapted to clinical practice. Statistical sources modeling using unsupervised spatial filtering (*e.g.,* ICA) are blind to the prediction target hence not optimal for regression. Sensor-space linear modeling leverages the power of linear models but are not optimal when predicting from brain rhythms, hence not adapted to our generative models assumptions. We will now present how to mitigate the distortion without biophysical source modeling when performing regression on M/EEG source power. We will see that we can indeed overcome volume conduction with an appropriate regression algorithm, adapted to the generative process.

1.5 A family of statistically consistent regression algorithms

To recall our generative model setup: our input is a P-dimensional signal $\boldsymbol{x}(t)$ arising from $Q < P$ sources and additive noise. Assuming the source and noise spaces are unmixed, shared across samples and of respective dimension Q and $P - Q$ then we can write $\boldsymbol{x}_i(t) = \boldsymbol{A}\boldsymbol{\eta}_i(t)$ where $\boldsymbol{A} \in \mathbb{R}^{P \times P}$ is an invertible matrix (which includes source and noise patterns) and $\boldsymbol{\eta}_i(t) \in \mathbb{R}^P$ is the concatenation of source $\boldsymbol{s}_i(t)$ and noise signals. If the sources are zero-mean, uncorrelated and uncorrelated from the noise, then the covariances are full-rank and writes $\boldsymbol{C}_i = \boldsymbol{A}\boldsymbol{E}_i\boldsymbol{A}^\top$ where $\boldsymbol{E}_i$ is a block diagonal matrix, whose upper $Q \times Q$ block is $\mathrm{diag}(\boldsymbol{p}_i)$ with $\boldsymbol{p}_i = \mathbb{E}[\boldsymbol{s}_i^2(t)] \in \mathbb{R}^Q$, the power of the sources of sample i. Our output is a continuous target y_i modelled as a function of these source powers $y_i = \boldsymbol{\beta}^\top f(\boldsymbol{p}_i)$ without additional noise. In this section, we assume these ideal conditions hold (fixed volume conduction across samples, full rank signals, no noise in target).

Given this generative model as our prior knowledge, our goal is to find a regression algorithm to predict the target y_i from sensor-space M/EEG signal $\boldsymbol{x}_i(t)$, that do not require to estimate the sources $\boldsymbol{s}_i(t)$ and that has no approximation error. Our input signal is assumed to depend linearly on the sources, whereas our target on the power of sources (*i.e.,* the squared amplitude of the source signal). This non-linear dependence hints at using non-linear models. Deep learning models have shown strong performance in learning non-linear functions but require lots of data. Scarcity of high-dimensional medical data would therefore favor the use of linear mechanics, applied to non-linear features of the M/EEG signal (Generalized Linear Model strategy). Its covariance matrix $\boldsymbol{C}_i$ seems a good candidate as it contains powers & cross-powers of sensors. This representation is the adequate approximation of the neural signals at the second order and is low-dimensional. These covariances, computed at the sensor-level, also contain information about the powers of the sources ($\in \boldsymbol{E}_i$) but this information is noisy and distorted through field spread $\boldsymbol{A}$. Can we get rid of it?

In this thesis, we introduce four different regression algorithms: Upper, Riemann, Wasserstein and SPoC. They are all based on a linear model, applied to carefully chosen vectorization $\boldsymbol{v}_i$ of the covariance $\boldsymbol{C}_i$. Showing that these models are statistically consistent amounts to proving that the real relationship between y_i and $\boldsymbol{v}_i$ is linear. In particular, we show that different functions f yield a linear relationship between the y_i's and the $\boldsymbol{v}_i$'s for different Riemannian metrics, hence show that these four regression models successfully achieve statistical consistency for different generative assumptions.

More specifically, throughout this study, we consider the following regression pipeline. Given a training set of samples $\boldsymbol{x}_1(t), \dots, \boldsymbol{x}_N(t) \in \mathbb{R}^P$ and target continuous variables $y_1, \dots, y_N \in \mathbb{R}$, we first compute the covariances of each sample $\boldsymbol{C}_1, \dots, \boldsymbol{C}_N \in \mathbb{R}^{P \times P}$. After computing their vectorization $\boldsymbol{v}_1, \dots, \boldsymbol{v}_N \in \mathbb{R}^K$ (*cf.* below), a linear regression technique (e.g. ridge regression) with parameters $\boldsymbol{\beta} \in \mathbb{R}^K$ can be employed assuming that $y_i \simeq \boldsymbol{v}_i^\top \boldsymbol{\beta}$.

1.5.1 Four statistically consistent regression algorithms

Proposition : Consistency of Upper regression algorithm

The **Upper regression algorithm** consists in taking the Euclidean vectorization:

$$\boldsymbol{v}_i = \text{Upper}(\boldsymbol{C}_i) \in \mathbb{R}^{\frac{P(P+1)}{2}} \quad , \tag{1.52}$$

where $\text{Upper}(\boldsymbol{M})$ is defined as the vector containing the upper triangular coefficients of $\boldsymbol{M}$, with off-diagonal terms weighted by a factor $\sqrt{2}$. This weighting ensures that the vector and the matrix have same norms ($\|\text{Upper}(\boldsymbol{M})\|_2 = \|\boldsymbol{M}\|_F$).
This model is statistically consistent in the particular case where $f =$ identity: the relationship between y_i and $\text{Upper}(\boldsymbol{C}_i)$ is linear.

Proof

We assume $f(p) = p$. Rewriting Eq. (1.49) as $\boldsymbol{E}_i = \boldsymbol{A}^{-1} \boldsymbol{C}_i \boldsymbol{A}^{-\top}$, and since the $p_{i,j}$ are on the diagonal of the upper block of $\boldsymbol{E}_i$, the relationship between the $p_{i,j}$ and the coefficients of $\boldsymbol{C}_i$ is also linear. Since the variable of interest y_i is linear in the coefficients of $\boldsymbol{p}_i$, it is also linear in the coefficients of $\boldsymbol{C}_i$, hence linear in the coefficients of $\boldsymbol{v}_i$. In other words, y_i is a linear combination of the vectorization of $\boldsymbol{C}_i$ *w.r.t.* the standard Euclidean distance, hence the 'upper' regression algorithm is statistically consistent for $f =$ identity.

Note that this method cannot be generalized to an arbitrary spectral function f because $f(\boldsymbol{C}_i) \neq \boldsymbol{A}\, f(\boldsymbol{E}_i)\, \boldsymbol{A}^\top$.

Proposition : Consistency of Riemann regression algorithm

The **Riemann regression algorithm** consists in taking the Geometric vectorization:

$$\boldsymbol{v}_i = \mathcal{P}_{\overline{\boldsymbol{C}}}(\boldsymbol{C}_i) = \text{Upper}\left(\log\left(\overline{\boldsymbol{C}}^{-1/2}\boldsymbol{C}_i\overline{\boldsymbol{C}}^{-1/2}\right)\right) \in \mathbb{R}^{\frac{P(P+1)}{2}} \quad , \tag{1.53}$$

the vectorization of $\boldsymbol{C}_i$ *w.r.t.* the geometric distance using as reference $\overline{\boldsymbol{C}} = \text{Mean}_G(\boldsymbol{C}_1, \ldots, \boldsymbol{C}_N)$ the geometric mean of the dataset.
This model is statistically consistent in the particular case where $f = \log$: the relationship between y_i and $\mathcal{P}_{\overline{\boldsymbol{C}}}(\boldsymbol{C}_i)$ is linear.

Proof

The proof relies crucially on the affine invariance property: using Riemannian embeddings of the $\boldsymbol{C}_i$'s, is equivalent to working directly with the $\boldsymbol{E}_i$'s. First, we note that by invariance, $\overline{\boldsymbol{C}} = \text{Mean}_G(\boldsymbol{C}_1, \ldots, \boldsymbol{C}_N) = \boldsymbol{A}\text{Mean}_G(\boldsymbol{E}_1, \ldots, \boldsymbol{E}_N)\boldsymbol{A}^\top = \boldsymbol{A}\overline{\boldsymbol{E}}\boldsymbol{A}^\top$, where $\overline{\boldsymbol{E}}$ has the same block diagonal structure as the $\boldsymbol{E}_i$'s, and $\overline{\boldsymbol{E}}_{jj} = (\prod_{i=1}^N p_{i,j})^{\frac{1}{N}}$ for $j \leq Q$. Denote $\overline{\boldsymbol{U}} = \overline{\boldsymbol{C}}^{\frac{1}{2}}\boldsymbol{A}^{-\top}\overline{\boldsymbol{E}}^{-\frac{1}{2}}$. By simple verification, we obtain $\overline{\boldsymbol{U}}^\top\overline{\boldsymbol{U}} = I_P$, i.e. $\overline{\boldsymbol{U}}$ is orthogonal.
Furthermore, we have:

$$\overline{\boldsymbol{U}}^\top\overline{\boldsymbol{C}}^{-\frac{1}{2}}\boldsymbol{C}_i\overline{\boldsymbol{C}}^{-\frac{1}{2}}\overline{\boldsymbol{U}} = \overline{\boldsymbol{E}}^{-\frac{1}{2}}\boldsymbol{E}_i\overline{\boldsymbol{E}}^{-\frac{1}{2}} \quad . \tag{1.54}$$

It follows that for all i,

$$\overline{\boldsymbol{U}}^\top \log(\overline{\boldsymbol{C}}^{-\frac{1}{2}}\boldsymbol{C}_i\overline{\boldsymbol{C}}^{-\frac{1}{2}})\overline{\boldsymbol{U}} = \log(\overline{\boldsymbol{E}}^{-\frac{1}{2}}\boldsymbol{E}_i\overline{\boldsymbol{E}}^{-\frac{1}{2}}) \tag{1.55}$$

Note that $\log(\overline{\boldsymbol{E}}^{-\frac{1}{2}}\boldsymbol{E}_i\overline{\boldsymbol{E}}^{-\frac{1}{2}})$ shares the same structure as the $\boldsymbol{E}_i$'s, and that $\log(\overline{\boldsymbol{E}}^{-\frac{1}{2}}\boldsymbol{E}_i\overline{\boldsymbol{E}}^{-\frac{1}{2}})_{jj} = \log(\frac{p_{i,j}}{\bar{p}_j})$ for $j \leq Q$.
Therefore, the relationship between $\log(\overline{\boldsymbol{C}}^{-\frac{1}{2}}\boldsymbol{C}_i\overline{\boldsymbol{C}}^{-\frac{1}{2}})$ and the $\log(p_{i,j})_j$ is linear.
Finally, since $\boldsymbol{v}_i = \text{Upper}(\log(\overline{\boldsymbol{C}}^{-\frac{1}{2}}\boldsymbol{C}_i\overline{\boldsymbol{C}}^{-\frac{1}{2}}))$, the relationship between the $\boldsymbol{v}_i$'s and the $\log(p_{i,j})_j$ is linear, and the result holds.

Proof : Alternative proof

First, we note that by invariance, $\overline{\boldsymbol{C}} = \boldsymbol{A}\overline{\boldsymbol{E}}\boldsymbol{A}^\top$, where $\overline{\boldsymbol{E}}$ has the same block diagonal structure as the $\boldsymbol{E}_i$'s, and $\overline{\boldsymbol{E}}_{jj} = (\prod_{i=1}^N p_{i,j})^{\frac{1}{N}} \triangleq \overline{p}_j$ for $j \leq Q$. The vectorization is $\boldsymbol{v}_i = \text{Upper}\left(\log(\overline{\boldsymbol{C}}^{-1/2}\boldsymbol{C}_i\overline{\boldsymbol{C}}^{-1/2})\right)$. We observe that $\overline{\boldsymbol{C}}^{-1/2}\boldsymbol{C}_i\overline{\boldsymbol{C}}^{-1/2} = \overline{\boldsymbol{C}}^{-1/2}\left(\boldsymbol{C}_i\overline{\boldsymbol{C}}^{-1}\right)\overline{\boldsymbol{C}}^{1/2} = \boldsymbol{B}\boldsymbol{E}_i\overline{\boldsymbol{E}}^{-1}\boldsymbol{B}^{-1}$ with $\boldsymbol{B} = \overline{\boldsymbol{C}}^{-1/2}\boldsymbol{A}$ invertible. Therefore, $\log(\overline{\boldsymbol{C}}^{-1/2}\boldsymbol{C}_i\overline{\boldsymbol{C}}^{-1/2}) = \boldsymbol{B}\log\left(\boldsymbol{E}_i\overline{\boldsymbol{E}}^{-1}\right)\boldsymbol{B}^{-1}$, since matrix logarithm is equivariant by similarity. The Q values on the diagonal part of $\log(\boldsymbol{E}_i\overline{\boldsymbol{E}}^{-1})$ are the $\log\left(p_{i,j}/\overline{p}_j\right)$. In particular, by denoting $\boldsymbol{b}_j^{-1}$ the j-th row of $\boldsymbol{B}^{-1}$ and $\boldsymbol{b}_j$ the j-th column of $\boldsymbol{B}$, we find:

$$\log(p_{i,j}) = (\boldsymbol{b}_j^{-1})^\top \log(\overline{\boldsymbol{C}}^{-1/2}\boldsymbol{C}_i\overline{\boldsymbol{C}}^{-1/2})\boldsymbol{b}_j + \log(\overline{p}_j) \ . \tag{1.56}$$

This equation means that $\log(p_{i,j})$ is obtained as a linear combination of the coefficients in $\log(\overline{\boldsymbol{C}}^{-1/2}\boldsymbol{C}_i\overline{\boldsymbol{C}}^{-1/2})$, i.e. the coefficients of the vectorization $\boldsymbol{v}_i$. Since y_i is itself a linear combination of the $\log(p_{i,j})$, the advertised result holds.

As a side note, we have that $\|\boldsymbol{v}_i\|_2 = \left\|\log\left(\overline{\boldsymbol{C}}^{-1/2}\boldsymbol{C}_i\overline{\boldsymbol{C}}^{-1/2}\right)\right\|_F = \left\|\log\left(\overline{\boldsymbol{C}}^{-1}\boldsymbol{C}_i\right)\right\|_F = \mathrm{d}(\boldsymbol{C}_i, \overline{\boldsymbol{C}}) = \mathrm{d}(\boldsymbol{E}_i, \overline{\boldsymbol{E}})$, by affine-invariance of the geometric distance $d(\cdot)$ (see Appendix 1.3.3): the norm of $\boldsymbol{v}_i$ does not depend on $\boldsymbol{A}$, but only on the log source powers and noise.

The Riemannian embedding yields a representation of sensor-level power and its correlation structure relative to a common reference. In the particular case where $f = \log$, the idea is to normalize each covariance $\boldsymbol{C}_i$ by a common reference $\overline{\boldsymbol{C}}$, the geometric mean of covariances $\boldsymbol{C}_i$. We normalize using $\overline{\boldsymbol{C}}^{-1/2}\boldsymbol{C}_i\overline{\boldsymbol{C}}^{-1/2}$, which has the advantage over $\boldsymbol{C}_i\overline{\boldsymbol{C}}^{-1}$ to be symmetrical while having the same eigenvalues. We showed that a linear model applied to feature vector $\boldsymbol{v}_i = \mathcal{P}_{\overline{\boldsymbol{C}}}(\boldsymbol{C}_i) = \text{Upper}\left(\log\left(\overline{\boldsymbol{C}}^{-1/2}\boldsymbol{C}_i\overline{\boldsymbol{C}}^{-1/2}\right)\right)$ leads to a consistent regression algorithm. This, essentially, means taking the log of $\boldsymbol{C}_i$ after it has been whitened by $\overline{\boldsymbol{C}}^{-1/2}$, making the quantity of interest relative to some reference $\overline{\boldsymbol{C}}$ that hopefully will get rid of mixing matrix A. In terms of Riemannian geometry this is the projection of covariance matrix $\boldsymbol{C}_i$ to a common Euclidean space: the tangent space at $\overline{\boldsymbol{C}}$. In particular the norm of $\boldsymbol{v}_i$ can be interpreted as the (geometric) distance between $\boldsymbol{C}_i$ and $\overline{\boldsymbol{C}}$ and does not depend on $\boldsymbol{A}$. Essentially, the Riemannian approach projects out fixed linear spatial mixing through the whitening with the common reference. Finally, even though the geometric mean is the most natural reference on the positive definite manifold, consistency of the Riemann regression algorithm still holds when using the Euclidean mean as the common reference point. Indeed, a recent study on fMRI-based predictive modeling has reported negligible differences between the two options [Dad+19, appendix A].

Proposition : Consistency of Wasserstein regression algorithm

The **Wasserstein regression algorithm** consists in taking the Wasserstein vectorization:

$$\boldsymbol{v}_i = \mathcal{P}_{\overline{\boldsymbol{C}}}(\boldsymbol{C}_i) \in \mathbb{R}^{\frac{P(P+1)}{2}} \quad , \tag{1.57}$$

the vectorization of $\boldsymbol{C}_i$ *w.r.t.* the Wasserstein distance using as reference $\overline{\boldsymbol{C}} = \text{Mean}_W(\boldsymbol{C}_1, \dots, \boldsymbol{C}_N)$ the Wasserstein mean of the dataset.
This model is statistically consistent in the particular case where $f = \sqrt{\cdot}$ and $\boldsymbol{A}$ orthogonal: the relationship between y_i and $\mathcal{P}_{\overline{\boldsymbol{C}}}(\boldsymbol{C}_i)$ is linear.

Proof

First, we note that $\boldsymbol{C}_i = \boldsymbol{A}\boldsymbol{E}_i\boldsymbol{A}^\top \in \mathcal{S}_P^{++} = \mathcal{S}_{P,P}^{+}$ so it can be decomposed as $\boldsymbol{C}_i = \boldsymbol{Y}_i\boldsymbol{Y}_i^\top$ with $\boldsymbol{Y}_i = \boldsymbol{A}\boldsymbol{E}_i^{\frac{1}{2}}$.
By orthogonal invariance, $\overline{\boldsymbol{C}} = \text{Mean}_W(\boldsymbol{C}_1, \dots, \boldsymbol{C}_N) = \boldsymbol{A}\text{Mean}_W(\boldsymbol{E}_1, \dots, \boldsymbol{E}_N)\boldsymbol{A}^\top = \boldsymbol{A}\overline{\boldsymbol{E}}\boldsymbol{A}^\top$, where $\overline{\boldsymbol{E}}$ has the same block diagonal structure as the $\boldsymbol{E}_i$'s, and $\overline{\boldsymbol{E}}_{jj} = (\sum_i \sqrt{p_{ij}})^2$ for $j \leq Q$. $\overline{\boldsymbol{C}}$ is also decomposed as $\overline{\boldsymbol{C}} = \overline{\boldsymbol{Y}}\overline{\boldsymbol{Y}}^\top$ with $\overline{\boldsymbol{Y}} = \boldsymbol{A}\overline{\boldsymbol{E}}^{\frac{1}{2}}$.
Further, $\boldsymbol{Q}_i^* = \boldsymbol{V}_i\boldsymbol{U}_i^\top$ with $\boldsymbol{U}_i$ and $\boldsymbol{V}_i$ coming from the SVD of $\overline{\boldsymbol{Y}}^\top\boldsymbol{Y}_i = \overline{\boldsymbol{E}}^{\frac{1}{2}}\boldsymbol{E}_i^{\frac{1}{2}}$ which has the same structure as the $\boldsymbol{E}_i$'s. Therefore $\boldsymbol{Q}_i^*$ has also the same structure with the identity matrix as its upper block.
Finally we have $\boldsymbol{v}_i = \mathcal{P}_{\overline{\boldsymbol{C}}}(\boldsymbol{C}_i) = \text{vect}(\boldsymbol{Y}_i\boldsymbol{Q}_i^* - \overline{\boldsymbol{Y}})$ so it is linear in $\sqrt{(p_{i,j})}$ for $j \leq Q$.

The restriction to the case where A is orthogonal stems from the orthogonal invariance of the Wasserstein distance. In the neuroscience literature square root rectifications are however not commonly used for M/EEG modeling. Nevertheless, it is interesting to see that the Wasserstein metric that can naturally cope with rank reduced data is consistent with this particular generative model.

Proposition : Consistency of SPoC regression algorithm

The **SPoC regression algorithm** consists in using all the P SPoC spatial filters:

$$\boldsymbol{v}_i = f\left(\text{diag}\left(\boldsymbol{W}_{\text{SPoC}}\,\boldsymbol{C}_i\,\boldsymbol{W}_{\text{SPoC}}^{\top}\right)\right) \in \mathbb{R}^P \ , \tag{1.58}$$

with $\boldsymbol{W}_{\text{SPoC}}$ a matrix $W \in \mathbb{R}^{P\times P}$ solution of the generalized eigenvalue problem:

$$\boldsymbol{C}_y\boldsymbol{W} = \overline{\boldsymbol{C}}\boldsymbol{W}\,\text{diag}(\lambda_1,\ldots,\lambda_P) \text{ subject to } \boldsymbol{W}^{\top}\overline{\boldsymbol{C}}\boldsymbol{W} = I_P \ , \tag{1.59}$$

with $\overline{\boldsymbol{C}} = \frac{1}{N}\sum_{i=1}^{N}\boldsymbol{C}_i$ the Euclidean average covariance matrix and $\boldsymbol{C}_y = \frac{1}{N}\sum_{i=1}^{N} y_i\boldsymbol{C}_i$ the weighted average covariance matrix, and $\lambda_1,\ldots,\lambda_P$ the generalized eigenvalues. We assume that the eigenvalues are all distinct, and therefore without loss of generality $\lambda_1 > \cdots > \lambda_P$.

This model is statistically consistent for any function f: the relationship between y_i and $\boldsymbol{v}_i$ is linear. It achieves consistency by taking a rather different approach than previous models: it recovers in $\boldsymbol{W}_{\text{SPoC}}$ the inverse of the mixing matrix $\boldsymbol{A}$.

Proof

If the eigenvalues are all distinct, the generalized eigenvalue problem has a unique solution $\boldsymbol{W}$. We recall the definition $\boldsymbol{E}_i = \mathbb{E}[\boldsymbol{\eta}_i(t)\boldsymbol{\eta}_i(t)^\top]$, which is block-diagonal with the sources powers $p_{i,j}$ as coefficient (j,j) when $j \leq Q$. We have $\boldsymbol{C}_i = \boldsymbol{A}\,\boldsymbol{E}_i\,\boldsymbol{A}^\top$, and therefore $\overline{\boldsymbol{C}} = \boldsymbol{A}\,\overline{\boldsymbol{E}}\,\boldsymbol{A}^\top$ and $\boldsymbol{C}_y = \boldsymbol{A}\,\boldsymbol{E}_y\,\boldsymbol{A}^\top$, with $\overline{\boldsymbol{E}} = \frac{1}{N}\sum_{i=1}^N \boldsymbol{E}_i$ and $\boldsymbol{E}_y = \frac{1}{N}\sum_{i=1}^N y_i\boldsymbol{E}_i$ sharing the same block-diagonal structure than the $\boldsymbol{E}_i$. Their lower $(P-Q)\times(P-Q)$ diagonal blocks, respectively $\overline{\boldsymbol{\Sigma}}$ and $\boldsymbol{\Sigma}_y$, are symmetric matrices. Further, $\overline{\boldsymbol{\Sigma}}$ is definite positive, as a linear combination with positive coefficients of definite positive matrices. Hence, $\overline{\boldsymbol{\Sigma}}$ and $\boldsymbol{\Sigma}_y$ are co-diagonalizable *i.e.,* there exists an invertible matrix $\boldsymbol{Z}$ such that $\boldsymbol{\Sigma}_y = \boldsymbol{Z}\boldsymbol{D}_y\boldsymbol{Z}^\top$ and $\overline{\boldsymbol{\Sigma}} = \boldsymbol{Z}\overline{\boldsymbol{D}}\boldsymbol{Z}^\top$. By denoting $\boldsymbol{A}' = \boldsymbol{A}\times\begin{bmatrix}\boldsymbol{I}_Q & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{Z}\end{bmatrix}$, we have that $\overline{\boldsymbol{C}}$ and $\boldsymbol{C}_y$ are co-diagonalized by $\boldsymbol{A}'$. Let $\boldsymbol{D}$ the diagonal matrix such that $\overline{\boldsymbol{C}} = \boldsymbol{A}'\boldsymbol{D}\boldsymbol{A}'^\top$. The matrix $\boldsymbol{W} = \boldsymbol{A}'^{-\top}\boldsymbol{D}^{-1/2}$ is solution of the generalized eigenvalue problem. By the unicity assumption, SPoC recovers $\boldsymbol{W}$ up to a permutation of its columns. The first Q rows of $\boldsymbol{W}^\top$ are the first Q columns of $\boldsymbol{A}'^{-\top}$, hence the first Q rows of $\boldsymbol{A}^{-1}$, up to scale. In particular, the transform $\boldsymbol{W}^\top\boldsymbol{x}_i$ recovers the Q sources $\boldsymbol{s}_i$, so that $\boldsymbol{W}_{\text{SPoC}}\,\boldsymbol{C}_i\,\boldsymbol{W}_{\text{SPoC}}^\top$ recovers the $\boldsymbol{p}_i$. Finally, since y_i is linearly related to the components of $f(\boldsymbol{p}_i)$ that themselves are linearly related to the components of $\boldsymbol{v}_i$, it will also be linearly related to the components of the feature vector $\boldsymbol{v}_i$, hence the consistency of the SPoC regression algorithm.

As an historical side note, the SPoC algorithm is a supervised spatial filtering algorithm simultaneously discovered by [CP14] and [Däh+14a]. The main idea of the SPoC algorithm is to use the information contained in the outcome variable to guide the decomposition, giving priority to source signals whose band power correlates with y. Note that it was originally developed for event-level regression, *e.g.,* in BCI, and we adapt it here to a general problem that can also accommodate subject-level regression, where one observation corresponds to one subject instead of one trial. More formally, the filters $\boldsymbol{W}$ are chosen to synthesize signals whose powers maximally covariates with the outcome y. Which may be a good idea since we supposed that our target is linearly related to the power of the sources. Denoting by $\overline{\boldsymbol{C}} = \frac{1}{N}\sum_{i=1}^N \boldsymbol{C}_i$ the Euclidean average covariance matrix and $\boldsymbol{C}_y = \frac{1}{N}\sum_{i=1}^N y_i\boldsymbol{C}_i$ the weighted average covariance matrix, the first filter $\boldsymbol{w}_{\text{SPoC}}$ is given by: $\boldsymbol{w}_{\text{SPoC}} = \arg\max_{\boldsymbol{w}} \frac{\boldsymbol{w}^\top\boldsymbol{C}_y\boldsymbol{w}}{\boldsymbol{w}^\top\overline{\boldsymbol{C}}\boldsymbol{w}}$. In practice, all the filters in $\boldsymbol{W}_{\text{SPoC}}$ are obtained by solving the generalized eigenvalue decomposition problem [Däh+14a]. The proof is quite straightforward and given below. Note that here we use all the P spatial filters ($R = P$).

Proof. We assume that the signal $\boldsymbol{x}(t)$ is band-pass filtered in one frequency band of interest, so that for each subject the band power of signal is approximated by the variance over time of the signal. We denote the expectation $\mathbb{E}$ and the variance $\mathbb{V}\text{ar}$ over time t or subject i by a corresponding subscript.

The source extracted by a spatial filter $\boldsymbol{w}$ for subject i is $\widehat{\boldsymbol{s}}_i = \boldsymbol{w}^\top \boldsymbol{x}_i(t)$. Its power reads: $\Phi_i^{\boldsymbol{w}} = \mathbb{V}\text{ar}_t[\boldsymbol{w}^\top \boldsymbol{x}_i(t)] = \mathbb{E}_t[\boldsymbol{w}^\top \boldsymbol{x}_i(t)\boldsymbol{x}_i^\top(t)\boldsymbol{w}] = \boldsymbol{w}^\top \boldsymbol{C}_i \boldsymbol{w}$ and its expectation across subjects is given by: $\mathbb{E}_i[\Phi_i^{\boldsymbol{w}}] = \boldsymbol{w}^\top \mathbb{E}_i[\boldsymbol{C}_i]\boldsymbol{w} = \boldsymbol{w}^\top \overline{\boldsymbol{C}}\boldsymbol{w}$, where $\overline{\boldsymbol{C}} = \frac{1}{N}\sum_i \boldsymbol{C}_i$ is the average covariance matrix across subjects. Note that here, $\boldsymbol{C}_i$ refers to the covariance of the $\boldsymbol{x}_i$ and not its estimate as in Sec. 1.2.3.

We aim to maximize the covariance between the target y and the power of the sources, $\mathbb{C}\text{ov}_i[\Phi_i^{\boldsymbol{w}}, y_i]$. This quantity is affected by the scaling of its arguments. To address this, the target variable y is normalized: $\mathbb{E}_i[y_i] = 0 \quad \mathbb{V}\text{ar}_i[y_i] = 1$. Following [Däh+14a], to also scale $\Phi_i^{\boldsymbol{w}}$ we constrain its expectation to be 1: $\mathbb{E}_i[\Phi_i^{\boldsymbol{w}}] = \boldsymbol{w}^\top \overline{\boldsymbol{C}}\boldsymbol{w} = 1$ The quantity one aims to maximize reads:

$$\begin{aligned}
\mathbb{C}\text{ov}_i[\Phi_i^{\boldsymbol{w}}, y_i] &= \mathbb{E}_i[\ (\Phi_i^{\boldsymbol{w}} - \mathbb{E}_i[\Phi_i^{\boldsymbol{w}}])\ (y_i - \mathbb{E}_i[y_i])\ ] \\
&= \boldsymbol{w}^\top \mathbb{E}_i[\boldsymbol{C}_i y_i]\boldsymbol{w} - \boldsymbol{w}^\top \overline{\boldsymbol{C}}\boldsymbol{w}\mathbb{E}_i[y_i] \\
&= \boldsymbol{w}^\top \boldsymbol{C}_y \boldsymbol{w}
\end{aligned}$$

where $\boldsymbol{C}_y = \frac{1}{N}\sum_i y_i \boldsymbol{C}_i$.

Taking into account the normalization constraint we obtain: $\widehat{\boldsymbol{w}} = \arg\max_{\boldsymbol{w}^\top \overline{\boldsymbol{C}}\boldsymbol{w}=1} \boldsymbol{w}^\top \boldsymbol{C}_y \boldsymbol{w}$. Note that it can be also written as a generalized Rayleigh quotient:

$$\widehat{\boldsymbol{w}} = \arg\max_{\boldsymbol{w}} \frac{\boldsymbol{w}^\top \boldsymbol{C}_y \boldsymbol{w}}{\boldsymbol{w}^\top \overline{\boldsymbol{C}}\boldsymbol{w}} \ .$$

Its Lagrangian reads $F(\boldsymbol{w}, \lambda) = \boldsymbol{w}^\top \boldsymbol{C}_y \boldsymbol{w} + \lambda(1 - \boldsymbol{w}^\top \overline{\boldsymbol{C}}\boldsymbol{w})$. Setting its gradient *w.r.t.* $\boldsymbol{w}$ to 0 yields a generalized eigenvalue problem: $\nabla_{\boldsymbol{w}} F(\boldsymbol{w}, \lambda) = 0 \implies \Sigma_y \boldsymbol{w} = \lambda \overline{\Sigma_{\boldsymbol{x}}} \boldsymbol{w}$ This equation has a unique closed-form solution called the generalized eigenvectors of $(\boldsymbol{C}_y, \overline{\boldsymbol{C}})$. The second derivative gives:

$$\nabla_{\boldsymbol{\lambda}} F(\boldsymbol{w}, \lambda) = 0 \implies \lambda = \boldsymbol{w}^\top \Sigma_y \boldsymbol{w} = \mathbb{C}\text{ov}_i[\Phi_i^{\boldsymbol{w}}, y_i] \tag{1.60}$$

This equation leads to an interpretation of λ as the covariance between $\Phi^{\boldsymbol{w}}$ and y, which should be maximal. As a consequence, $\boldsymbol{W}_{\text{SPoC}}$ is built from the generalized eigenvectors of the generalized eigenvalue problem above, sorted by decreasing eigenvalues. □

Link between the regression algorithms It is noteworthy that both SPoC and Riemann models have in common to whiten the covariances $\boldsymbol{C}_i$ with a common reference covariance (the Euclidean mean for SPoC and the geometric mean for

Riemann): Riemann explicitly with $\overline{\boldsymbol{C}}^{-1/2}\boldsymbol{C}_i\overline{\boldsymbol{C}}^{-1/2}$, SPoC implicitly by solving generalized eigenvalue problem of $(\boldsymbol{C}_y, \overline{\boldsymbol{C}})$, or equivalently of $(\boldsymbol{C}_i, \overline{\boldsymbol{C}})$ which is equivalent to solving the regular eigenvalue problem of $\boldsymbol{C}_i$ after whitening with $\overline{\boldsymbol{C}}$ ([Fuk90; NNC11] eq. 13-16). SPoC retrieves the eigenvectors of $(\boldsymbol{C}_i, \overline{\boldsymbol{C}})$. Riemann produces vectors whose size depend on the log eigenvalues of $(\boldsymbol{C}_i, \overline{\boldsymbol{C}})$. They both produce non-linear features that measure powers relative to a common reference.

'Upper' and Riemann models both avoid inverting $\boldsymbol{A}$ by being insensitive to it. More precisely, they consist in building from $\boldsymbol{C}_i$ a $P \times P$ symmetric matrix $\boldsymbol{M}_i$ mathematically congruent to a block-diagonal matrix $\boldsymbol{D}_i$ whose $Q \times Q$ upper block is $\mathrm{diag}(f(\boldsymbol{p}_i))$ *i.e.,* that writes $\boldsymbol{M}_i = \boldsymbol{B}\,\boldsymbol{D}_i\,\boldsymbol{B}^\top$, for some invertible matrix $\boldsymbol{B}$. Indeed, if this holds, the coefficients of $f(\boldsymbol{p}_i)$ are a linear combination of coefficients of $\boldsymbol{M}_i$ which implies that the outcome y_i is linear in the coefficients of $\boldsymbol{M}_i$. Therefore a linear model applied to the features $\boldsymbol{v}_i = \mathrm{Upper}(\boldsymbol{M}_i)$ is statistically consistent. For 'Upper' $\boldsymbol{M}_i = \boldsymbol{C}_i$ and for Riemann $\boldsymbol{M}_i = \log(\overline{\boldsymbol{C}}^{-1/2}\boldsymbol{C}_i\overline{\boldsymbol{C}}^{-1/2})$.

Finally, these two models amounts to estimating Q parameters (the powers of each sources) from $P(P+1)/2$ parameters (the upper part of a symmetric matrix). Again, it is important to emphasize that we are not aiming for explicitly estimating the most probable model parameters $\hat{\beta}$ but rather a function that has the smallest approximation error possible, even if over-parametrized. These approaches achieve consistency without inverting $\boldsymbol{A}$ at a price of over-parametrization: the number of parameters will always be a lot higher than the number of samples N. Learning in this under determined high dimensional setting requires regularizing the linear model to stabilize learning. We will thus use a Ridge regression algorithm with linear kernel, but in a data-driven fashion with nested generalized cross-validation, leading to effective degrees of freedom less than numerical rank of the input data.

In this thesis we will compare these four regression algorithms to the inconsistent **diag regression algorithm** as baseline. This model is probably the historically most frequently used model in M/EEG research in countless publications. Here, powers are considered on the sensor array while the correlation structure is being ignored. This consists in taking only the diagonal elements of the covariance matrix $\boldsymbol{C}_i$ as features, *i.e.,* the powers (variances) of sensor-level signals:

$$\boldsymbol{v}_i = f(\mathrm{diag}(\boldsymbol{C}_i)) \in \mathbb{R}^P \quad . \tag{1.61}$$

1.5.2 Model violations

The current theoretical analysis implies that the mixing matrix $\boldsymbol{A}$ must be common to all subjects and the covariance matrices must be full rank. However, this is rarely the case in practice. If these conditions are not satisfied, the consistency guarantees are lost, rendering model performance an empirical question. This will be addressed

with simulations (Section 1.5.3), in which we know the true brain-behaviour link, and real-world data analysis (Section 2.1) in which multiple uncertainties arise at the same time.

Noise in the target variable. Most often the target variable is corrupted by a small additive random perturbation ε_i. The noise in the outcome variable depends on the context: it can represent intrinsic measurement uncertainty of y_i, for example sampling rate and latency jitter in behavioral recordings, inter-rater variability for a psychometric score, or simply a model mismatch. The true model may not be linear for example. It can also regroup all the other dependencies we consider non-discriminative.

Individual mixing matrix. A model where the mixing matrix $\boldsymbol{A}$ is subject-dependent reads: $\boldsymbol{x}_i(t) = \boldsymbol{A}_i \boldsymbol{s}_i(t) + \boldsymbol{n}_i(t)$. Such situations typically arise when performing subject-level regression due to individual head geometry, individual head positions in MEG and individual locations of EEG electrodes. In this setting, we loose consistency guarantees, but since the $\boldsymbol{A}_i$ cannot be completely different from each other (they all originate from human brains), we can still hope that our models perform reasonably well.

Rank-deficient signal. We have seen that M/EEG data is in practice often rank-reduced for mainly two reasons. First, popular techniques for cleaning the data amounts to reduce the noise by projecting the data in a subspace, supposed to predominantly contain the signal. Second, a limited amount of data may lead to poor estimation of covariance matrices. This leads to rank-deficient covariance matrices.

Riemann regression algorithms must be adapted since singular matrices are at infinite distance from any regular matrices. Assuming the rank $R < P$ is the same across subjects, the corresponding covariance matrices do not belong to the $\mathcal{S}_P^{++}$ manifold anymore but to the $\mathcal{S}_{P,R}^{+}$ manifold of positive semi-definite matrices of fixed rank R. To handle the rank-deficiency one can then project the covariance matrices on $\mathcal{S}_R^{++}$, the manifold of full-rank matrices of reduced size R to make them full rank, and then use the geometric distance. To do so, a common strategy is to project the data into a subspace (here of size R) that captures most of its variance. This is achieved by Principal Component Analysis (PCA) applied to the average covariance matrix across subjects. We denote the filters in this case by $\boldsymbol{W}_{\text{UNSUP}} = \boldsymbol{U} \in \mathbb{R}^{R \times P}$, where $\boldsymbol{U}$ contains the eigenvectors corresponding to the top R eigenvalues of the average covariance matrix $\overline{\boldsymbol{C}} = \frac{1}{N} \sum_{i=1}^{N} \boldsymbol{C}_i$. This step is blind to the values of y and is therefore unsupervised. Note that under the assumption that the time series across subjects are independent, the average covariance $\overline{C}$ is the covariance of the data over the full population. The Riemann regression algorithm is then applied to the

spatially-filtered covariance matrix $\boldsymbol{W}_{\text{UNSUP}}\,\boldsymbol{C}_i\,\boldsymbol{W}_{\text{UNSUP}}^{\top} \in \mathbb{R}^{R\times R}$, of rank R hence belonging to $\mathcal{S}_R^{++}$.

If covariances are rank-deficient, we will use low-rank versions of both SPoC and Riemann models where only the first components of the spatial filters up to their rank R are kept. In SPoC, components are ordered by covariance with the outcome (supervised algorithm). In Riemann, components are ordered by explained variance in the predictors, not the outcome (unsupervised algorithm). By construction, we can then expect that SPoC achieves similar best performance than Riemann with fewer components: the variance related to the outcome can be represented with fewer dimensions.

Besides helping to cope with rank-reduced data, the effect of spatial filtering can be difficult to predict: it helps the regression algorithm by reducing the dimensionality of the problem making it statistically easier, but it can also destroy information if the individual covariance matrices are not aligned (if they span different spaces). We expect to observe this trade-off via low-rank optima, with a plateau after the effective rank R of the data (see Section 2.1.2).

To summarize: we presented four regression algorithms, linear models applied to four particular vectorization of the covariance matrix $\boldsymbol{C}_i$: 'Upper' takes its upper-part (powers & cross-powers of sensors), Riemann takes its Riemannian embedding *w.r.t.* the geometric distance *i.e.,* its projection in a common tangent space taken at their geometric mean, Wasserstein takes its Riemannian embedding *w.r.t.* the Wasserstein distance, and finally SPoC takes the diagonal of the covariance matrix of spatially-filtered signals where the filters are generalized eigenvectors of $(\boldsymbol{C}_i, \overline{\boldsymbol{C}})$ with $\overline{\boldsymbol{C}}$ the mean of the $\boldsymbol{C}_i$'s. They will be compared to the 'diag' baseline model that takes its log-diagonal (log-powers of sensors).

We demonstrated mathematically that, if we have constant volume conduction across samples and if the signal is full rank, these regression algorithms are statistically consistent for some function f defined in our generative model. Each of them uses a different strategy to bypass the mixing effect of the matrix $\boldsymbol{A}$ and can then replace the need for source localization. When f = Identity (y_i is linear in the coefficient of source powers $\boldsymbol{p}_i$) Upper achieves consistency by taking the coefficients of $\boldsymbol{C}_i$ as input vector. When f = Log (y_i is log-linear in the coefficient of source powers $\boldsymbol{p}_i$), Riemann achieves consistency by being simply insensitive to $\boldsymbol{A}$, blind to linear projections. Finally, SPoC learns to invert the mixing matrix $\boldsymbol{A}$ by incidentally recovering a co-block-diagonalization basis of all the $\boldsymbol{C}_i$'s, hence recovering the statistical sources: it is therefore consistent for any function f. In those idealized conditions, we demonstrated that source localization can be replaced with either spatial filters or with Riemannian embeddings.

1.5.3 Validation with simulations

We now consider simulations to illustrate these mathematical guarantees and investigate theoretical performance as model violations are gradually introduced (noise in target, individual mixing matrices). We focused on the 'linear-in-powers' and the 'log-linear-in-powers' generative models (Eq. 1.50 with $f = \text{identity}$ and $f = \log$) and compared the performance of the proposed approaches by measuring their score as the average mean absolute error (MAE) obtained with 10-fold cross-validation. Independent identically distributed covariance matrices $\boldsymbol{C}_1, \ldots, \boldsymbol{C}_N \in S_P^{++}$ and variables $y_1, \ldots, y_N$ were generated following each generative model. The mixing matrix $\boldsymbol{A}$ was defined as $\exp(\mu \boldsymbol{B})$ with the random matrix $\boldsymbol{B} \in \mathbb{R}^{P \times P}$ and the scalar $\mu \in \mathbb{R}$ to control the distance of $\boldsymbol{A}$ from identity ($\mu = 0$ yields $\boldsymbol{A} = \boldsymbol{I}_P$, increased μ means more mixing). This is not a model violation but a way to validate that the affine invariance property of the geometric Riemmanian distance indeed make the Riemann model blind to $\boldsymbol{A}$. The outcome variable was linked to the source powers (*i.e.,* the variance): $y_i = \sum_j \alpha_j f(p_{ij})$ with $f(x) = x$ or $\log(x)$. We chose $P = 5$, $N = 100$ and $Q = 2$. In these idealized conditions (no noise in target, fixed $\boldsymbol{A}$ and full rank signals), our consistent regression algorithms should show perfect out-of-sample prediction (no generalization error). Then, to investigate more realistic scenarios, we corrupted the clean ground truth data in two ways. First we add Gaussian noise in the outcome variable: $y_i = \sum_j \alpha_j f(p_{ij}) + \varepsilon_i$ where $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ is a small additive random perturbation. Second, we consider individual mixing matrices deviating from a reference: $\boldsymbol{A}_i = \boldsymbol{A} + \boldsymbol{E}_i$, where entries of $\boldsymbol{E}_i$ sampled i.i.d. from $\mathcal{N}(0, \sigma^2)$. The reference $\boldsymbol{A}$ can then be thought of as representing the head of a mean-subject. With these more realistic assumptions our mathematical guarantees break but our simulations will reveal the robustness of our regression algorithms to model violations.

Fig. 1.3 **A** displays the results for the linear generative model ($f = \text{identity}$ in Eq. (1.50)). The left panel shows the scores of each method as the parameters μ controlling the distance from the mixing matrix $\boldsymbol{A}$ to the identity matrix $\boldsymbol{I}_P$ increases (more mixing). We see that the Riemannian method is not affected by μ (orange), which is consistent with the affine invariance property of the geometric distance. At the same time, it is not the correct method for this generative model as is revealed by its considerable prediction error greater than 0.5. Unsurprisingly, the 'diag' method (green) is highly sensitive to changes in $\boldsymbol{A}$ with errors proportional to the mixing strength. On the contrary, both 'upper' (blue) and SPoC (dark orange) methods lead to perfect out-of-sample prediction (MAE $= 0$) even as mixing strength increases. This demonstrates consistency of these methods for the linear generative model. They both transparently access the statistical sources, either by being blind to the mixing matrix $\boldsymbol{A}$ ('upper') or by explicitly inverting it (SPoC). Hence, they may enable regression on M/EEG brain signals without source localization. When we

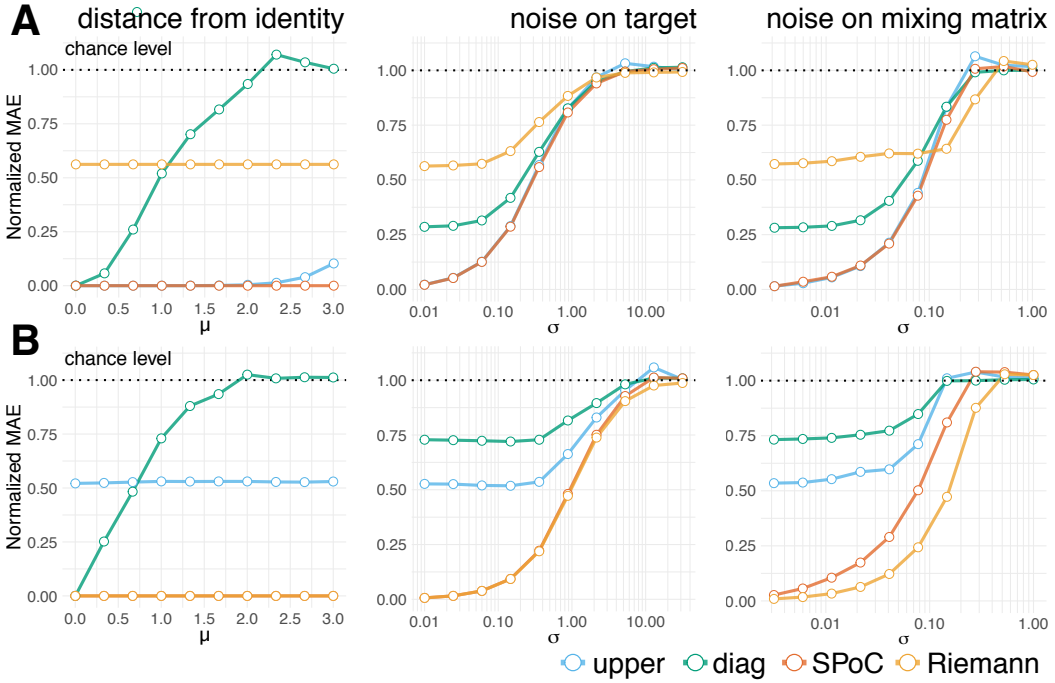


Fig. 1.3: Simulation-based model comparison across generative models We focused on four regression algorithms (indicated by color) each of which learnt from the covariance in distinct ways. The simulations performance across three types of model violations: the distance $\boldsymbol{\mu}$ between the mixing matrix $\boldsymbol{A}$ and the identity matrix $\boldsymbol{I}_P$ (left), noise on the outcome y (middle) and individual noise on $\boldsymbol{A}$ (right). (**A**) Results for generative model in which y depends linearly on source variance. All but the Riemannian model achieve consistent regression when no mixing occurs (left). SPoC remained consistent throughout the simulated range. The 'upper' and SPoC models performed best as noise on the outcome (center) and noise on $\boldsymbol{A}$ (right) increased. (**B**) Results for generative model in which y depends log-linearly on source powers. The SPoC and Riemannian models achieve consistent regression across all simulated values (left). Likewise, both methods are more robust to noise on y (center). Finally, the Riemannian model is most resilient to individual noise on the mixing matrix $\boldsymbol{A}$ (right).

add noise in the outcome variable y (middle), *i.e.,* introduce worse supervision, or individual noise in the mixing matrix (right), *i.e.,* individual volume conduction, we have no theoretical guarantee of optimality for those methods. Yet, we see that both 'upper' and SPoC are equally sensitive to these model violations. The Riemannian method seems to be more robust than any other method to individual noise in $\boldsymbol{A}$, in the sense that its performance is decaying at a slower rate.

Fig. 1.3 **B** displays the results for the log-linear generative model ($f = \log$ in Eq. (1.50)). In this case Riemann and SPoC performed best (left), as expected by consistency of these methods in this generative model. Both were equally sensitive to noise in outcome variable (middle) but, again, the Riemann method was more robust than other methods as individual noise on the mixing matrix increased (right). The simulations show that, under these idealized circumstances, 'upper', and SPoC are equivalent when the outcome y depends linearly on source powers. When y depends linearly on the log-powers, SPoC and Riemann are equivalent. However,

when every data point comes with a different mixing matrix, Riemann may be the best default choice, irrespective of the generative model of y. The Wasserstein regression algorithm has not been pictured here to avoid overloading but has a similar behaviour to the 'Upper' model with the same MAE profile, yet a marginally better performance. The same experiment with $f(p) = \sqrt{p}$ yields comparable results, yet with Wasserstein distance performing best and achieving perfect out-of-sample prediction when $\sigma \to 0$ and A is orthogonal.

To summarize we have proposed four different regression algorithms and showed, both theoretically and with simulated data, that they can each perfectly approximates the true function asymptotically hence supports perfect prediction with enough data, under a particular generative model. For this consistency to hold, we need 2 conditions: A must be fixed across samples and the signal hence the covariance matrix must be full rank. If they are satisfied, these models, as they stay in sensor-space, overcome volume conduction problem and avoid the need of costly source localization. These models can readily accommodate rank-deficient covariance matrices. Other models are working, for instance [Sch+17] uses deep learning method that does implicit filtering to handle the fieldspread. Our models have the advantages to be simple, to work with smaller quantities of data and above all propose an explicit explanation of why they work by making explicit the data generative model to which they are adapted. Let's now apply these regression algorithms to real-life empirical data: in ideal laboratory conditions first, then in clinical conditions.

Application with laboratory data

2

Contents

List of acronyms and notations of the chapter

BCI	brain-computer interface
BAD	Brain Age Delta
Cam-CAN	Cambridge Center of Aging
ERM	empirical risk minimization
EOG	electro-oculogram
ECG	electro-cardiogram
EMG	eletromyogram
M/EEG	magneto- and electroencephalography
ML	machine learning
MAE	mean absolute error
MNE	mnimium norm estimate
MRI	magnetic resonance imaging
OAS	Oracle Approximation Shrinkage
PCA	principal component analysis
SPoC	source power comodulation
SSS	Signal Space Separation
SSP	Signal Space Projection
TUH	Temple University Hospital

In Chapter 1, we detailed our first (theoretical) contribution [Sab+19a]: when faced with the particular task of predicting, from M/EEG signals, an outcome that is (log)-linear in sources powers, we have found three learning algorithms with no approximation error. They are all based on a linear function class, applied to particular vectorization on the spatial covariance matrix:

$$\mathcal{H} = \{\boldsymbol{X} \in \mathbb{R}^{P \times T} \mapsto \boldsymbol{w}^\top \boldsymbol{v}(\frac{\boldsymbol{X}\boldsymbol{X}^\top}{T}), \forall \boldsymbol{w}\} \ , \tag{2.1}$$

where $\boldsymbol{v}$ has been previously defined as per our Upper, Riemann and SPoC regression algorithms. We leveraged our domain prior knowledge of M/EEG signals to design these hypothesis classes and show that they are perfectly adapted to the data-generating mechanism, depending on how outcome could linearly depends on source powers: Upper for linear-in-powers outcome, SPoC or Riemann for linear-in-logpowers outcome. Yet, this guarantee of no approximation error holds only if two model assumptions are verified: the mixing matrix $\boldsymbol{A}$ is fixed across samples, and the signal (hence the covariance matrix) is full-rank. Besides, these algorithms show no optimization error. Indeed, since the hypothesis classes are linear in the parameter $\boldsymbol{w}$, they are all learnable with the ERM learning rule and no optimization algorithm is necessary to implement it since there is an analytic solution to the optimization problem. Additionally, if the outcome is noiseless (no irreducible error) and the data sample is infinite (no estimation error), then our algorithms have perfect generalization error.

In this Chapter 2, presenting our second (experimental) contribution [Sab+20], we now confront our theoretical results to real-life regression tasks of predicting a continuous neuro-outcome from real M/EEG data. In real-life experiments, we do not have access to the actual sources and do not know a priori which generative model, hence, which regression algorithm performs best. The true outcome-generating mechanism can be linear in sources powers, log-linear in sources powers or something completely different. Likewise, even if our algorithms are adapted to the data distribution, we cannot expect perfect out-of-sample prediction (no generalization error): the outcome variable may be noisy (leading to irreducible error), data samples are finite (leading to estimation error) and, most importantly, the two main modeling assumptions that guarantee consistency may not hold (leading to approximation error). Indeed, the mixing matrix $\boldsymbol{A}$ is not fixed when predicting at the subject-level as each individual has her own head and brain. Also the M/EEG data is often rank-reduced in practice, for mainly two reasons. First, popular techniques for cleaning the data amounts to reduce the noise by projecting the data in a subspace, supposed to predominantly contain the signal. Second, a limited amount of data may lead to poor estimation of covariance matrices. However, by performing model comparisons based on cross-validation errors, we can potentially infer which model provides the better approximation.

This chapter is organized as follows. To validate our algorithms on real M/EEG data, we first benchmarked them against a prediction task from M/EEG band-limited covariances (*i.e.,* computed separately within different frequency bands) on three datasets, for which multiple model violations occur at once. Then we will further inspect our regression models to gain physiological insights. Finally, in order to further clear the road towards their clinical translation, we will probe their robustness, both to low-fidelity devices and to signal preprocessing.

Section 2.1 details the motivation, methods and results obtained by our regression algorithms on three experiments, chosen to cover a wide range of model violations. In the first experiment, we focussed on predicting muscle contraction of a single subject from MEG beta activity on the FieldTrip data, for which we can consider the mixing matrices to be constant across samples. Compared to the ideal conditions necessary to obtain the mathematical guarantees of consistency, this experiment nevertheless presents one model violation: rank-deficient covariance matrices due to limited amount of data. In the other two experiments, we apply our methods to infer age from brain signals. Age is a dominant driver of cross-person variance in neuroscience data and a serious confounder [SN18]. As a consequence of the globally increased average lifespan, ageing has become a central topic in public health that has stimulated neuropsychiatric research at large scales. The link between age and brain function is therefore of utmost practical interest in neuroscientific research. The second experiment consists in predicting the age from MEG on Cam-CAN data, where we have the two model violations at once: variable mixing matrices and rank-reduced data, this time due to preprocessing. Finally, the third experiment also focus on age prediction but using EEG signals from the TUH dataset, where the covariances are full rank but the mixing matrices are individual.

Section 2.2 will dvelve into model inspection beyond performance. First, some of our algorithms use spatial filters, hence support inspection of the corresponding spatial patterns. This will allow to check that the patterns they learnt are physiologically plausible but more importantly that they are informative on the brain regions potentially involved in the task. Second, we will perform a sensitivity analysis of the algorithms to assess the individual relative influence of the data generating factors of head geometry, uniform global power and topographic information. We will show that all methods learn from anatomy but the Riemannian embeddings better capture individual head geometry, suggesting a complementary use of M/EEG to MRI. One important strength of our algorithms is that they avoid source reconstruction, facilitating their translation to the clinic.

In Section 2.3 we further investigate the robustness of our algorithms in order to assess their potential usage in the clinic, where only low-density EEG devices are really practical and a light preprocessing pipeline is conceivable. We will see that, again, the Riemannian model is particularly robust to preprocessing options and

performs well even when no preprocessing is done. This will clear the way to translating our algorithms to the clinic, which will be the focus of the next chapter.

Statistical modeling. Note that since our problem is high-dimensional (the number of dimensions will always be much higher than the number of available samples) we will have to stabilize the ERM learning rule using regularization, hence using a RLM learning rule (for instance ridge or Lasso). Here, we used ridge regression [HK70] to predict from the vectorized covariance matrices and tuned its regularization parameter by generalized cross-validation [GHW79] on a logarithmic grid of 100 values in $[10^{-5}, 10^{3}]$ on each training fold of a 10-fold cross-validation loop. For each model described in previous sections ('diag', Upper, SPoC, Riemann), we standardized the features enforcing zero mean and unit variance. This preprocessing step is standard for penalized linear models. To compare models against chance, we estimated the chance-level empirically through a dummy-regressor predicting the mean outcome of the training data. Uncertainty estimation was obtained from the cross-validation distribution. Note that formal hypotheses testing for model comparison was not available for any of the datasets analyzed as this would have required several datasets, such that each average cross-validation score would have made one observation. To improve conditioning of the covariance estimates, across all analyses, additional low-rank shrinkage for spatial filtering with SPoC and unsupervised spatial filtering with Riemann was fixed at the mean of the value ranges tested in [Sab+19a] *i.e.,* 0.5 and 10^{-5}, respectively.

Software. All numerical analyses were performed in Python 3.7 using the Scikit-Learn software [Ped+11], the MNE software for processing M/EEG data [Gra+14], the PyRiemann package [CBA13] for manipulating Riemannian objects, and the open-source Python library 'Coffeine' (`https://github.com/coffeine-labs/coffeine`) that provides a high-level interface to the predictive modeling techniques we develop in this chapter. We used the R-programming language and its ecosystem for visualizing the results [R C19; AUT19; Wic16; CSM17].

2.1 Empirical validation with real M/EEG data

In Chapter 1 we have seen that Riemannian approaches has turned out to be promising for enhancing classification of event-level data [Bar+12] for BCI classification. Yet, systematic comparisons against additional baselines and competing regression algorithms on larger datasets and other outcomes are missing. Importantly, the majority of approaches have focused on event-level prediction problems instead of subject-level prediction and have never been been systematically compared in terms of their statistical properties and empirical behavior. Here we will explicitly mainly focus on subject-level as contrasted to event-level prediction, both, theoretically and at the level of data analysis. Note that this thesis does not focus on event-level prediction with generalization across subjects [HP18; Wes+18; OKA14], which is a distinct and more complex problem inheriting its structure from, both, event-level and subject-level regression.

2.1.1 Predicting muscle contraction from MEG on FieldTrip data

In a first step, we considered a problem where the unit of observation was individual behavior of one single subject with some unknown amount of noise affecting the measurement of the outcome. The problem is an event-level regression task of predicting continuous electromyogram (EMG) from brain beta activity captured concomitantly with MEG. In this scenario, the mixing matrix is fixed to the extent that the subject avoided head movements, which was enforced by the experimental design. At the time of the analysis, individual anatomical data was not available, hence we constrained the analysis to the sensor-space.

Data acquisition. We analyzed one anonymous subject from the data presented in [Sch+11] and provided by the FieldTrip website to study cortico-muscular coherence [Oos+11]. The MEG recording was acquired with 151 axial gradiometers and the Omega 2000 CTF whole-head system. EMG of forceful contraction of the wrist muscles (bilateral musculus extensor carpi radialis longus) was concomitantly recorded with two silver chloride electrodes. MEG and EMG data was acquired at 1200Hz sampling-rate and online-filtered at 300Hz. For additional details please consider the original study [Sch+11].

Data processing and feature engineering. The analysis closely followed the continuous outcome decoding example from the MNE-Python website [Gra+14]. We considered 200 seconds of joint MEG-EMG data. First, we filtered the EMG above $20\,\mathrm{Hz}$ using a time-domain firwin filter design, a Hamming window with 0.02 passband ripple, $53\,\mathrm{dB}$ stop band attenuation and transition bandwidth of $5\,\mathrm{Hz}$ (-$6\,\mathrm{dB}$ at $17.5\,\mathrm{Hz}$) with a filter-length of $661\,\mathrm{ms}$. Then we filtered the MEG between 15 and $30\,\mathrm{Hz}$ using an identical filter design, however with $3.75\,\mathrm{Hz}$ transition bandwidth

for the high-pass filter (-6 dB at 13.1 Hz) and 7.5 Hz for the low-pass filter (-6 dB at 33.75 Hz). The filter-length was about 880 ms. Note that the transition bandwidth and filter-length was adaptively chosen by the default procedure implemented in the filter function of MNE-Python. We did not apply any artifact rejection as the raw data was of high quality. The analysis then ignored the actual trial structure of the experiment and instead considered a sliding window-approach with 1.5 s windows spaced by 250 ms. Allowing for overlap between windows allowed to increase sample size. We then computed the covariance matrix in each time window and applied Oracle Approximation Shrinkage (OAS) [Che+10] to improve conditioning of the covariance estimate. The outcome was defined as the variance of the EMG in each window.

Model evaluation. For event-level regression with overlapping windows, we applied 10-fold cross-validation without shuffling such that folds correspond to blocks of neighboring time windows preventing data-leakage between training and testing splits. The initialization of the random number generator used for cross-validation was fixed, ensuring identical train-test splits across models. Note that a Monte Carlo approach with a large number of splits would lead to significant leakage, hence, optimistic bias [Var+17]. This, unfortunately, limits the resolution of the uncertainty estimates and precludes formalized inference. As we did not have any a priori interest in the units of the outcome, we used the R^2 metric, a.k.a. coefficient of determination, for evaluation. Compared to the ideal conditions necessary to obtain the mathematical guarantees of consistency, this experiment presents one model violation: rank-deficient covariance matrices (along with noise in outcome).

Results: models performance. The results are depicted in Fig. 2.1. The analysis revealed that only models including the cross-terms of the covariance predicted visibly better than chance (Fig. 2.1**A**). For the methods with projection step (SPoC and Riemann) we reported the performance using the full 151 components, equal to the total number of gradiometer channels. Importantly, extensive search for model order for SPoC and Riemann revealed important low-rank optima (Fig. 2.1**B**) with performance around 50% variance explained on unseen data. This is not surprising when considering the difficulty of accurate covariance estimation from limited data. Indeed, low-rank projection is one important method in regularized estimation of covariance [EG15]. Interestingly, SPoC showed stronger performance with fewer components than Riemann (4 vs 42). This is not surprising: SPoC is a supervised algorithm, constructed such that its first components concentrate most of the covariance between their power and the outcome variable. The variance related to y can hence be represented with fewer dimensions than Riemann that uses unsupervised spatial filtering. However, it remains equivocal which statistical model best matches this regression problem. The best performing models all implied the log-linear model. Yet, compared to the linear-in-power Upper model, the low-

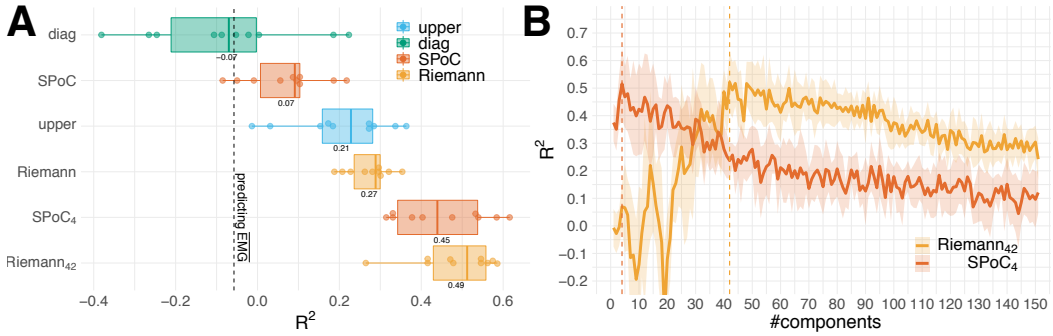


Fig. 2.1: Predicting continuous muscular activity on single-subject MEG. (A) model comparison using cross-validation with 10 consecutive groups of approximately overlapping 80 epochs from one single subject. Models are depicted along the y-axis, expected out-of-sample performance (R^2) on the x-axis. The distribution is summarized by standard boxplots. Split-wise prediction scores are represented by dots. The model type is indicated by color. SPoC and Riemann (without subscript) includes spatial filtering with full 151 components, equal to the total number of gradiometer channels. As covariance estimation is necessarily inaccurate with the short 1.5 second epochs, models may perform better when fit on a reduced subspace of the covariance. For these models we reported alternative low-rank models (model order indicated by subscripts). **(B)** Exhaustive search for model order in pipelines with projection step. All values from 1 to the total number of 151 gradiometer channels were considered. One can spot well defined low-rank regimes in both models. However, SPoC supports a lower model order than Riemann. Only models explicitly considering the between-sensor correlation were successful. The best performance was achieved when projecting into a lower dimensional space with optima for SPoC and Riemann of order 4 and 42, respectively.

rank SPoC and Riemann models also implied massive shrinkage on the covariances, leaving unclear if the type of model or regularized covariance estimation explains their superior performance.

2.1.2 Predicting age from MEG on Cam-CAN data

In a second MEG data example, we considered a subject-level regression problem in which we focused on age prediction from brain signals using the currently largest publicly available MEG dataset provided by the Cam-CAN (Cambridge Center of Aging) [Tay+17; Sha+14]. In this problem, each sample consists of resting-state MEG signals recorded from different persons, hence different brains. On theoretical grounds, one may therefore expect individual cortical folding, size and proportional composition of the head and its tissues to induce important distortions to the signal that may pose severe problems to purely data-driven approaches. Here, each data point can be said to have its own mixing matrix inducing unique distortions in each observation. To investigate this point explicitly, we further conducted source localization to obtain power estimates that corrected for individual head geometry based on biophysical prior knowledge. On the other hand, 8 minutes of MEG support accurate covariance estimation, hence, rendering model-order search less important for shrinkage. Covariance matrices are nevertheless rank-deficient due to the preprocessing steps, explained below. Thus, this problem imposes two important model violations of varying source geometry due to individual anatomy and rank-deficient covariances, while providing a clean outcome - the age - with virtually no

measurement noise. Other sources of noise can nevertheless still be present in the outcome.

Data acquisition. We considered task-free MEG recordings during which participants were asked to sit still with eyes closed in the absence of systematic stimulation. The recording lasted about eight minutes, sampled at $1000\,\mathrm{Hz}$. We then drew $T \simeq 520,000$ time samples from $N = 643$ subjects, between 18 and 89 years of age. MEG was acquired using a 306 VectorView system (Elekta Neuromag, Helsinki). This system is equipped with 102 magnetometers and 204 orthogonal planar gradiometers inside a light magnetically shielded room. During acquisition, an online filter was applied between around $0.03\,\mathrm{Hz}$ and $1000\,\mathrm{Hz}$. To support offline artifact correction, vertical and horizontal electrooculogram (VEOG, HEOG) as well as electrocardiogram (ECG) signal was concomitantly recorded. Four Head-Position Indicator (HPI) coils were used to track head motion. For subsequent source-localization the head shape was digitized. For additional details on MEG acquisition, please consider the reference publications on the Cam-CAN dataset [Tay+17; Sha+14].

Data processing and feature engineering. This large dataset required more extensive data processing. We composed the preprocessing pipeline following current good practice recommendations [Gro+13; Jas+18; Per+18]. The full procedure comprised the following steps: suppression of environmental artifacts, suppression of physiological artifacts (EOG/ECG) and rejection of remaining contaminated data segments. Each of them are detailed below. First, to mitigate contamination by high-amplitude environmental magnetic fields, we applied the signal space separation method (SSS) [TK05], as detailed in 1.2.4. SSS requires a comprehensive sampling (more than about 150 channels) and a relatively high calibration accuracy that is machine/site-specific. For this purpose we used the fine-calibration coefficients and the cross-talk correction information provided in the Can-CAM repository for the 306-channels Neuromag system used in this study. We used the temporal SSS (tSSS) extension [TK05], where both temporal and spatial projection are applied to the MEG data. For the spatial part, SSS decomposes the MEG signal into extracranial and intracranial sources and renders the data rank-deficient. We kept the default settings of eight and three components for harmonic decomposition of internal and external sources, respectively ($l = L_{in} = 8$ for the $\boldsymbol{S}_{in}$ basis, and up to $l = L_{out} = 3$ for the $\boldsymbol{S}_{out}$ basis). The origin of internal and external multipolar moment space was estimated based on the head-digitization hence specified in the 'head' coordinate frame and the median head position during 10s sliding windows is used. Once applied, magnetometers and gradiometers are projected back from a common lower dimensional SSS subspace, hence become linear combinations of approximately $R_i = 65$ common SSS components in our experiments. As a result, both sensor types contain highly similar information (which also modifies the inter-channel correlation structure), hence become interchangeable [Gar+17]. For

name	low	δ	θ	α	β_{low}	β_{high}	γ_{low}	γ_{mid}	γ_{high}
range (Hz)	$0.1-1.5$	$1.5-4$	$4-8$	$8-15$	$15-26$	$26-35$	$35-50$	$50-74$	$76-120$

Tab. 2.1: Definition of frequency bands

simplicity, we therefore conducted all analyses on signals from magnetometer sensors ($P = 102$), using a scale factor of 100 to bring the magnetometers to approximately the same order of magnitude as the gradiometers, as they have different units (T vs T/m). For the temporal part, we used 10-second sliding windows. To discard segments in which inner and outer signal components were poorly separated, we applied a correlation-threshold of 98%, in concert with basis regularization. Since no continuous head monitoring data were available at the time of our study, we performed no movement compensation. Second, to mitigate physiological ocular and cardiac artifacts, we applied the signal space projection method (SSP) [UI97]. This method learns principal components on data-segments contaminated by artifacts and then projects the signal into the subspace orthogonal to the artifact. To reliably estimate the signal space dominated by the cardiac and ocular artifacts, we excluded data segments dominated by high-amplitude signals using the 'global' option from autoreject [Jas+17]. To preserve the signal as much as possible, we only considered the first SSP vector based on the first principal component. As a final preprocessing step, we epoch the resulting data in 30s non overlapping windows and identify bad data segments (i.e. trials containing transient jumps in isolated channels) that have a peak-to-peak amplitude exceeding a certain global threshold, learnt automatically from the data using the autoreject (global) algorithm [Jas+17].

Concerning feature engineering, we considered a wide range of frequencies, as the most important source of variance is not a priori known for the problem of age prediction. To capture age-related changes in cortical brain rhythms [BJF10; Voy+15; Cla+04], we bandpass filtered the data into nine conventional frequency bands (*cf.* Tab. 3.2) adapted from the Human-Connectome Project [LP+13], and computed the band-limited covariance matrices with the OAS estimator [Che+10], hence focusing on the power spectral topography and between-sensor covariance as features. We verify that the covariance matrices all lie on a small portion of the manifold, justifying projection in a common tangent space. Then we ran the covariance pipelines independently in each frequency band and concatenated the ensuing features after the vectorization step.

Model evaluation. We used ridge regression and tuned its regularization parameter by generalized cross-validation [GHW79] on a logarithmic grid of 100 values in $[10^{-5}, 10^{3}]$ on each training fold of a Monte Carlo (shuffle split) cross-validation loop with 100 splits and 10% testing data. The initialization of the random number generator used for cross-validation was fixed, ensuring identical train-test splits

across models. This choice also allowed us to obtain more fine-grained uncertainty estimates than was possible with the time-series data used for subject-level regression. As absolute changes of the unit of the outcome is meaningful, we used the mean absolute error (MAE) as evaluation metric. The proposed method, including all data preprocessing, applied on the 500GB of raw MEG data from the Cam-CAN dataset, runs in approximately 12 hours on a regular desktop computer with at least 16GB of RAM. The preprocessing for the computation of the covariances is embarrassingly parallel and can therefore be significantly accelerated by using multiple CPUs. The actual predictive modeling can be performed in less than a minute on standard laptop.

Results: models performance in sensor space. Fig. 2.2 displays the results for our different regression models. The analysis revealed that all models performed clearly

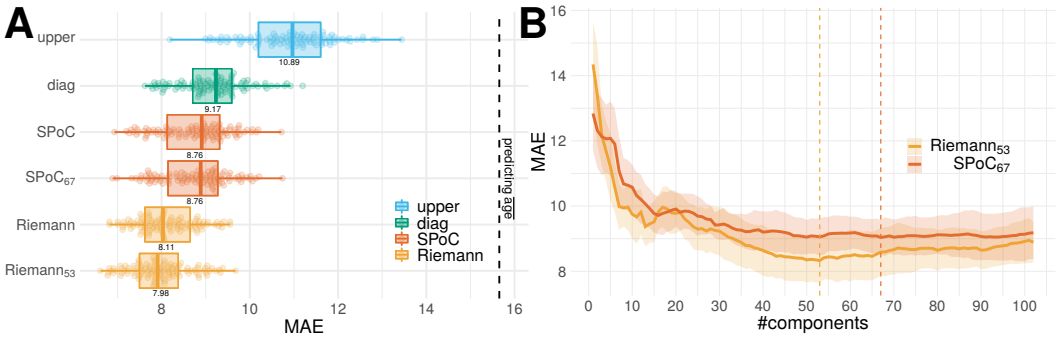


Fig. 2.2: Predicting age from subject-level MEG in sensor space. **(A)** model comparison using Monte Carlo cross-validation with 100 splits sampled from 596 subjects. Models are depicted along the y-axis, expected out-of-sample performance (mean absolute error) on the x-axis. The distribution is summarized by standard boxplots. Split-wise prediction scores are depicted by dots. The model type is indicated by color. Here, covariance estimation was based on 8 minutes of MEG, hence, the impact of shrinkage should be small. For comparison with the single-subject data (Fig. 2.1), we nevertheless reported the alternative low-rank models (model order indicated by subscripts, no subscript meaning an order of 65, the minimum rank of covariances). **(B)** Exhaustive search for model order in pipelines with projection step. All values from 1 to the total number of 102 magnetometer channels were considered. One can see that performance starts to saturate around 40 to 50. No striking advantage of model-order search was evident compared to deriving the order from prior knowledge on rank deficiency at a value of about 65. We see no low-rank minima, best performance is obtained with the full signals. All models performed better than chance, however, models consistent with log-linear model and using correlation terms performed better. On real data, when multiple model violations occur, the Riemannian models performed best.

better than chance. The Riemannian model (orange) yielded the best performance (8y MAE), followed by SPoC (dark orange, 8.8y MAE) (2.2**A**). The diagonal (green) and upper-triangle (blue) models performed worse. The chance level was $\sim$ 16y MAE. Model-order search did not reveal striking low-rank optima. Models above rank 40 seem approximately equivalent, especially when considering the estimation uncertainty of standard deviation above 1 year of MAE. For both SPoC and Riemann, the best low-rank model was close to the model at the theoretically derived rank of 65 (due to preprocessing with SSS, see [TK05]). For subsequent analyses, we, nevertheless, retained the best models.

One first important observation suggests that the log-linear model is more appropriate in this regression problem, as the only model not implying a log transform, the Upper model, performed clearly worse than any other model. Yet, important difference in performance remain to be explained among the log-linear models.

This points at the cross-terms of the covariance, which turns out to be an essential factor for prediction success: The 'diag' model ignores the cross-terms and performed worst among all log-linear models. The SPoC and Riemann models performed better than 'diag' and both analyzed the cross-terms, SPoC implicitly through the spatial filters. This raises the question why the cross-terms were so important. One explanation would be that they reveal physiological information regarding the outcome. Alternatively, the cross-terms may expose the variability due to individual head geometry. To further investigate this point we conducted the same regression analysis on source localized M/EEG signals, *i.e.,* after having corrected for distortions induced by individual head geometry with a biophysical model.

Results: models performance in source space. To compare the data-driven statistical models against a biophysics-informed method, for this dataset, we included a regression pipeline based on anatomically constrained minimum norm estimates (MNE) informed by the individual anatomy. The MNE approach has been detailed in Section 1.4.2. Following common practice using the MNE software, we used $Q = 8196$ candidate dipoles positioned on the cortical surface, and set the regularization parameter to $1/9$ [Gra+14]. Concretely, we used the MNE inverse operator as any other spatial filter by multiplying the covariance with it from both sides to obtain source-space covariance matrices. We then retained the diagonal elements which provides estimates of the source power. To obtain spatial smoothing and reduce dimensionality, we averaged the MNE solution using a cortical parcellation encompassing 448 regions of interest from [Kha+18]. For preprocessing of structural MRI data we used the FreeSurfer software ([Fis12], `http://surfer.nmr.mgh.harvard.edu/`). Results are depicted in Fig. 2.3.

Now, the optimal number of components for prediction remarkably dropped: 11 for Riemann and 20 for SPoC in source space, as compared to 53 and 67, respectively, in sensor space. This may suggest that the inflated number of components in sensor space is related to extra directions in variance accounting for individual head geometry. Second, 'diag' (green) is now by far the best regression model with performance at ~ 7.7y MAE. This model only takes the log powers into account and discards the cross-terms. This suggests that the outcome does not depend on the cross-terms or at least that the potential gain of the cross-terms is inaccessible due to the inflated dimensionality of feature space. The 'diag' score is also the highest among all the models that we considered so far, illustrating that the MNE solution to the inverse problem provides superior unmixing of brain signals.

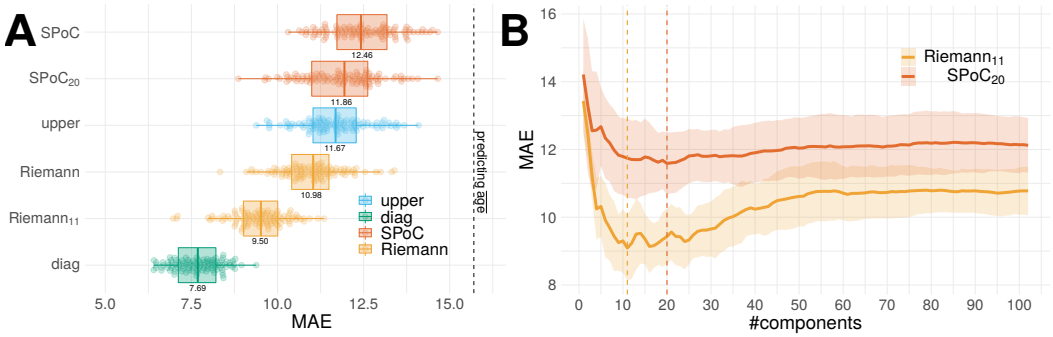


Fig. 2.3: Predicting age from subject-level MEG in source space. (A) model comparison applied to sources and using Monte Carlo cross-validation with 100 splits sampled from 596 subjects. It follows the same layout conventions than Fig. 2.2. The sources are estimated by MNE that exploits biophysical prior knowledge. **(B)** Exhaustive search for model order in pipelines with projection step. All values from 1 to the total number of 102 magnetometer channels were considered. One can see that performance starts to saturate around 40 to 50. But contrary to sensor space analysis of Fig. 2.2 projection models show clear low-rank minima. All models performed better than chance, however, the 'diag' model that only considers sources' log-powers clearly outperforms other models.

2.1.3 Predicting age from EEG on TUH data

The results on subject-level regression based on MEG suggest the importance of model violations due to individual head geometry. Importantly, with traditional cryogenic MEG, the sensor array is not fixed relative to the head, rendering head-positioning and head-movements factors contributing to model violations due to individual signal geometry. How would the present results generalize to clinical EEG where sensors are fixed relative to the head but, in general, fewer sensors are used? To investigate this question, we applied our subject-level age regression setting of Cam-CAN to clinical EEG: We analyzed resting-state EEG (21 sensors) from about 1000 subjects of the Temple University Hospital (TUH) EEG dataset [Har+14], one of the largest publicly available database of clinical EEG recordings. This ongoing project currently includes over 30,000 EEGs spanning the years from 2002 to present. As with previous analysis of the Cam-CAN data, each data point had its own mixing matrix. As with the Cam-CAN, the EEG recordings from TUH were sufficiently long to support accurate covariance estimation, hence, rendering model-order search less important for shrinkage. We did not preprocess the data on purpose to ensure having full-rank signals. This experiment is therefore appropriate to primarily investigate the particular model violation of sample-dependent mixing matrices with constrained degrees of freedom for the sensor-positioning as well as the generalization from MEG to EEG. Unfortunately the absence of associated MRI data prevented us to conduct source localization to correct for individual head geometry.

Data acquisition. We used the TUH "Abnormal EEG Corpus", a subset of TUH EEG Corpus that have been annotated as normal or abnormal by medical experts. From this dataset we focussed on the 1385 healthy patients, from both training and evaluation sets, whose EEG has been annotated as normal. Their age ranges

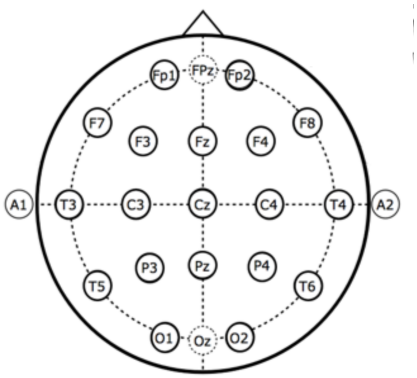


Fig. 2.4: Position of the 21 EEG electrodes selected for our experiment using TUH EEG Corpus dataset.

between 10 and 95 years (mean 44.3y, std 16.5y, 775 females, 610 males). EEG was acquired using several generations of Nicolet EEG system (Natus Medical Inc.), equipped between 24 and 36 channels. All sessions have been recorded with an average reference electrode configuration, sampled at 250Hz minimum. The minimal recording length for each session was about 15 minutes. For additional details on EEG acquisition, please consider the reference publications on the TUH dataset [Har+14].

Data processing, feature engineering and model evaluation. We applied minimal preprocessing to the raw EEG data. We first selected the subset of 21 electrodes common to all subjects (A1, A2, C3, C4, CZ, F3, F4, F7, F8, FP1, FP2, FZ, O1, O2, P3, P4, PZ, T3, T4, T5, T6), see Figure 2.4. We then discarded the first 60 seconds of every recording to avoid artifacts occurring during the setup of the experiment. For each patient we then extracted the first eight minutes of signal from the first session, to be comparable with Cam-CAN. EEG recordings were downsampled to 250Hz. Finally, we excluded data segments dominated by high-amplitude signals using the 'global' option from autoreject [Jas+17] that computes adaptive rejection thresholds. Note that the absence of linear projection to preprocess raw data (as SSS or SSP in Cam-CAN) ensures the data are full rank. While the rank was reduced by one by the use of a common average reference, as we used a subset of channels common to all subjects, the data are actually full rank. Otherwise, we followed the same feature engineering and modeling pipeline used for the Cam-CAN data (See Section 2.1.2).

Results: models performance. Fig. 2.5 displays the results for different regression models. Model-order search did not reveal any clear low-rank optima. This was expected considering the absence of preprocessing and accurate covariance estimation. Strikingly, the only model not implementing a log transform, the Upper model, performed at chance level, clearly worse than any other model. All other models performed better than chance, with Riemann clearly leading, followed by SPoC and diag. Those results are consistent with our simulations in Fig. 1.3**(B)** in which the only model violation comes from individual mixing matrices. The performance and

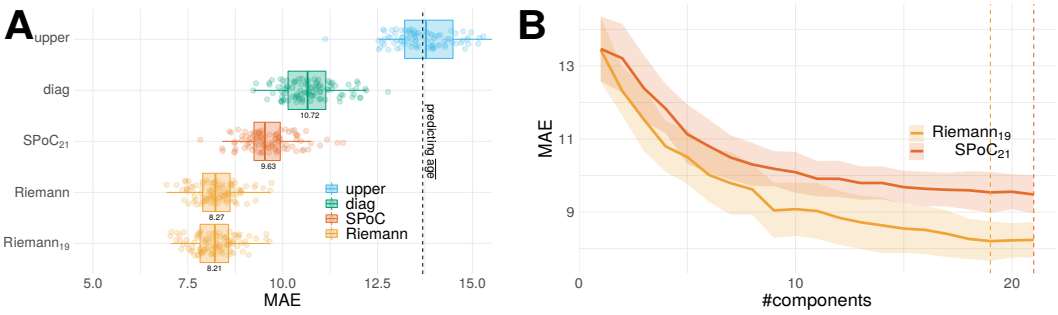


Fig. 2.5: Predicting age from subject-level EEG in sensor space. (A) model comparison applied to sensors and using Monte Carlo cross-validation with 100 splits sampled from 1000 subjects. It follows the same layout conventions than Fig. 2.2. Here, covariance are full rank, the impact of shrinkage should be small. We nevertheless reported the alternative low-rank models (model order indicated by subscripts). **(B)** Exhaustive search for model order in pipelines with projection step. All values from 1 to the total number of 21 electrodes were considered. Model-order search did not reveal striking low-rank optima. All models except 'upper' performed better than chance, however, models consistent with log-linear model and using correlation terms performed better. The Riemannian models performed best.

ordering of the models in the TUH data is also consistent with the results obtained on the Cam-CAN dataset. This strongly suggests that the log-linear model is more appropriate in this regression problem. It is noteworthy, that the best performance based on the Riemannian model was virtually identical to its performance with MEG on the Cam-CAN data. However, it remains open to which extent the benefit of constrained signal geometry due to fixed sensor positioning is cancelled out by reduced spatial sampling with 21 instead of 306 sensors.

As a conclusion, we considered three experiments presenting a variety of complementary model violations. Across all these experiments, our Riemannian algorithm is a clear winner, leading the data-driven methods in sensor-space, showing a strong performance and robustness to model violations.

2.2 Model inspection

Beyond assessing pure performance, it is important to inspect our regression algorithms to check that they yield physiologically plausible explanations of performance. We will focus our model inspection analysis on the Cam-CAN dataset, as it allows to assess every models presented so far including the source-localized MNE regression model. In our first analysis we will leverage the fact that both the SPoC and MNE regression models use spatial filters, respectively informed by the outcome or by the individual anatomy. Hence, they readily support inspection of the corresponding spatial patterns, which is not the case for the Riemannian model [1]. In our second analysis we will perform a sensitivity analysis of 'diag', SPoC and Riemann models to assess individual relative influence of head geometry, uniform global power and topographic information in performance.

2.2.1 Spatial patterns

Fig. 2.6 depicts the marginal patterns [Hau+14] from the SPoC supervised filters and the MNE source-level filters (the rows of the linear operator $\boldsymbol{W}_{\text{MNE}}$ derived in Section 1.4.2), respectively. The sensor-level results suggest predictive dipolar patterns in the theta to beta range roughly compatible with generators in visual, auditory and motor cortices. Note that differences in head-position can make the sources appear deeper than they are (distance between the red positive and the blue negative poles). Similarly, the MNE-based model suggests localized predictive differences between frequency bands highlighting auditory, visual and premotor cortices. While the MNE model supports more exhaustive inspection, the supervised patterns are still physiologically informative. For example, one can notice that the pattern is more anterior in the β-band than the α-band, potentially revealing sources in the motor cortex.

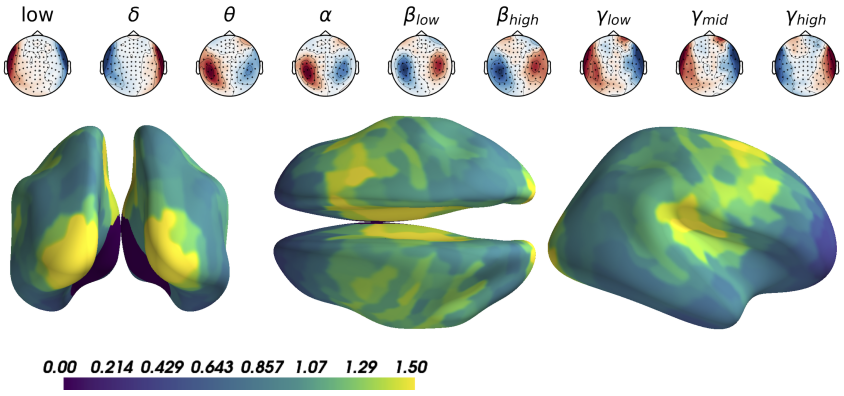


Fig. 2.6: Model inspection. Upper panel: sensor-level patterns from supervised projection. One can notice dipolar configurations varying across frequencies. Lower panel: standard deviation of patterns over frequencies from MNE projection highlighting bilateral visual, auditory and premotor cortices.

[1] Recent methods [XGWJ20; Kob+21], published after this work, have then been proposed to perform such introspections for Riemannian models, notably for tangent space linear models

2.2.2 MEG and EEG as a cheap MRI?

The Cam-CAN study revealed that the performance of Riemann in sensor space (Figure 2.2) is close to 'diag' in source space (Figure 2.3), suggesting that the cross-term models, in sensor space, have learnt to some extent what 'diag', in source space, receives explicitly from source localization. Still, the good performance of 'diag' in source space may be due to two independent factors that are not mutually exclusive: It could be that source localization standardizes head geometry, hence, mitigates the variability of mixing. On the other hand, if the anatomy itself covaries with the outcome, which is a safe assumption to make for the case of aging [Lie+17], the leadfields will also covary with the outcome. Source amplitudes may then change as a result of dampening-effects (See methods in Khan et al. [Kha+18]).

To disentangle the factors explaining model performance, and understand how Riemannian model partially handle individual volume conduction, we devised a novel error-decomposition method derived from the proposed statistical framework (Fig. 1.2). The link between the data-generating mechanism and the proposed regression models allows us to derive an informal analysis of variance [Gel+05] for estimating the importance of the data generating factors such as head geometry, uniform global power and topographic, i.e., spatial information. Given the known physics from Eq. (1.44), the data covariance can be written $\boldsymbol{C}_i = \boldsymbol{G}_i \boldsymbol{C}_i^z \boldsymbol{G}_i^\top$, where $\boldsymbol{C}_i^z$ is the covariance matrix of the physiological sources in a given frequency band. The input to the regression model is therefore affected by both the head geometry expressed in $\boldsymbol{G}_i$, and the covariance of the sources. Using a simulation-based approach, we can therefore compute *degraded* observations, *i.e.,* versions of the full individual covariance $\boldsymbol{C}_i^D$ that were either exclusively influenced by the individual anatomy in terms of the leadfields, or also by additive uniform power. Subsequent model comparisons against the full models then allow isolating the relative merit of each of these specific components. Following common practice, we considered electrical dipolar sources $\boldsymbol{z}_i(t) \in \mathbb{R}^M$, with $M \approx 8000$, and we computed the leadfield matrix $\boldsymbol{G}_i$ with a boundary element model (BEM) [Gra+14]. We then defined two alternative models which are only based on the anatomical information or, additionally, on the global signal power in a given frequency band without topographic structure. This simulation will therefore allow us to estimate to which extent the log-linear models have learnt from anatomical information, global signal power of the MEG and topographic details.

Model using anatomy only. Assuming the physiological sources are Gaussian, uncorrelated and of unit variance (power) $\boldsymbol{z}_i^D(t) \sim \mathcal{N}(0, \boldsymbol{I}_M)$, we can re-synthesize their covariance matrix from individual leadfields alone without taking into account the actual covariance structure:

$$\boldsymbol{C}_i^D = \boldsymbol{G}_i \boldsymbol{G}_i^\top \quad . \tag{2.2}$$

Model using anatomy and spatially uniform power. Assuming the physiological sources are Gaussian, uncorrelated and of uniform power $\boldsymbol{z}_i^D(t) \sim \mathcal{N}(0, \sigma_i^2 \boldsymbol{I}_M)$, where σ_i is a scaling factor, we can re-synthesize their covariance matrix from individual leadfields and subject-specific source power, again, ignoring the actual covariance structure:

$$\boldsymbol{C}_i^D = \sigma_i^2 \boldsymbol{G}_i \boldsymbol{G}_i^\top \quad . \tag{2.3}$$

Specifically, we chose $\sigma_i^2 = \mathrm{Tr}(\boldsymbol{C}_i)/\mathrm{Tr}(\boldsymbol{G}_i\boldsymbol{G}_i^\top)$, such that $\mathrm{Tr}(\boldsymbol{C}_i^D) = \mathrm{Tr}(\boldsymbol{C}_i)$: the sum of powers of the signals is the same. This corresponds to taking into account the total power of the sources in a given frequency band and anatomy in the ensuing regression model. Note that we omitted frequency-specific notation for simplicity.

Results: error decomposition To perform our sensitivity analysis we repeated the Cam-CAN analysis with spatio-spectral information progressively removed. Fig. 2.7 compares three log-linear models based on the original observations (black) and the degraded covariances (orange): the 'diag' model and the best low-rank models previously found for SPoC and Riemann methods.

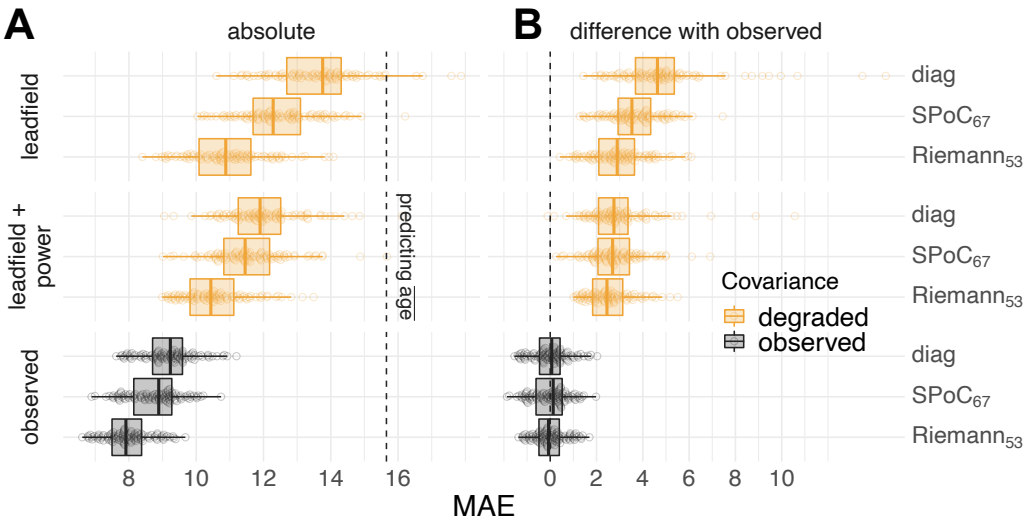


Fig. 2.7: Simulation-based error decomposition. We performed model comparisons for the observed data (black) and degraded data (orange) for which spatio-spectral information was progressively removed: 'leadfield + power' muted topographic information keeping only spatially uniform power and information from the individual leadfields (Eq. 2.3), 'leadfield' muted all electrophysiological variance (Eq. 2.2). (**A**) depicts absolute performance, (**B**), differences with the full observation, correspondingly, for each model. One can see that all models learnt to predict age from all three components: anatomical variation across subjects, electrophysiological signal power and topographic information. However, the relative importance of each error component was clearly different across models. Riemannian model was most responsive to the leadfield component (that explains ∼5y of performance) and least responsive to the uniform power (∼0.5y or performance). A large portion of the prediction performance (∼2.5y) was also explained by fine-grained spatial patterns.

One can see that all three error components improved overall prediction in similar ways, each improving performance between 2 and 4 years on average (Fig. 2.7**A**). The best performance with the leadfields-only was obtained by the Riemannian

model scoring an MAE of about 11y on average. Adding spatially uniform power, the Riemann model kept leading and improved by about 0.5y. Predictions based on the observed data with full access to the covariance structure improved performance by up to about 3y, suggesting that age prediction clearly benefits from information beyond the leadfields.

Generally, the choice of algorithm mattered across all levels of the data generating scenario with Riemann always leading and the 'diag' model always trailing (Fig. 2.7**A**). Finally, the results suggest the presence of an interaction effect where both the leadfields and the uniform power components were not equally important across models (Fig. 2.7**A,B**). For the Riemannian model, when only learning from leadfields, performance got as close as three years to the final performance of the full model (Fig. 2.7**B**). The 'diag' model, instead, only arrived at 5 years of distance from the equivalent model with full observations (Fig. 2.7**B**). On the other hand, the Riemannian model extracted rather little additional information from the uniform power and only made its next leap forward when accessing the full non-degraded covariance structure. Please note that these analyses are based on cross-validation. The resulting resampling splits do not count as independent samples. This precludes formal analysis of variance with an ANOVA model.

Overall, error decomposition suggests that all methods learn from anatomy and that indeed, the leadfield in isolation is predictive of age. Models considering cross-terms of the covariance were however more sensitive. It turns out that the leadfield contained some information on aging and that Riemannian embeddings were most sensitive to this information. Riemannian model was most responsive to the leadfield component and least responsive to the uniform power. This information was not explained by head positioning, pointing at differences in brain anatomy. It's conceivable that the Riemannian embedding better exposed this anatomical information, facilitating deconfounding for the ridge model and/or contributing unique information. The fact that Riemannian embeddings seem to capture individual head geometry justifies the use of EEG in the clinic beyond availability: brain age EEG and brain age MRI must be correlated because half the variance in covariances is anatomically explained. But also the fact that anatomical variations does not fully explain the performance means that neuronal activity captured by M/EEG contributes to the prediction: something is unique about M/EEG. This is an additional hint, beyond our article [Eng+20], and using a different method, that M/EEG is complementary with MRI.

2.3 Model robustness

Our first contribution led to the development of a regression model from M/EEG signals, with mathematical guarantees of optimality under certain conditions, and most importantly that avoids source reconstruction, facilitating its translation to the clinic. Yet, two important roadblocks remain before considering a clinical usage of our algorithms: they seem to require 1) a high-density 306 channels MEG device to acquire the signals and 2) an heavy preprocessing pipeline to clean the signals from environmental & physiological artefacts, both conditions inadequate to clinical practice.

Robustness to low-fidelity EEG. We already investigated the first issue with our clinical EEG experiment in Section 2.1.3 whereby we applied our Riemannian model to the analysis of ~ 1000 low-fidelity 21-channels clinical EEGs from the Temple University dataset and found remarkably similar performance levels with 306-channel high-density lab MEG, again with the Riemannian embeddings leading to the best performance. Our regression models seem therefore robust to signal low-fidelity and a clinical-grade device. The second issue is that our methods seem to require a heavy preprocessing pipeline: does it work with noisy signals?

Robustness to signal preprocessing. Commonly used preprocessing in M/EEG analysis is based on the idea to enhance signal-to-noise ratio by removing signals of non-interest, often using dedicated signal-space decomposition techniques [UI97; TK05; HKO04]. For instance, in our Can-CAM age regression problem, we used the preprocessing pipeline detailed in paragraph *Data processing and feature engineering* of section Section 2.1.2, consisting in environmental denoising (via signal space separation SSS), physiological ECG/EOG artifacts removal (via signal space projection SSP), and rejection of bad segments (via automatic peak-to-peak amplitude thresholding). However, it is perfectly imaginable that such preprocessing removes information useful for predicting. At the same time, predictive models may learn the signal subspace implicitly, which could render preprocessing unnecessary. To investigate this issue in Can-CAM, we sequentially repeated the analysis after activating the essential preprocessing steps one by one, and compared them to the baseline of extracting the features from the raw data with no preprocessing at all. This allows to compare regression models across different combinations of preprocessing steps. For this purpose, we considered an alternative preprocessing pipeline in which we kept all steps unchanged but the SSS [TK05] for removal of environmental artifacts. We used instead a data-driven PCA-based SSP [UI97] computed on empty room recordings. Results are depicted in Fig. 2.8.

The analysis revealed that the Riemannian model performed reasonably well when no preprocessing was done at all (Fig. 2.8**A**), almost as good as the other algorithms

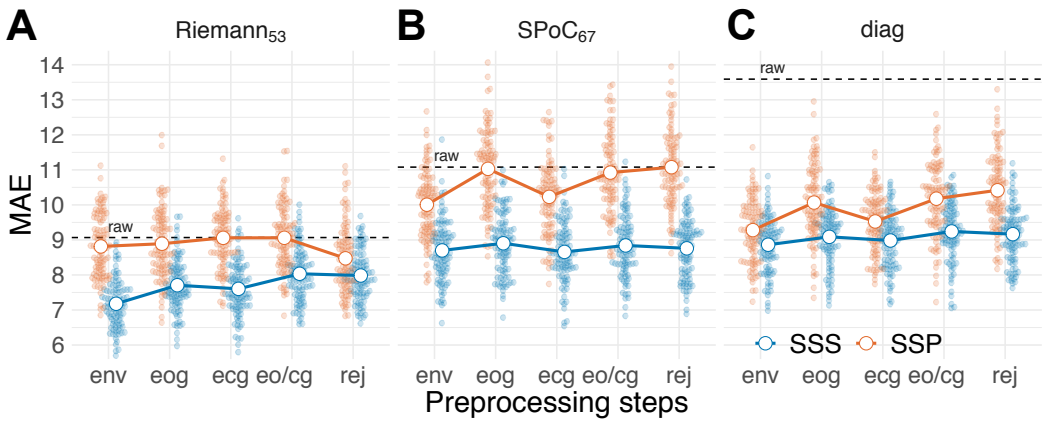


Fig. 2.8: Impact of preprocessing. Model comparison across cumulative artifact removal steps: environmental artifacts (env), environmental + occular (eog), environmental + cardiac (ecg), environmental + occular + cardiac (eo/cg), environmental + occular + cardiac + bad segments (rej). Results are compared to the baseline of extracting features from raw data with no preprocessing (depicted by vertical dashed lines). The method for removal of environmental artifacts is indicated by color, *i.e.,* , blue and red for SSS and SSP respectively. Note that the endpoint `rej` is identical to the full preprocessing conducted in previous analyses. Panels depict performance for the best Riemannian model **(A)**, the best SPoC model **(B)**, and the 'diag' model **(C)**. One can see that the Riemann model, but not the 'diag' model, is relatively robust to preprocessing and its details.

with full preprocessing. It also turned out to be relatively robust to particular preprocessing choices. On the other hand, whether preprocessing was done or not turned out decisive for the 'diag' model and to some extent for the SPoC model (Fig. 2.8**B,C**). A few common tendencies became apparent. Across all models, while improving above baseline, SSP as a first step consistently led to worse performance than SSS. Second, performance was also slightly degraded by removing ocular and cardiac artifacts, suggesting that both shared variance with age. Removing EOG seemed to consistently degrade performance. On the other hand, removing ECG had virtually no impact for SPoC and the 'diag' model. SSP, whether used for environmental or physiological artefacts, has virtually no impact on Riemann. This was expected, as SSP involves orthogonal projections and Riemann has a built-in ability to discard noise orthogonal to signal. Yet, for Riemann, both removing ECG and EOG after SSS additively deteriorated performance which suggests these artefacts share variance with the age outcome. Finally bad epochs rejection had a negligible and inconsistent effect. Overall, the results suggest that the importance of preprocessing depended on the model, while minimal denoising with SSP or SSS always helped improve performance. Of note, with minimal preprocessing using SSS, the Riemannian model performed at least as well as the 'diag' model after source localization (Fig. 2.3), here showing the best performance observed so far around 7y MAE. This contribution, the analysis of the impact to preprocessing on Can-CAM, reveals that Riemannian model seem to be a good bet across a wide range of settings with considerable robustness to extreme noise.

Finally, to summarize our two contributions so far, we found that when predicting from M/EEG power spectra is the priority, the capacity of linear models can be extended optimally by Riemannian embeddings despite model violations. Strikingly, applied to age prediction, this Riemannian regression algorithm has the potential to be used in the clinic: it operates in sensor-space (avoiding costly source localization), it is robust to environmental and physiological artefacts and it accommodates cheap EEG recordings. This optimal, robust and light model is therefore a good candidate to develop our clinical Brain Age Delta (BAD) biomarker.

2.4 Discussion

What distinguishes event-level from subject-level prediction in the light of model violations?

Unsurprisingly, no model performed perfectly when applied to empirical data for which the data generating mechanism is by definition unobservable, multiple model violations may occur and information is only partially available. One important source of differences in model violation is related to whether outcomes are defined at the event-level or at the subject-level. When predicting outcomes from ongoing segments of neural time-series within a subject, covariance estimation becomes non-trivial as the event-level time windows are too short for accurate estimation. Even if regularized covariance estimates provide an effective remedy, there is not one shrinkage recipe that works in every situation [EG15]. In this study, we have relied on the oracle approximating shrinkage (OAS) [Che+10] as a default method in all analyses. Yet, we found that additional low-rank shrinkage [EG15; Woo+11; TB99; RCJ18], as implied by the SPoC method [Däh+14a], or the unsupervised projection for the Riemannian model [Sab+19a], improved performance considerably for event-level prediction. A spatial-filter method like SPoC [CP14; Däh+14a] can be particularly convenient in this context. By design, it concentrates the variance most important for prediction on a few dimensions, which can be easily searched for, ascending from the bottom of the rank spectrum. Riemannian methods can also be operated in low-rank settings [Sab+19a]. However, model-order search may be more complicated as the best model may be anywhere in the spectrum. This can lead to increased computation times, which may be prohibitive in realtime settings such as BCI [Lot+07; Lot+18; Tan+08].

Issues with the numerical rank of the covariance matrix also appear when predicting at the subject-level. The reason for this is fundamentally different and rather unrelated to the quality of covariance estimation. Many modern M/EEG preprocessing techniques focus on estimating and projecting out the noise-subspace, which leads to rank-deficient data. In our analysis of the Cam-CAN dataset [Sha+14; Tay+17], we applied the SSS method [TK05] by default, which is the recommended way when no strong magnetic shielding is available, as is the case for the Cambridge MEG-system on which the data was acquired (see also discussion in [Jas+18]). However, SSS massively reduces the rank down to about 64 out of 306 dimensions, which may demand special attention when calibrating covariance estimation. Our results suggest that projection can indeed lead to slightly improved average prediction once a certain rank value is reached. Yet, thoughtful search of optimal model order may not be worth the effort in practice when a reasonably good guess of model order can be derived from the understanding of the preprocessing steps applied. Our findings, moreover, suggest, that a Riemann-based model is, in general, a reasonably

good starting point, even when no model-order search is applied. What seems to be a much more important issue in subject-level prediction from M/EEG are the model violations incurred by individual anatomy. Our mathematical analysis and simulations demonstrated that not even the Riemannian approach is immune to those, for MEG and EEG.

What explains the performance in subject-level prediction?

Our results suggested that, for the current regression problems with MEG and EEG, the log-linear model was more appropriate than the linear-in-powers ones. This is well in line with practical experience and theoretical results highlighting the importance of log-normal brain dynamics [BM14]. On the other hand, on the Cam-CAN data, we observed substantive differences in performance within the log-normal models highlighting a non-trivial link between the cross-terms of the covariance and subject-level variation. Indeed, the 'diag' model, both in sensor and source space, ignored the cross-terms of the covariance, yet in source space, it performed about 1.5 years better on average than in sensor space. This is rather unsurprising when recapitulating the fact that subject-level regression on M/EEG implies individual anatomy. Indeed, our mathematical analysis and simulations identified this factor as important model violation. MNE source localization, by design, uses the head and brain geometry to correct for such violations. On the other hand, if leadfields are correlated with the outcome, the source localization, which depends on the leadfields, will be predictive of the outcome too, even if no brain source is actually relevant to the outcome. This suggests that the cross-term models that were more successful than the 'diag' model may either convey biological information relevant to predict the outcome, or expose forward information on head geometry to the regression model, which then improved prediction by de-confounding for head geometry. Our findings on source localization strongly suggested that correcting for geometrical misalignment was the driving factor, evidenced by the fact that after source localization the simple 'diag' model performed best. Yet, these findings did not rule out that leadfields themselves were not predictive of the outcome.

We, therefore, derived a novel error-decomposition technique from the statistical framework presented in Fig. 1.2 to estimate the sensitivity of our M/EEG regression models to anatomy, spatially uniform power and topographic details. We applied this technique on the Cam-CAN dataset to investigate the subject-level prediction problem. While all models captured anatomical information and the Riemannian models were the most sensitive to it, anatomical information did not explain the performance based on the full data. At the same time, this demonstrated that MEG captures age-related anatomical information from the individual leadfields and raises the question of which aspects of anatomy were concerned. Neuroscience of aging has suggested important alterations of the cortical tissues [Lie+17], relevant for generating M/EEG signals, such as cortical surface area, cortical thickness or

cortical folding. Yet, more trivially, head size or posture are a common issue in MEG and could explain the present effect, which would be potentially less fascinating from a neuroscientific standpoint. We investigated this issue post-hoc by predicting age from the device-to-head transform describing the position of the head relative to the helmet and the coregistration transforms from head to MRI. Compared to the Riemannian model applied to the leadfields-only surrogate data, this resulted in three years lower performance of around 14 years error, which is close to the random guessing error and may at best explain the performance of the 'diag' model. Moreover, translating our approach to EEG for which sensor placement relative to the head is less variable, we did not witness improvements over MEG. On the other hand, this may be due to the smaller number of sensors available in EEG. Future work will have to show, how these two factors interact in practice across prediction problems and EEG-configurations.

Interestingly, also the SPoC model was more sensitive to anatomy than the 'diag' model. This suggests that by learning adaptive spatial filters from the data to best predict age, SPoC may implicitly also tune the model to the anatomical information conveyed by the leadfields. This seems even more plausible when considering that from a statistical standpoint, SPoC learns how to invert the mixing matrix $\boldsymbol{A}$ to get the statistical sources implied by the predictive model. This must necessarily yield a linear combination of the columns of $\boldsymbol{G}$. As a consequence, SPoC does not learn to invert the leadfields $\boldsymbol{G}$ but directly yields an imperfect approximation to $\boldsymbol{G}$. Theoretically, unique SPoC solution can be found with arbitrary outcomes as long as the data is full-rank and the target is noise-free. In practice, this is rarely the case. Therefore, the SPoC solution empirically depends on the choice of the outcome. This also motivates the conjecture that differences between SPoC and Riemann should become smaller when the $\boldsymbol{G}_i$ are not correlated with the outcome (Riemann should still enjoy an advantage due to increased robustness to model violations) or even vanish when $\boldsymbol{G}$ is constant and no low-rank issues apply. The latter case is what we encountered in the event-level analysis where SPoC and Riemann where roughly on par, suggesting that both handled the distortions induced by $\boldsymbol{G}$.

Unfortunately, the current analysis did not elucidate the precise mechanism by which different models learnt from the individual anatomy and why the Riemannian model was so much more proficient. As a speculation, one can imagine that changes in the leadfields translate into simple topographic displacements that the 'diag' model can easily capture. This would be in line with the performance of the 'diag' model on the leadfields-only surrogate data, which matched prediction performance based on the device-to-head transforms or the coregistration matrices previously mentioned. With cross-terms included in the modeling, SPoC and, in particular, Riemann may better unravel the directions of variation with regard to the outcome by considering the entire geometry presented in the leadfields. Instead, for the case of the leadfields-

only surrogates, SPoC attempts capturing sources which literally do not exist, hence must yield a degraded view on $\boldsymbol{G}$.

Overall, our results suggest that Riemannian models may also be the right choice when the anatomy is correlated with the outcome and the primary goal is prediction. The enhanced sensitivity of the Riemannian model to source and head geometry may be precisely what brings them so close to performance based on source localization. Indeed, the TUH experiment shows that these properties render Riemannian models particularly helpful in the case of EEG, where the leadfields should be less variable as the sensor cap is affixed to the head, which strongly limits variation due to head posture.

How important is preprocessing for subject-level prediction?

It is up to now equivocal how important preprocessing is when performing predictive modeling at the subject-level. Some evidence suggests that preprocessing may be negligible when performing event-level decoding of evoked responses as a linear model may well learn to regress out the noise-subspace [Hau+14]. Our findings suggest a more complex situation when performing subject-level regression from M/EEG signal power. Strikingly, performing no preprocessing was clearly reducing performance, for some models even dramatically, SPoC and in particular 'diag'. The Riemann model, on the other hand, was remarkably robust and performed even reasonably well without preprocessing. Among the preprocessing steps, the removal of environmental artifacts seemed to be most important and most of the time led to massive improvements in performance. Removing EOG and ECG artifacts mostly reduced performance suggesting that age-related information was present in EOG and ECG. For example, one can easily imagine that older subjects produced less blinks or showed different eye-movement patterns [Tha+15] and also cardiac activity may change across the lifespan [Att+19].

Interestingly, our results suggest that the method used for preprocessing was highly important. In general, performance was clearly enhanced when SSS was used instead of SSP. Does this mean that SSP is a bad choice for removing environmental artifacts? Our results have to be interpreted carefully, as the situation is more complicated when considering how fundamentally different SSP and SSS are in terms of design. When performing SSS, one actually combines the information of independent gradiometer and magnetometer sensor arrays into one latent space of roughly 65 dimensions, less than the dimensionality of both sensor arrays (306 sensors in total). Even when analyzing the magnetometers only after SSS, one will also access the extra information from the gradiometers [Gar+17]. SSP on the other hand is less invasive and is applied separately to magnetometers and gradiometers. It commonly removes only few dimensions from the data, yielding a subspace greater than 280 in practice. Our results therefore conflate two effects: 1)

learning from magnetometers and gradiometers versus learning from magnetometers only and 2) differences in strength of dimensionality reduction. To disentangle these factors, careful experimentation with more targeted comparisons is indicated. To be conclusive, such an effort may necessitate computations at the scale of weeks and should be investigated in a dedicated study. For what concerns the current results, the findings simply suggest that SSS is a convenient tool as it allows one to combine information from magnetometers and gradiometers into a subspace that is sufficiently compact to enable efficient parameter estimation. It is not clear though, if careful processing with SSP and learning on both sensors types would not lead to better results.

Conclusion

Our study has investigated learning continuous outcomes from M/EEG signal power from the perspective of generative models. Across datasets and electrophysiological modalities, the log-linear model turned out to be more appropriate. In the light of common empirical model violations and preprocessing options, models based on Riemannian geometry stood out in terms of performance and robustness. The overall performance level is remarkable when considering the simplicity of the model. Our results demonstrate that a Riemannian model can actually be used to perform end-to-end learning [Sch+17] involving nothing but signal filtering and covariance estimation and, importantly, without deep-learning [Roy+19]. When using SSS, performance improves beyond the current benchmark set by the MNE model but probably not because of denoising but rather due to the addition of gradiometer information. Moreover, we observed comparable performance on minimally processed clinical-EEG with only 21 channels instead of 306 MEG-channels, suggesting that the current approach may well generalize to certain clinical settings. This has at least two important practical implications. First, this allows researchers and clinicians to quickly assess the limits of what they can hope to learn in an economical and eco-friendly fashion [SGM19]. In this scenario, the Riemannian end-to-end model rapidly delivers an estimate of the overall performance that could be reached by extensive and long processing, hence, support practical decision making on whether a deeper analysis is worth the investment of time and resources. Second, this result suggests that if prediction is the priority, availability of MRI and precious MEG expertise for conducting source localization is not any longer the bottleneck. This could potentially facilitate data collection and shift the strategy towards betting on the law of large numbers: assembling an MEG dataset in the order of thousands is easier when collecting MRI is not a prerequisite.

It is worthwhile to consider important limitations of this study. Unfortunately, we have not had access to more datasets with other interesting continuous outcomes. In particular the conclusions drawn from the comparison between event-level and subject-level regression may be expanded in the future when considering larger

event-level datasets and other outcomes for which the linear-in-powers model may be more appropriate. Second, one has to critically acknowledge that the performance benefit for the Riemannian model may be partially explained by increased sensitivity to anatomical information, which might imply reduced specificity with regard to neuronal activity. In this context it is noteworthy that recent regression pipelines based on a variant of SPoC [Däh+14b] made use of additional spatial filtering for dimensionality reduction, *i.e.,* , SSD [NNC11] to isolate oscillatory components and discard arrhythmic (1/f) activity. This raises the question if the specificity of a Riemannian model could be enhanced in a similar way. Ultimately, what model to prefer, therefore, clearly depends on the strategic goal of the analysis [Bzd+18; BI19] and cannot be globally decided.

We hope that this study will provide the community with the theoretical framework and tools needed to deepen the study of regression on neural power spectra and safely navigate between regression models and geometric distortions governing M/EEG observations.

3 Application with clinical data: general anaesthesia

Contents

List of acronyms and notations of the chapter

AI	artificial intelligence
APHP	assistance publique - hôpitaux de paris
ASA	american society of anesthesiology
BA	brain age
BIDS	brain imaging data structure
BS	burst suppression
BSI	bispectral index
Cam-CAN	Cambridge center of aging
CV	cross validation
EDF	european data format
EEG	electroencephalography
GA	general anaesthesia
GABA	gamma-aminobutyric acid
MAE	mean absolute error
M/EEG	magneto- and electroencephalography
MEG	magnetoencephalography
MOCA	Montreal cognitive assessment
ML	machine learning
MRI	magnetic resonance imaging
OAS	oracle approximating shrinkage
PSD	power spectral density
PSI	patient state index
SEF95	spectral edge frequency (under which 95% of power is contained)
STD	standard deviation
TCI	target controlled infusion
TUH	Temple University hospital

3.1 Intraoperative brain age: from population modeling to anaesthesia

Neurovascular and neurodegenerative diseases are among the top causes of worldwide mortality hence a major public health concern [Eura]. Their impact can be mitigated by early diagnosis using predictive measures of neurodegenerative risk. Some of these measures assessing individual brain health have been developed in research laboratory from large population using machine learning, *e.g.,* recently brain-predicted age to predict brain aging from M/EEG data. Yet, they require biomedical exams that are indicated only when suffering is present, making the diagnosis often too late. Moreover, population studies suffer from selection bias since mostly healthy people participate in lab studies, leading to demographic stratification of predictive accuracy.

One rarely investigated solution to these problems consists in using monitoring data from General Anaesthesia (GA). Indeed, contrary to research studies, GA is driven by medical indication and performed at massive scales. This procedure concerns people from anywhere in society leading to millions of recordings, including EEG as monitoring EEG during GA is a general recommendation by learned societies to monitor depth of anaesthesia. Also, early evidence suggests that EEG during GA can reveal neurodegenerative risk factors [Fri+20]. By revealing pathologies of brain function, GA-based modeling could revolutionize preventive medicine if the monitoring data from millions of annual operations were scientifically actionable.

Brain age (BA) has been shown to be a better indicator of cognitive disorders than age [Col+19]. Yet, BA, EEG and GA have rarely been investigated together. BA has been mostly investigated in literature through MRI not EEG [Col+18; Fra+12; Jón+19; Col+17]. We also know that deriving biomarkers from EEG can be done optimally under certain conditions [Sab+19a], and has recently been applied to BA [Sab+20], yet never in a GA context. GA is a pertinent moment to extract signals from patients with minimal artefacts due to muscle-inhibitor drugs, hence a particularly adapted moment to build biomarkers from EEG, yet it has never been used to estimate BA.

Multiple challenges stand in the road to building brain age models from GA data due to specificities of the anaesthesia period. First, EEG during GA is done for monitoring purposes, not research. Doctors read out multiple cues in the signal, *e.g.,* flat EEG periods to detect burst suppression, spectrograms to judge signal quality by the presence of beta power or BIS/PSI indices to monitor anesthetic depth [Sch+03]. These quantitative factors are used to make clinical decisions on individual fragility, and adjust anaesthetic drug accordingly. Hence, data is not collected in a fully controlled environment with explicit protocol, eventually

leading to confounding between preoperative health, drug dosage and intraoperative biosignals. Moreover, the scarcity of research-grade equipment could invalidate existing research-based approaches. Second, despite muscle-inhibitor drugs, the EEG signal during anaesthesia can be noisy: certain surgical events can induce artefacts, transient electrode loss is not uncommon due to high amplitude environmental artefacts, clipping on subset of electrodes can happen in certain EEG devices, etc. Third, doctors use different drugs during the GA procedure, that are known to modify EEG brain spectral signatures [Pur+15b]. Finally, defining with a single criteria the stable anaesthesia period to optimally extract biomarkers pose difficulties as young/old, healthy/less healthy people can have very different EEG landmarks, *e.g.,* alpha peaks, making it likely to throw away useful data or include noise.

Brain age, the age of the brain organ, is an easy notion to grasp for the general public, and could therefore be an appealing communication tool for discussing brain health in the clinic [Den+21]. Yet its clinical value should be thoroughly demonstrated before introducing it as a new biomarker. Also, knowing that BA is trained by predicting age of healthy patients, the complementarity of BA with age is a particularly important question to investigate: BA should not be redundant with age, an already ubiquitous marker of multiple clinical outcomes [Dad+21]. The information extracted from biological sources by the BA should at least be complementary to age to predict important clinical outcomes like health status or neurocognitive disorders. One possible way to assess cognitive disorders for instance is to detect Burst Suppression (BS), an EEG pattern of alternating periods of iso-electric suppression and high amplitude waves, that has been linked to postoperative cognitive dysfunctions [Fri+16; Wil+19], a disorder that touches about one third of anaesthetized 60y+ people. Leveraging the signals collected in the operating room, brain age could be an important prospective biomarker to predict BS and a prevention tool candidate of postoperative complications.

Hence the questions we seek to answer: Can we predict the BA in the clinic from EEG during GA, *i.e.,* is the translation of lab-developed BA valid in GA settings? Does the BA have a clinical diagnostic value, can it be considered as a biomarker of neurocognitive disorders? Does the drug impact BA prediction under GA and how to take it into account? To investigate these questions we collaborated with the anaesthesia-critical care department of Lariboisière hospital in Paris. With ∼14 000 anaesthesia every year and 14 operating rooms, Lariboisière hospital is one of the major hospitals in France.

This chapter is organized as follows. Section 3.2 describes the methods used to investigate our research questions, with a focus on the impediments met to transform raw clinical data into an actionable dataset. Section 3.3 is devoted to data exploration in order to gain some intuitions on the relationships between age, brain age, EEG power, health status and drugs. The next Section 3.4 is devoted to the results of our

study, answering each of our three scientific questions: Section 3.4.1 the feasibility of clinical translation of BA to GA-settings, Section 3.4.2 the validation of model predictions against the occurrence of BS, and Section 3.4.3 the impact of GA drugs on BA prediction. Section 3.5 proposes a discussion on the results and a perspective on future work.

3.2 Methods

3.2.1 General anaesthesia setting

General Anesthesia can be roughly defined as an artificial state of coma, induced by a specific hypnotic drug for the time of a surgery. In this state, the protective reflexes are lost: body temperature is not regulated anymore, breathing is not automatic [BLS10]. GA is therefore a technical procedure during which an anesthetist should stand in to preserve the normal functioning of the organs and keep the patient in a stable state.

GA is composed of three phases: induction ($\sim 25\,\mathrm{min}$), maintenance (duration of surgery) and awakening ($\sim 1\,\mathrm{h}$). In the induction phase the patient is administered numerous anaesthetic drugs as an intravenous continuous flow (not push or bolus) to reach desired stable anaesthesia state. In the maintenance phase the main physiological constants of the patient are monitored by the anaesthetist (oxygen saturation, arterial pressure, heart rate and BS) to make sure the patient stays in a stable state.

During induction, among the multiple anesthetic agents used during GA, the hypnotic drug is responsible for inducing sedation and promoting loss of consciousness. Common hypnotic drugs include propofol, ketamine or halogenated gas (desflurane, sevoflurane), each of them with their own spectral brain signature [Pur+15b]. At Lariboisière, patients are usually induced using propofol, and maintained with either propofol or sevoflurane.

During maintenance, to monitor the depth of anaesthesia, doctors are guided by EEG-derived commercial indices like the bispectral index BIS or the Patient State Index PSI. Keeping these indices in the standard range of values is recommended to maintain the patient in a stable anaesthesia state. The depth of anaesthesia can also be monitored using the Spectral Edge Frequency index SEF95, defined as the frequency under which 95% of the cumulative sum of the normalized spectral power is contained. Hence, the smaller the SEF95, the less conscious the patient and the deeper the anaesthesia. A stable state of anaesthesia hypnosis is usually defined by a SEF95 index contained in the [8-13]Hz range [Bru+03].

We focused on general anaesthesia procedures only, discarding local and loco-regional procedures, all using total intravenous propofol as induction drug and then either propofol or sevoflurane as maintenance drug (we excluded 15 patients maintained by desflurane). Only patients undergoing neuroradiology interventions and orthopedic surgeries were included in this translational study. These two types of intervention cover patients with varied age and health status: neuroradiology interventions mostly concern young and healthy populations without neurological

antecedents, orthopedic surgeries are more often performed on aged patients known for cognitive disorders.

3.2.2 Data collection

Cerebral activity during GA was monitored using Masimo device with a 4-frontal electrodes EEG montage (Fp1, Fp2, F7 and F8), sampled at $63\,\mathrm{Hz}$ by default. EEG cap is placed on the patient's scalp a few minutes before injection of the drugs and removed at the end of anaesthesia. Peroperative EEG data are then extracted from the device, anonymized and stored on a file server in EDF format. APHP security policy enforced a restricted access to this private and secured server, only accessible from the Lariboisière local network to a small number of individuals. Each EDF file is completely anonymized and stored under a directory named after the corresponding operating room. Because of this anonymization procedure we had to match each patient with its EDF files using only the date/time of surgery and the name of the operating room, recorded in a separate file.

We also had access to non-EEG information collected during the mandatory anaesthesia-specific medical consultation, including demographic data (age, gender, weight, height, BMI), clinical information (type of surgery, type of anaesthesia, type of drug, ASA score, blood pressure) and medical information (neurological and cardiovascular antecedents, neurological and cardiovascular treatments, cardiovascular risk factors). These binary medical scores were either too scarcely collected or too unbalanced to be used in our statistical study. By contrast the ASA score $\in \{1, 2, 3\}$, a standardized score indicating preoperative physical health status (the lower the healthier), was more systematically collected. We also had access to information manually recorded by anesthetists during the intervention, among which the quantity and timing of administered drugs, timing of specific events and peroperative variables. Nevertheless, lack of digitization prevented us from easily using this information.

To detect iso-electrical suppressions (first part of the BS pattern) from intraoperative EEG we adapted the method from [Car+19]. For each EEG, a trained clinician identified intraoperative periods based on the alpha-band. From this intraoperative EEG signal $S(t)$, we first discarded flat artifacts searching for segments below $0.1\,\mu\mathrm{V}$ amplitude, lasting at least $1\,\mathrm{s}$. Similarly, periods of high amplitude voltage above $80\,\mu\mathrm{V}$ were also removed. To anticipate possible amplitude drift during the intervention, we rescaled the signal by first computing a rolling standard deviation over $30\,\mathrm{s}$-long time windows $S_{\mathrm{std}}(t)$, then building $\hat{S}(t) = S(t)\left(1 + 2\frac{\langle S_{\mathrm{std}}(t)\rangle}{S_{\mathrm{std}}(t)}\right)/3$, where $\langle\cdot\rangle$ indicates the temporal mean. A mask was constructed based on regions where $|\hat{S}(t)| < 2.5\,\mu\mathrm{V}$, then we applied in series a $0.2\,\mathrm{s}$ erosion, a $1\,\mathrm{s}$ dilation and an $0.8\,\mathrm{s}$ erosion. This output was used to estimate time and fraction of time spent in iso-electrical suppression during the entire intraoperative period including the induction

(the first $25\,\mathrm{min}$) and the maintenance phasess (after $25\,\mathrm{min}$). We focused on the maintenance period for the better statistical properties of the signal (more samples due to a longer period, closer to stationarity and less artefacts) and potentially lower false positive rate (drug-induced BS can arise at induction more often). However, this lengthy period is also more prone to confounding effects of drugs used to stabilize the patient's vital signs. Once detected, these episodes of iso-electrical suppressions were automatically discarded from the signal via artefact rejection, hence never used in subsequent steps including feature extraction and modeling.

3.2.3 Data curation

Neuroscience data provided by research consortia are usually cleaned, processed and curated before made public, which facilitates further analysis. In this study, we directly worked with raw clinical data, originally collected for monitoring purposes. High quality monitoring is indeed the priority of anaesthetists to guarantee the quality of the GA procedure in a highly controlled clinical environment. To reuse the data for research purposes, an essential part of my thesis work was to develop a data curation strategy.

First, since the EEG data acquisition process has not been carried out to answer a dedicated research question, the conditions of data collection were not ideally controlled, which is common in this situation. Doctors used different strategies to administer propofol. Manual infusion consists in injecting a single dose of drug (bolus) and monitoring some physiological constants (heart rate, arterial pressure) for potential adjustments. Target Controlled Infusion (TCI) leverages modern syringe drivers that administer the drug continuously and automatically so that the cerebral concentration saturates at a desired target value, using only age and weight as input. Stable Anaesthesia requirement let the doctor decide the dosage to meet a certain criteria like SEF95 index in the [8-13]Hz range or to compensate for perceived resistance to anaesthesia. Besides intervention in drug administration, we had to deal with other impediments in the data acquisition process. Some recordings failed for practical reasons resulting in several small unusable EDF files. Some metadata was lost as some students, helping with the process, used their own Excel template for their thesis with different formats, which ended up not being consolidated with the main metadata file. Finally, the file server storing EEG data was not ideally organized, with inconsistent directory naming structure, duplicated patients under different directories, nested patient directories, ghost directories containing patients operated in an unreported room number, etc. To mitigate this last obstacle, largely unavoidable, the Lariboisière team developed a script running every day on the file server and providing an updated view on all EDF files in a searchable database. A dedicated script has also been developed to match each patient with its EDF files folder. Finally, we assigned a unique identifier to each patient for each visit at the

hospital, as it is important to distinguish between two visits of the same patient, which can then present different characteristics (age, health status, etc.).

Regarding the EEG data quality, we also had to overcome multiple obstacles. Masimo devices impose a maximum file size to each recorded EDF file, leading to multiple EDF files per patient. The concatenation of these files was non-trivial because of two reasons. First, the data collection procedure led to excluding patients associated with only one EDF file of small size (considered as recording failure) and patients with multiple files but whose first and second file are incomplete. For the remaining patients we excluded the first file if it was incomplete and if the second file is at least 80% of its max size (a symptom that the device has been unplugged for some reason, generally to facilitate the device transfer between the induction and operating room). Second, the EEG device used in the operating room has not been designed for research but for monitoring usage. Hence the raw data it records strictly follows data visualized on its screen. For instance, if the extreme traces on the visualisation screen (which corresponded to the two electrodes Fp1 and F8) were clipped, the recorded raw data were also clipped. The raw data are therefore modified whenever the device operator changes the visualization settings. Hence, the calibration factor between two consecutive files could be different due to Masimo's operator changing y-scale on the monitor, *e.g.,* to avoid clipping of the extreme raw EEG traces during recording. For this reason we also excluded files that do not have the same calibration factor than the first file in order to avoid error at concatenation. Finally the remaining files were downsampled to a common sampling frequency of 63Hz before concatenation. The downsampling was necessary because sampling frequency could change between two consecutive files due to Masimo's operator changing x-scale on monitor during recording.

Regarding the metadata collection process, these data were recorded by hand by each medical doctor on a paper sheet, before being further consolidated in an Excel file. This process inevitably led to missing values: the anesthetists helped us to go through the archive department of the hospital and look at the paper record of a few dozens of patients one by one to retrieve their drugs and their ASA score. We also encountered incorrect values (*e.g.,* due to a change of convention between cohorts in the encoding of the gender attribute), inconsistent values (*e.g.,* some patients having two different maintenance drugs, which could go unnoticed when filtering for the presence of a particular drug), and finally unnormalized values (*e.g.,* the letter O in place of the number 0). These errors were corrected by hand, which is a common reality in clinical settings. As a side note of interest, the growing number of research projects within the anaesthetists team has recently led to the hiring of a dedicated data manager, leading the data curation effort.

Finally, from the 518 patients with demographic information, 473 underwent general (and not loco-regional) anaesthesia among which 435 had their maintenance drug

informed. Among those patients, 348 had a successful EDF-files matching and 345 with successful EDF files concatenation, leading to a dataset of 345 patients with both a proper concatenated EEG recording (in FIF format) and metadata information. Finally we converted the resulting EEG dataset into the standard and anonymized BIDS format [Per+19] with EDF files in Brainvision format using the MNE-BIDS Python library [App+19]. The statistics of the cohort is summarized in Table 3.1, along with a stratified view by the grouping variable ASA and drug.

3.2.4 Data processing and feature extraction

The data processing has been carried out using 'mne-bids-pipeline' (https://mne.tools/mne-bids-pipeline/). This tool is not a Python library but a suite of Python scripts that generate processed data in a FIF format from BIDS files. Processing instructions are contained in a single configuration file that can be tailored to particular needs. For our study, we minimally used it to generate epochs: for every patient, the signal is epoched into $60\,\mathrm{s}$ sliding windows shifted by $10\,\mathrm{s}$.

Intermediate power representations of the signal were computed using the open-source Python library 'Coffeine' (https://github.com/coffeine-labs/coffeine). In each window, the power spectral density, the covariance matrices and the cross-frequency covariance matrix are averaged across internal Hamming windows of $8\,\mathrm{s}$ shifted by $4\,\mathrm{s}$. The power spectral density was estimated in $244\,\mathrm{f}$requency bins between $0\,\mathrm{a}$nd $32\,\mathrm{Hz}$ and its averaging has been further robustified by trimming distribution from both tails at $25\,\%$ cutoff. The covariance matrices of the four EEG channels are estimated using the 'OAS' shrinkage method[Che+10] in each of five frequency bands as described in Table 3.2 leading to five 4×4 matrices. The cross-frequency covariance matrix is the covariance matrix of the 20 'virtual' channels constructed from each four real channels filtered in each five frequency bands, leading to a single 20×20 matrix. This matrix allows to investigate coupling between frequency bands, while still enjoying the same mathematical guarantees obtained in Chapter 1 as it is a covariance matrix. Besides these covariance features, EEG-characteristics commonly used to judge the signal quality during monitoring are also computed: the SEF95 index and the maximum peak-to-peak amplitude, as they are used to perform epochs-selection.

Epochs with peak-to-peak amplitude lower than $0.1\,\mu\mathrm{V}$ on any one electrode were discarded on all electrodes, avoiding learning from BS periods or clipped segments. Epochs are then selected within the stable anaesthesia period, defined by the SEF95 index belonging to the [8-13]Hz range. We focused on the longest period of consecutive epochs of stable anaesthesia using the average SEF95 across the 4 channels, which gave better results than concatenating all epochs satisfying the SEF95 constraint. The PSD, the covariance matrices and the cross-frequency covariance matrix

	Overall (N=330)
Age	
Mean (SD)	54.41 (19.43)
Median (Q1, Q3)	56.00 (37.25, 69.00)
Min - Max	16.00 - 99.00
Missing	0
Gender	
female	213 (64.5%)
male	117 (35.5%)
Missing	0
Height	
Mean (SD)	167.58 (9.74)
Median (Q1, Q3)	168.00 (160.00, 174.00)
Min - Max	123.00 - 195.00
Missing	4
Weight	
Mean (SD)	74.39 (17.27)
Median (Q1, Q3)	73.00 (60.00, 85.00)
Min - Max	37.00 - 145.00
Missing	6
BMI	
Mean (SD)	26.35 (6.52)
Median (Q1, Q3)	25.39 (22.32, 29.54)
Min - Max	0.00 - 54.20
Missing	8
ASA	
ASA1	69 (21.3%)
ASA2	185 (57.1%)
ASA3	70 (21.6%)
Missing	6
Neurological Antecedents	
No	162 (68.4%)
Yes	75 (31.6%)
Missing	93
Cardiovascular Antecedents	
No	194 (87.4%)
Yes	28 (12.6%)
Missing	108
Cardiovascular Risk Factor	
No	109 (42.1%)
Yes	150 (57.9%)
Missing	71
Cardiovascular Treatment	
No	151 (64.3%)
Yes	84 (35.7%)
Missing	95
Neurological Treatment	
No	175 (80.6%)
Yes	42 (19.4%)
Missing	113
Drug	
propofol	220 (66.7%)
sevoflurane	110 (33.3%)
Missing	0
Prop. time spent in BS during induction	
Mean (SD)	4.74 (9.55)
Median (Q1, Q3)	0.96 (0.08, 4.14)
Min - Max	0.00 - 72.09
Missing	20
Prop. time spent in BS during maintenance	
Mean (SD)	3.65 (7.85)
Median (Q1, Q3)	0.62 (0.05, 3.33)
Min - Max	0.00 - 77.17
Missing	20

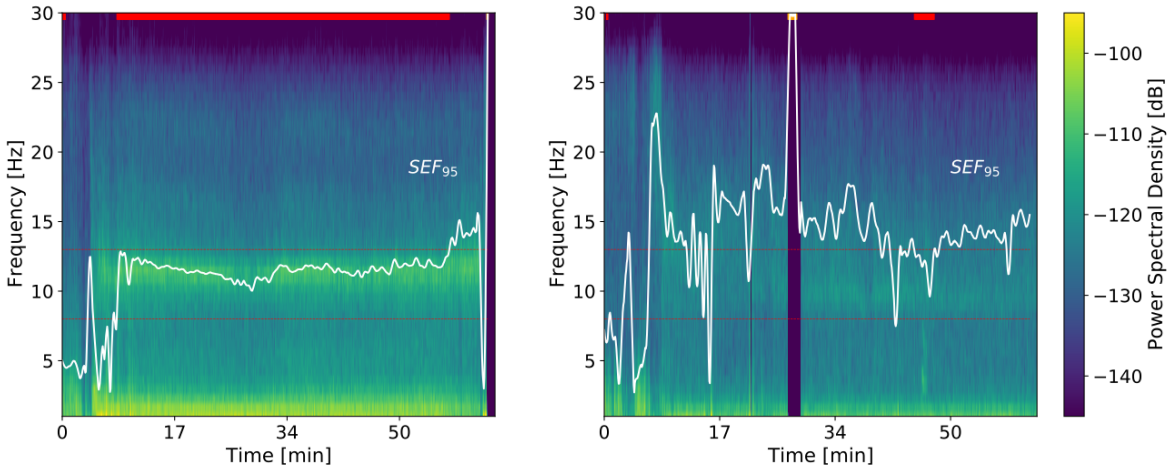
Tab. 3.1: Descriptive summary statistics table of Lariboisière data.

	ASA1.propofol (N=43)	ASA2.propofol (N=130)	ASA3.propofol (N=41)	ASA1.sevoflurane (N=26)	ASA2.sevoflurane (N=55)	ASA3.sevoflurane (N=29)	Total (N=324)	p value
Age								< 0.001
Mean (SD)	41.37 (14.84)	51.73 (18.76)	65.17 (14.60)	35.12 (12.39)	62.40 (17.21)	70.79 (11.85)	54.24 (19.35)	
Median (Q1, Q3)	44.00 (30.00, 51.50)	54.00 (33.25, 68.00)	66.00 (54.00, 75.00)	35.00 (25.25, 43.00)	63.00 (47.50, 75.00)	71.00 (60.00, 78.00)	56.00 (37.00, 69.00)	
Min - Max	16.00 - 85.00	19.00 - 89.00	27.00 - 91.00	17.00 - 59.00	22.00 - 98.00	48.00 - 99.00	16.00 - 99.00	
Missing	0	0	0	0	0	0	0	
Gender								0.092
female	28 (65.1%)	90 (69.2%)	29 (70.7%)	19 (73.1%)	29 (52.7%)	14 (48.3%)	209 (64.5%)	
male	15 (34.9%)	40 (30.8%)	12 (29.3%)	7 (26.9%)	26 (47.3%)	15 (51.7%)	115 (35.5%)	
Missing	0	0	0	0	0	0	0	
Height								0.074
Mean (SD)	168.81 (9.33)	167.04 (10.18)	164.90 (7.61)	171.85 (9.03)	168.17 (9.85)	167.86 (10.18)	167.66 (9.71)	
Median (Q1, Q3)	169.00 (162.00, 174.50)	166.00 (160.00, 172.75)	162.50 (160.00, 170.25)	170.50 (166.50, 180.00)	170.00 (160.50, 175.00)	167.00 (160.00, 172.00)	168.00 (160.00, 174.00)	
Min - Max	152.00 - 187.00	123.00 - 190.00	150.00 - 180.00	154.00 - 191.00	142.00 - 192.00	150.00 - 195.00	123.00 - 195.00	
Missing	0	0	1	0	1	1	3	
Weight								0.040
Mean (SD)	67.49 (12.03)	76.13 (16.78)	76.75 (19.57)	69.31 (12.81)	74.45 (17.36)	78.29 (22.43)	74.40 (17.21)	
Median (Q1, Q3)	68.00 (58.00, 75.00)	76.00 (65.00, 85.00)	78.00 (60.00, 89.00)	69.50 (60.00, 79.50)	69.00 (65.00, 85.00)	77.00 (60.00, 88.50)	73.00 (60.50, 85.00)	
Min - Max	45.00 - 93.00	40.00 - 145.00	37.00 - 128.00	48.00 - 90.00	48.00 - 116.00	44.00 - 137.00	37.00 - 145.00	
Missing	0	1	1	0	2	1	5	
BMI								< 0.001
Mean (SD)	23.61 (3.44)	27.20 (6.66)	28.21 (7.88)	23.31 (3.50)	25.70 (6.29)	27.67 (7.43)	26.31 (6.46)	
Median (Q1, Q3)	22.89 (21.26, 25.75)	26.06 (23.44, 30.64)	27.43 (22.26, 33.15)	23.57 (20.55, 25.59)	25.39 (22.57, 28.48)	25.83 (23.36, 31.75)	25.39 (22.31, 29.59)	
Min - Max	18.52 - 32.18	0.00 - 54.20	14.45 - 50.00	17.01 - 30.10	0.00 - 39.90	17.19 - 54.19	0.00 - 54.20	
Missing	0	1	3	1	1	1	7	
Neurological Antecedents								0.029
No	30 (78.9%)	76 (66.7%)	22 (66.7%)	2 (100.0%)	23 (76.7%)	5 (33.3%)	158 (68.1%)	
Yes	8 (21.1%)	38 (33.3%)	11 (33.3%)	0 (0.0%)	7 (23.3%)	10 (66.7%)	74 (31.9%)	
Missing	5	16	8	24	25	14	92	
Cardiovascular Antecedents								< 0.001
No	38 (100.0%)	107 (93.9%)	19 (61.3%)	2 (100.0%)	21 (91.3%)	3 (33.3%)	190 (87.6%)	
Yes	0 (0.0%)	7 (6.1%)	12 (38.7%)	0 (0.0%)	2 (8.7%)	6 (66.7%)	27 (12.4%)	
Missing	5	16	10	24	32	20	107	
Cardiovascular Risk Factor								< 0.001
No	34 (89.5%)	51 (44.3%)	7 (19.4%)	2 (66.7%)	14 (37.8%)	0 (0.0%)	108 (42.5%)	
Yes	4 (10.5%)	64 (55.7%)	29 (80.6%)	1 (33.3%)	23 (62.2%)	25 (100.0%)	146 (57.5%)	
Missing	5	15	5	23	18	4	70	
Cardiovascular Treatment								< 0.001
No	34 (89.5%)	74 (67.3%)	13 (37.1%)	3 (100.0%)	21 (70.0%)	4 (28.6%)	149 (64.8%)	
Yes	4 (10.5%)	36 (32.7%)	22 (62.9%)	0 (0.0%)	9 (30.0%)	10 (71.4%)	81 (35.2%)	
Missing	5	20	6	23	25	15	94	
Neurological Treatment								0.126
No	35 (92.1%)	84 (79.2%)	26 (76.5%)	2 (100.0%)	20 (90.9%)	6 (60.0%)	173 (81.6%)	
Yes	3 (7.9%)	22 (20.8%)	8 (23.5%)	0 (0.0%)	2 (9.1%)	4 (40.0%)	39 (18.4%)	
Missing	5	24	7	24	33	19	112	
Prop. time spent in BS during induction								< 0.001
Mean (SD)	2.10 (4.85)	5.45 (9.68)	10.55 (15.41)	0.84 (1.30)	2.04 (3.48)	4.55 (11.31)	4.67 (9.55)	
Median (Q1, Q3)	0.19 (0.02, 1.27)	0.99 (0.11, 6.54)	5.04 (1.27, 14.50)	0.03 (0.00, 1.34)	0.73 (0.18, 1.72)	0.84 (0.31, 2.63)	0.95 (0.07, 3.94)	
Min - Max	0.00 - 24.06	0.00 - 42.38	0.05 - 72.09	0.00 - 5.09	0.00 - 18.21	0.00 - 48.41	0.00 - 72.09	
Missing	1	2	3	4	6	4	20	
Prop. time spent in BS during maintenance								< 0.001
Mean (SD)	1.43 (3.73)	4.41 (7.28)	8.27 (14.23)	0.34 (0.77)	1.19 (2.27)	3.31 (9.16)	3.58 (7.86)	
Median (Q1, Q3)	0.28 (0.01, 1.33)	1.01 (0.14, 5.27)	3.47 (0.95, 8.01)	0.01 (0.00, 0.19)	0.36 (0.04, 1.25)	0.63 (0.06, 1.70)	0.57 (0.05, 3.12)	
Min - Max	0.00 - 23.04	0.00 - 32.33	0.00 - 77.17	0.00 - 3.18	0.00 - 10.86	0.00 - 44.02	0.00 - 77.17	
Missing	1	2	3	4	6	4	20	

name	low	δ	θ	α	β
range (Hz)	$0.1-1.5$	$1.5-4$	$4-8$	$8-15$	$15-30$

Tab. 3.2: Definition of frequency bands

are then averaged across the selected epochs. The epochs-selection mechanism is illustrated on Fig. 3.1.



.

Fig. 3.1: Illustration of the epochs selection mechanism. Power spectral density of a subject over the entire time of his surgery (spectrogram). The subject is maintained in GA under propofol. The SEF95 index is drawn in white. The epochs with peak-to-peak amplitude lower than $0.1\,\mu$V on any one electrode are identified by an orange mark at the top of the spectrogram. They correspond to burst suppression episodes and are discarded from subsequent analysis. The stable anaesthesia period is defined by the largest consecutive period for which the SEF95 index is in the [8-13]Hz range. This range is depicted by red dashed lines. The corresponding epochs are identified by a red mark at the top of the spectrogram. We used this period to compute the covariance matrices. **(Left)** A young subject in good pre-operative health (aged 28 with an ASA1 score) shows no episode of burst suppression, a long period of stable anaesthesia and a strong increase in alpha power during sedation. **(Right)** An old subject with pathologies (aged 74 with an ASA3 score) entered a long episode of burst suppression before a rather short period of stable anaesthesia. Their alpha power was weak during the whole surgery.

As features, we used the previous EEG signatures of brain aging developed in Chapter 1: the upper part, the log-diagonal or the Riemannian embedding of the covariance matrices. In this chapter we also introduce the Riemannian embedding of the cross-frequency covariance matrix as an interesting and statistically tractable option due to the fewer number of channels available in the clinic. For this 4-channels EEG experiment, it leads to a vector of size 210, still statistically acceptable compared to our sample size. For a 21-channels EEG TUH experiment its size would have been 5050.

As a side note of interest, feature extraction was performed using the open-source Python library 'Coffeine' (https://github.com/coffeine-labs/coffeine) we developed in the team as a result of this thesis. I contributed the core features of this library implementing all the methods developed during my PhD and presented in our articles [Sab+19a; Sab+20; Eng+20]. In particular, this library provides a high-level interface to the predictive modeling techniques we developed and presented in 2 using the M/EEG covariance matrix as representation of the signal.

3.2.5 Machine learning and statistical modeling

For the age prediction task of this Lariboisière experiment, we will benchmark the regression algorithms previously introduced in Chapter 1 *i.e.,* simple regularized linear regression model (ridge regression) applied to particular vectorizations of the covariance matrices: its upper part ('upper' model), its log-diagonal ('log-diag' model), its Riemannian embedding ('Riemann' model) and the Riemannian embedding of the cross-frequency covariance matrix, as defined in Section 3.2.4. The Riemannian model was defined with no projection step since the covariance matrices are here full rank. Indeed, they are estimated from sufficiently large chunk of signals and are not rank-reduced by preprocessing steps (see Chapter 1). Classical EEG average reference, that amounts to projecting the signal into the subspace orthogonal to the average signal, reduces the rank by one, but is not used in this study. The regularization parameter of ridge is tuned by generalized cross-validation [GHW79] on a logarithmic grid of 100 values in $[10^{-5}, 10^{3}]$ on each training fold of a 10-fold cross-validation loop. For each model we standardized the features enforcing zero mean and unit variance, a standard preprocessing step for penalized linear models. To compare models against chance, we estimated the chance-level empirically through the performance of a dummy-regressor predicting the mean outcome of the training data without trying to find patterns, thus equivalent to random guess. Uncertainty estimation was obtained from the cross-validation distribution.

Note that formal hypothesis testing for model comparison was not available for any of the datasets analyzed as this would have required several datasets, such that each average cross-validation score would have made one observation. For data exploration and statistical modeling, we used standard classical two-sample statistical tests: Welch two sample t-test of difference means for continuous/categorical variables and Pearson's chi2 of independence for categorical/categorical variables.

To assess the impact of drug on BA estimation we compared the performance of our brain age model when learning from two different drugs (propofol or sevoflurane). To investigate their cross-effects, we developed different classes of models. Drug-specific models learn from patients under one drug and predict on patients under the same (propo/propo, sevo/sevo). The drug-agnostic model learns and predicts from all patients without being informed on their drug (all/all). Drug-crossed models

learn from patients under one drug and predict age of patients under the other (propo/sevo or sevo/propo). Drug-aware models learn jointly from all patients with drug-interaction effects and predict on either drug (joint/propo, joint/sevo): they learn a compromise between a global model that ignores drugs, and a specific model handling exceptions for the drugs, leveraging what the two drugs have in common. Drug-specific models and the drug-agnostic model are simple restrictions of our model to different sub-populations. To compute CV-based uncertainty estimates of the performance of drug-crossed models, we splitted the training population into 10 folds (with shuffling), trained the model on 9 folds and tested this fitted model on the all testing population. This allows to probe the variance due to random training population. To implement the drug-aware model (joint model with interaction effect) we expanded the original p-feature vector to a new $(2p + 1)$-feature vector where the first p features are the original features, then the drug indicator variable (1 for propofol, 0 for sevoflurane) then the product between the two, leading to either a copy of the p features if propofol or a p-vector of zeros if sevoflurane. All these models have been implemented in the 'Coffeine' library.

3.3 Data exploration

Before diving into predictive modeling we visualized our sample to explore the link between EEG during stable anaesthesia and age: is the age indeed visible in the brain? The effect of age on EEG-brain signals could nevertheless be confounded by external factors.

We identified two major potential confounders: the health status and the maintenance drug. Intuitively, the health status is likely to be associated to both the age and the EEG power, which has been confirmed by our anaesthetists collaborators and statistical tests. The maintenance drug has a known effect on EEG power, *e.g.,* propofol induces an anteriorization of the alpha rhythm during loss of consciousness [Vij+13]. Running basic statistical analysis we also discovered its association with age: the age distribution under propofol is significantly shifted towards younger people compared to sevoflurane (t=4.73 p=8.065e-06). After investigation, we discovered that this link was not causal (the doctors confirmed the choice of maintenance drug is not driven by age) but more likely due to random circumstances: halogenated gas happened not to be available to neuroradiology interventions which mostly concerns young and healthy subjects. Also, the data partially originates from a study focused on propofol effect, hence the biggest proportion of this drug in our sample. In general, the still most commonly used hypnotic drugs are halogenated gas like sevoflurane for their ease of use, which amounts to opening a valve, and their lighter monitoring requirement. Yet propofol allows for a quicker and finer control of the anaesthetic state via TCI with a cleaner and more stable EEG spectral signature [Pur+15b], hence was most commonly used in our study. We'll therefore look at how the EEG-age relationship is shifted by ASA and by drug, and run basic statistical tests to assess whether these relationships are likely to generalize to the population.

In Fig. 3.2 we explore how the link between EEG and age is modulated by health status.

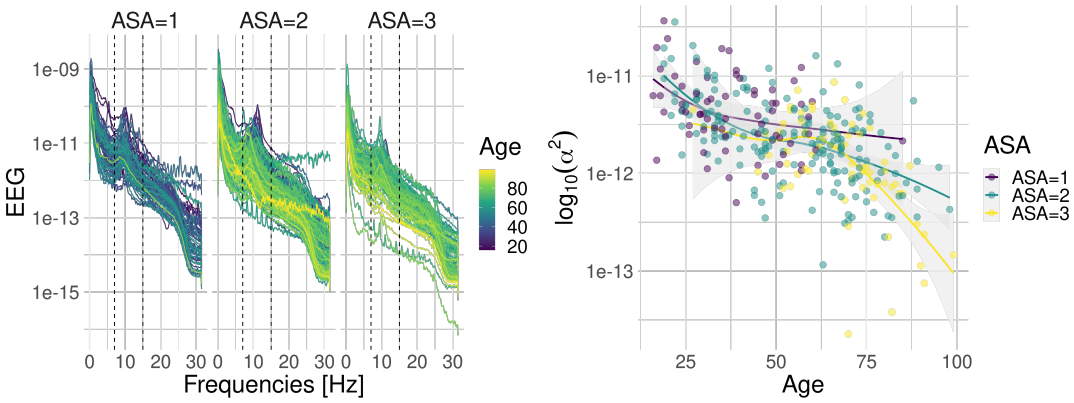


Fig. 3.2: Exploring how age impacts EEG power, depending on general health conditions. (Left) EEG power spectra during stable general anaesthesia of all patients in the Lariboisière dataset. Each line represents the PSD of a patient, color-coded by age (the older the brighter) and grouped by ASA. **(Right)** Alpha power during stable general anaesthesia vs age of all patients in the Lariboisière dataset. Each dot represents the log alpha power of a patient, color-coded by ASA score.

Left panel shows that ASA seems to strongly modulate the age-EEG link. Older patients tend to have a higher ASA score, a lower alpha peak and a lower power across frequencies. It gives us a hint that a brain age estimator that doesn't take the health status into account won't be optimal. Right panel shows, as expected, that alpha power decreases with age. ASA1 vs ASA3 overlap only on young patients.

In Fig. 3.3 we explore how the link between EEG and age is modulated by the drug used during the maintenance phase.

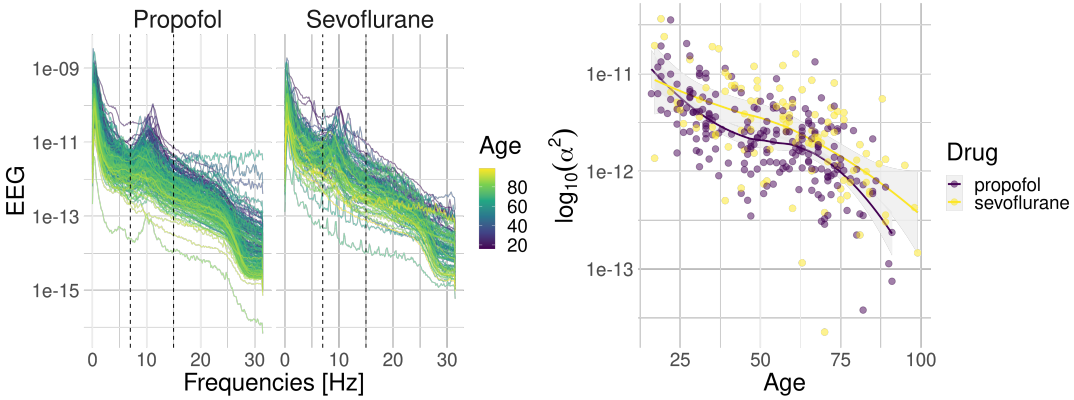


Fig. 3.3: Exploring how age impacts EEG power, depending on the drug used to maintain GA state. (Left) EEG power spectra during stable general anaesthesia of all patients in Lariboisière dataset. Each line represents the PSD of a patient, color-coded by age (the older the brighter) and grouped by drug. **(Right)** Alpha power during stable general anaesthesia vs age of all patients in Lariboisière dataset. Each dot represents the log alpha power of a patient, color-coded by drug.

In the left panel we see a clear dependence of the PSD on age: the younger the more power across all frequency bands and across drugs. Drugs seem to modulate this

link essentially in alpha band, with an apparent stronger power under sevoflurane. In the right panel we thus focus on the alpha log power where there seem to be systematic differences in alpha power under both drugs, confirmed by hypothesis testing (t=-2.1, p= 0.035). Also, a linear regression analysis of the alpha log power on age gender and drug for healthy patients suggests that age and drug have independent (additive) effects in opposite direction but of similar magnitude: age reduces (and sevoflurane increases) power by 0.5 STD per STD (beta_age=-0.62, t=-13.8, p<2e-16 | beta_drug=0.44, t=4.7, p=4e-6). We therefore find systematic differences in EEG between propofol and sevoflurane, when taking age into account. It confirms that drugs may influence prediction and calls for a thorough drug impact study, which will be developed in the results Section 3.4.3.

3.4 Results

3.4.1 Brain age during General Anesthesia

In this section, we aim to answer our first question: can we predict a brain's age in the clinic from EEG during anaesthesia? So far, we estimated the brain age from research-grade MEG or EEG devices with multiple channels and high time/frequency resolution, under a carefully controlled data collection procedure, on patients in resting state with no chemical substance known to modify brain activity (see Chapter 2). The particular setup of anaesthesia described earlier could challenge brain age estimation. Here we restrict our analysis to healthy patients (ASA1 or ASA2 score), maintained under propofol, within the stable anaesthesia period defined as the longest consecutive period for which the SEF95 index is in [8-13]Hz.

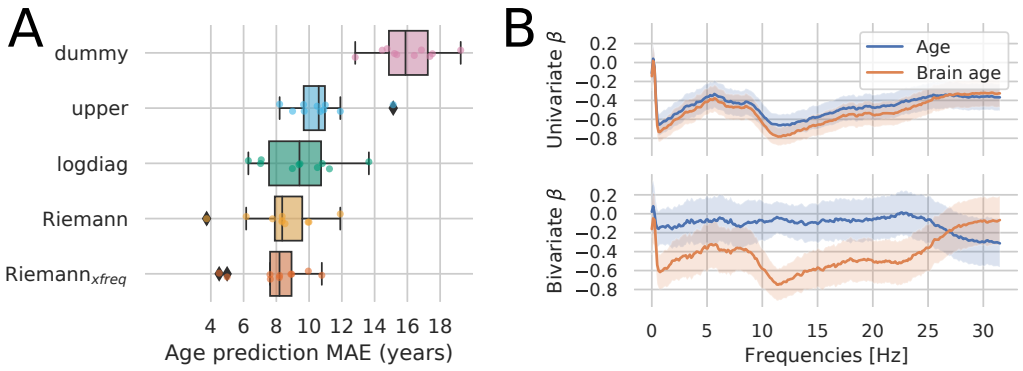


Fig. 3.4: Brain age during general anaesthesia. **(A)** *Brain age models comparison.* Performance of five different regression models in predicting age of healthy patients during the stable maintenance period of general anaesthesia under propofol, expressed as Mean Absolute Error (MAE). Uncertainty estimates are based on a 10 fold cross-validation procedure. The Riemannian model applied to cross-frequency covariances, that allows to explore the interaction between multiple frequency bands, shows the best performance with a MAE of 7.9y, compared to the 16y MAE of predicting the population average (dummy model). This model is used for all subsequent analyses. It demonstrates that brain age during GA can be estimated in the clinic. **(B)** *Brain age and EEG power spectrum.* Coefficients of Age (blue curve) and cross-validated Brain age (orange curve) in predicting the log EEG power of healthy patients under propofol for each frequency bin, using univariate (upper figure) or bivariate (lower figure) linear regression models. Uncertainty estimates of the coefficients are based on 95% confidence intervals. As expected, average powers decrease with brain age and age across the frequency spectrum. Brain age is complementary to age for predicting EEG powers, mostly replacing it in low and alpha ranges, while being redundant in theta and high beta.

Brain age models comparison We compared the performance of five regression strategies: 'upper', 'log-diag', 'Riemann' and 'Riemann on cross-frequency matrix'. The results are depicted in Fig. 3.4(A). The Riemannian model applied to the cross-frequency covariance matrix is a clear winner. It is noteworthy that the best performance based on the Riemannian model and the ordering of the models in this clinical experiment (4 EEG-channels for 345 subjects) is consistent with the results obtained on the experimental datasets studied in Chapter 2: Cam-CAN (102

MEG magnetometers channels for 596 subjects) and for TUH (21 EEG channels for 1385 subjects). It is an additional and independent hint that the Riemannian methods are the most efficient and robust models to perform M/EEG regression. This finding suggests that BA can be estimated during GA from routine monitoring EEG within the operative theater. All subsequent analyses are conducted using the best regression model of Riemann on cross-frequency covariance.

Brain age and EEG power spectrum. Does this new measure of brain-predicted age contain additional information on the EEG power spectrum than age? To investigate the complementarity of age and brain age we ran a statistical linear regression analysis of the log EEG power in every frequency bin using the age and the brain age as predictors. Their coefficients, as a function of frequency, are reported in Fig. 3.4(B).

Upper figure shows the coefficients of the two univariate linear regression models. As expected the marginal effect of age/brain age on the log powers (hence the correlations) are negative: the average brain power decreases with age and brain age across the frequency spectrum, the older the subject the less powerful his brain signal. This effect is more pronounced around low and alpha frequencies. Also brain age alone is at least as correlated to EEG powers than age. Therefore, when looking at EEG powers, anaesthetists can 'trust' this new biomarker as behaving similarly to the usual marker of age. We should note that since BA itself is built from powers as features it is not surprising to see this correlation. However, the brain age as input of this statistical model is cross-validated *i.e.,* corresponds to the age predicted out-of-sample (see [Eng+20; HT08] for the methodology of using cross-validated predicted scores for statistical analysis). Also, it is predicted from only five frequency bands while we here investigate fine-grained frequency patterns in EEG powers to see at which specific single frequency the model responds best.

These univariate models do not allow to disentangle both factors. The effect of brain age could be mainly due to the age itself. To regress out the effect of age on brain age we conducted the corresponding multiple regression analysis. Lower figure shows the coefficients of the multiple linear regression model using both age and brain age as predictors. They are again all negative across the frequency spectrum. Importantly, when brain age is included in the model, there is no unique contribution of age to the variation of power: brain age captures all the information about brain powers. This effect holds in the entire frequency spectrum with the exception of the high beta range. Therefore, brain age mostly replaces age for predicting EEG powers.

3.4.2 Clinical impact of GA-based Brain age

In this section, we aim to validate our brain age measure and answer our second question: does the BA have a clinical diagnostic value of perioperative complications?

Perioperative complications are the third leading cause of mortality in the world [Nep+19], potentially affecting any person undergoing GA (today one of the most frequent procedures in medicine, more than 300 million GA worldwide in 2020 [Csj], and 12 million in France). Although anesthesia and surgery are safer than before, it remains a risky procedure. They occur in about 25% of surgical operations involving general anesthesia [GBD09; Khu+05] and they have a strong impact on the patients' health: 9 million per year will die within 30 days of the intervention, 66 additional million will die 9 years younger on average [Khu+05]. In fact, perioperative complications have an even greater impact on the survival rate than the preoperative condition in major surgical operations [Khu+05]. Patients over the age of 60 are more likely to suffer from complications, the severity of which is in general greater. This age group represents about one quarter of the global population – but more than 40% of all anesthetic procedures in France [Dad+15] – and will continue to grow at a fast pace in the coming decades [Eurc; Eurb; Vol+20]: we expect twice more 65y+ humans in 2050 than today globally [Un2].

Perioperative cognitive disorders don't have precise definition in the US psychiatry reference manual DSM-5 classifying mental disorders, yet a possible classification proposed by US anaesthetists and based on the DSM begins to emerge [ZLJZS+19]. We can distinguish two main families. First, perioperative neurocognitive disorders (PND) are composed of three disorders of increasing duration of postoperative symptoms: postoperative delirium (POD), delayed neurocognitive recovery (DNCR) previously called postoperative cognitive dysfunction POCD, and neurocognitive disorders (NCD) that persist 30days after surgery. Second, cognitive decline (CD) is a loss of cognitive function (memory, language or thinking) not necessarily linked to surgery, scored by a psychometric test vs an anterior baseline, *e.g.,* by Montreal cognitive assessment (MOCA). It has recently been shown that patients experiencing CD have a higher incidence of PND [Fri+20]. Knowing patients' cognitive status would therefore allow the doctors to adapt anesthesia and postoperative care. Unfortunately, due to the large proportion of elderly patients going through GA, it is not practical to perform the neurocognitive evaluations necessary to assess CD on a large scale in the clinic: in France, approximately one third of 50y+ patients are scheduled for surgery every year. We propose to take advantage of GA to address this issue and investigate if our EEG-based BA (that can be estimated in the clinic during GA as seen in the previous section) can be used to provide an early diagnosis.

Some complications are directly related to surgery but the majority are related to the anesthesia management. Although ML can help surgeons to better prepare for

upcoming procedures with access to simulations beforehand and to monitor blood flow, anatomy, and physiology in real-time in the operating theater, few attempts have been made to introduce ML in anesthesia procedures. The usual role of the anesthesia team is 1) to allow surgery to proceed by administering drugs which have the side effects of deteriorating the entire cardiovascular system and 2) at the same time maintaining the patient's cardiovascular, pulmonary, renal and other status as stable as possible. Aside from the rare intraoperative complications such as allergy or difficult intubation, brain and cardiovascular complications represent the main perioperative complications and many recent clinical trials have tested various ways to reduce them. As certain patients are more likely to experience complications because of their age, medical history or risk of the surgery, a pre-anesthesia consultation was established in many countries to reduce risks related to anesthesia. However, pre-anesthesia consultation is a clinical exam that cannot properly explore brain function in detail. Pilot studies have shown that the risk of complications is linked with precise events happening during the operation including hypotension periods (blood pressure below a threshold of 65 mmHg), insufficient cerebral perfusion, and burst suppression patterns on the EEG. A few attempts have been made to prevent hypotensive episodes using a predictive algorithm based solely on blood pressure waves, but to the best of our knowledge, no device or algorithm has assessed brain viability and possible detrimental effect of hypotensive episodes on brain function. In the usual anesthetic management only depth of anesthesia is considered to prevent arousal states but the brain functional state is often not monitored in the operating room.

There is growing evidence that occurrence of intraoperative EEG BS patterns is associated with poor postoperative cognitive trajectories: it is an independent risk factor of POD [Fri+16], and can even predict PND in general (with the more permanent NCD) [Wil+19]. POD itself has been associated with a long term CD [Sac+12; Ino+16] and with an increased morbidity and healthcare costs [Mar17]. At last, peroperative BS and then postoperative cognitive dysfunction (POD and DNCR) appear to be linked to pre-existing CD that could be established before or during the surgery [Cul+17; Spr+17; BI+16]. Can GA BA be used to evaluate patient propensity to BS? We have early evidence that we can extract measures from EEG during GA that capture patient propensity to BS [Car+19] and that can be linked to preoperative CD as assessed with psychological measures like MOCA [Tou+20; JCPPV21]. Also, several studies have revealed the association between intraoperative alpha waves measured during maintenance and pre-existing CD or BS [Sha+20; Gia+17; Koc+19; Kre17]. Recently, the decrease of power spectral density in the alpha band (8–13Hz), collected on the frontal EEG under general anesthesia has been associated not only to chronological age [Pur+15a; Pur+15b] but also to preoperative CD [Koc+19].

In this study we will focus on how BA relates to two clinical targets: preoperative health status and postoperative neurocognitive disorders. Can the Brain age be considered a biomarker of postoperative neurocognitive disorders? To answer these questions we conducted a linear regression analysis of the BS rate during maintenance (denoted as BS) against age, brain age and health status. General health status was assessed by the clinical measure of ASA score. The physiological measure of the BS rate during maintenance has been preferred over the total time spent in burst suppression that depends on the total time of the GA procedure, which is highly variable between patients and that could have introduced a bias in our study (the length of surgery may itself be correlated with age for instance).

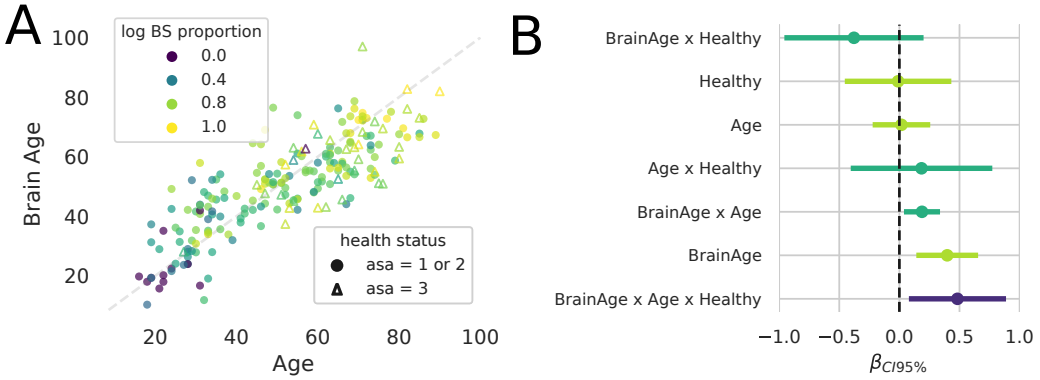


Fig. 3.5: Clinical impact of brain age during general anaesthesia. (A) *Brain age and burst suppression.* Scatter plot of age and brain age of 345 patients under propofol, color-coded by the log proportion of time spent in burst suppression during the maintenance phase of stable anaesthesia, both for healthy (circle) and non-healthy (triangle) patients. The dashed grey line separates over from under-estimated age. Older patients tend to spend more time in burst suppression. Model tends to overestimate age on young patients and underestimate age on old patients, possibly due to the interventional nature of the clinical protocol. **(B)** *Brain age and age show complementary effects on burst suppression.* To formalize observations of panel C we developed a linear statistical model of the log proportion of time spent in burst suppression during maintenance using as regressors the brain age, the age, the health status, and their interaction terms, to account for the observed non-linear trends. The Healthy binary variable is 1 for ASA1 and ASA2 patients, 0 for ASA3. This interaction model was selected by a Likelihood Ratio Test from different models as showing the best model fit (see Tables 3.3 and 3.4 for the full model comparison). We depict the coefficients of all predictors with their 95% confidence intervals. Brain age is a major factor influencing burst suppression. First, burst suppression is associated with higher Brain Age across all patients. Additionally the results suggest the importance of non-linear interaction terms: the interaction between age and brain age explains additional variance, implying that as age increases, the impact of brain age further increases non-linearly. Finally, the interplay between brain age and age depended, in addition, on the health status: it takes an even bigger role for healthy patients. Other model terms, including the age and health status, have far less consistent effects.

Lariboisière sample exploration. We first explored our sample to gain insights about BA relationship to these two measures. The results are depicted in Fig. 3.5(A). As expected, older patients (both chronologically and physiologically) tend to spend more time in burst suppression, whether healthy (ASA1 or ASA2) and non-healthy (ASA3). Healthy patients show a large spectrum of BS rates, whereas most unhealthy patients have a high rate of BS. The dashed gray line delineates subjects with older

brains (above the line) from subjects with younger brain (under the line). Restricting our attention to healthy subjects, on which our brain age model has been fitted, we see that the model tends to overestimate age on young patients and underestimate on old patients. Even if the mean absolute error of the model is 7.9y, it tends to be positive for young people and negative for old people. This suggests a non-linearity that is not captured by the brain age model. We then used this fitted model to predict the age of unhealthy subjects. Surprisingly, pathologies (ASA3) seem to be associated with a younger brain. This is most probably due to a confounding effect of age. A common problem in establishing brain-behavior correlations for brain age is spurious correlations due to shared age-related variance in the brain age delta and the score of interest (here the ASA) [Smi+19]. By definition, the brain age delta is the age residual. A perfect estimator of age should be orthogonal to age. If not the case, then brain age delta would still depend on age. The relation between brain age delta and ASA should therefore be interpreted with caution. This calls for a proper deconfounding analysis. Moreover, besides the confounding effect of age, we have two main hypotheses on the source of the problem: 1/ Lariboisière ASA1 and ASA2 subjects, on which we fitted our model, should not be considered 'healthy' 2/ we deal with an interventional dataset in which the drug dosage is changed by doctors depending on age, which bias the observations of the link between age and EEG with a canceling effect: older subjects have less EEG-power but are administered less drugs which increases their power. Both hypotheses are discussed in Section 3.5. But even if we can't interpret the sign of Brain Age Delta, we can show that it is a clinically useful complement to age when predicting BS [Dad+21], for which we now develop a statistical model.

Burst Suppression modeling. To account for the observed nonlinearity while still enjoying the interpretability of linear models, we developed a linear statistical model of the log proportion of time spent in burst suppression during maintenance using as predictors the brain age, the age, and the health status. The Healthy binary variable is 1 for ASA1 and ASA2 patients (considered healthy), 0 for ASA3 (considered unhealthy). We compared five different models: the two univariate models using age and brain age, the two multivariate models (age, brain age) and (age, brain age, health status) and finally the model taking all the predictors and their interaction terms. The main statistics of the linear regression analysis for all five models are summarized in Table 3.3.

```
Model 1: scale(BS_main_frac) ~ scale(age)
Model 2: scale(BS_main_frac) ~ scale(brain_age)
Model 3: scale(BS_main_frac) ~ scale(age) + scale(brain_age)
Model 4: scale(BS_main_frac) ~ scale(age) + scale(brain_age) + healthy
Model 5: scale(BS_main_frac) ~ scale(age) * scale(brain_age) * healthy
```

	Dependent variable:				
	Burst suppression				
	(1)	(2)	(3)	(4)	(5)
Age	0.371***		0.103	0.064	0.017
	(0.065)		(0.111)	(0.113)	(0.121)
Brain Age		0.411***	0.327***	0.332***	0.398***
		(0.064)	(0.111)	(0.111)	(0.130)
Healthy				0.277	−0.011
				(0.175)	(0.225)
Age x Brain Age					0.188**
					(0.076)
Age x Healthy					0.184
					(0.299)
Brain Age x Healthy					−0.378
					(0.294)
Brain Age x Age x Healthy					0.483**
					(0.205)
Constant	0.000	−0.000	0.000	−0.049	−0.204**
	(0.065)	(0.064)	(0.064)	(0.071)	(0.090)
Observations	207	207	207	207	207
R^2	0.137	0.169	0.173	0.183	0.274
Adjusted R^2	0.133	0.165	0.165	0.171	0.248
Note:					*p<0.1; **p<0.05; ***p<0.01

Tab. 3.3: Results of linear regression analysis. We regressed the fraction of time spent in BS during maintenance with five different models of growing complexity and summarized the results.

Since the significant coefficients are not the only sign of variable importance we conducted a model comparison using a log-likelihood statistical test. The interaction model was selected by a Likelihood Ratio Test from the different models as showing the best model fit (see Table 3.4 for the full model comparison).

	Res.Df	RSS	Df	Sum of Sq	Pr(>Chi)	
1	205	177.72				
2	205	171.15	-0	6.57		
3	204	170.44	1	0.71	0.3296	
4	203	168.36	1	2.07	0.0967	
5	199	149.58	4	18.79	0.0001	***

Tab. 3.4: Variance analysis: results of the ANOVA model selection procedure. Introducing brain age instead of age in the univariate regression induces the first big improvement in model fit. Brain age alone is a better biomarker of BS than age alone. Used in concert with age, this EEG-based biomarker enriches the information given by age, showing complementary effect on BS. Brain age explains some variance of BA that is not explained by age. The model incorporating all predictors and their interaction terms is the statistically most solid model, providing the best data fit.

The coefficients of the interaction model are depicted with their 95% confidence intervals on Fig. 3.5(B). Brain Age is a major factor influencing BS as it is present on all three significant predictors: Brain Age x Age x Healthy, Brain Age and Brain Age x Age. First, BS is associated with higher Brain Age across all patients at fixed age: elevated Brain Age increases fraction of time spent in BS by 0.4 STD per STD (equivalent to 3.4% additional time spent in BS per 15.8y of brain age). When taking age into account, brain age is associated with BS in a semantically correct way. People with older brains tend to spend more time in burst suppression, possibly uncovering postoperative complications. Additionally the results suggest the importance of non-linear interaction terms: the interaction between age and brain age explains additional variance, implying that as age increases, the impact of brain age further increases non-linearly. Finally, the interplay between brain age and age depended, in addition, on the health status: it takes an even bigger role for healthy patients. All the other model terms, including the age and the health status, are not significant.

3.4.3 Drug impact on Brain age prediction during GA

Statistical tests run during Data Exploration in section 3.3 hinted that drugs may be a confounder of the effect of EEG on age, potentially affecting both. Hypnotic drugs are indeed known to modify brain spectral signatures [Pur+15b] and are significantly associated with age due a systematic intervention of the anaesthetist adapting the drug dosage to age. Therefore we have to deal with two distinct sub-populations: propofol and sevoflurane-maintained patients. Probing the impact of the anesthetic drug is all the more interesting as it has been under-explored in literature. To further investigate the effect of drug on brain age prediction we compared the performance of our Riemannian model when learning from the two different drugs (propofol and sevoflurane), using the different classes of models described in Methods section 3.2.5. Results are depicted in Fig. 3.6

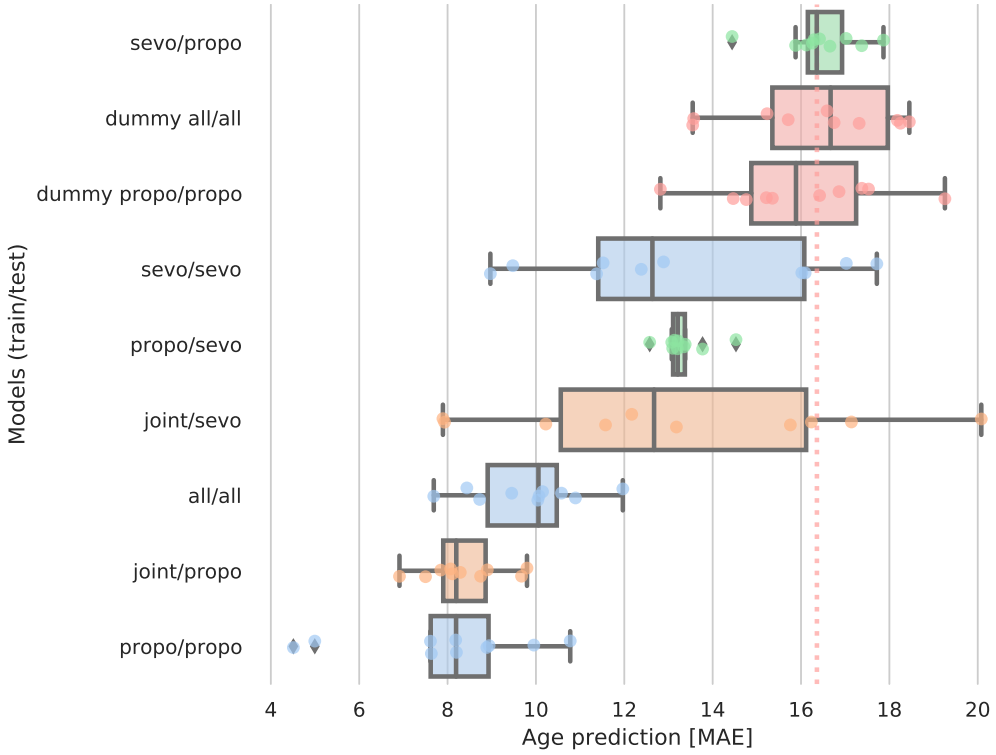


Fig. 3.6: Drug impact on brain age. Performance of the Riemannian model applied to cross-frequency covariances, when learning from two different drugs: propofol or sevoflurane. Separate models (learning and predicting on patients under the same drug) are depicted in blue, cross models (learning under one drug and predicting under the other) in green, joint models (learning an interaction model under any drug and predicting under either propofol or sevoflurane) in orange, dummy models (that do not learn) in red, with a vertical dotted line representing dummy model MAE *i.e.,* chance level. The separate models are drug-specific models and show that propofol allows to better discriminate the age from EEG-powers compared to sevoflurane. The joint models are drug-agnostic models and show at least as good performance as drug specific models. The cross models' poor performances indicate that sevoflurane must have a very different EEG spectral signature than propofol when related to age, but can still be used in principle. Therefore, propofol and sevoflurane have to be considered separately. Our joint model offers a promising framework that can still pool them, and might be revealed as the model of choice with a larger amount of data.

Drugs indeed seem to greatly influence age prediction under GA: we can't interpret EEG without drug information. The propofol-specific (propofol/propofol) model has a better performance than the drug-agnostic model (all/all) that is not informed by the drug. This drug-agnostic model as a single global linear model does not take into account that the two drugs influence EEG signals based on distinct data generating mechanisms in the two sub-populations hence does not perform optimally. This confirms that propofol and sevoflurane have to be considered separately.

The propofol-specific model is also much better than the sevoflurane-specific model (sevo/sevo), hinting that propofol allows better discriminate of age from EEG-powers compared to sevoflurane. This difference still holds when learning from a propofol population resampled to match age distribution and sample size of sevoflurane, discarding the datashift hypothesis to account for the performance difference. This result is also consistent with latest research that shows that intraoperative EEG alpha-band is a better proxy of preoperative cognitive function under propofol compared to sevoflurane [recently submitted work]. One hypothesis for this difference comes from the different action mechanism of both drugs: propofol only acts on the GABA receptors (it is a pure GABA agonist) whereas sevoflurane has several other action mechanisms (mainly GABA agonist and NMDA antagonist) [Tra+00; CMF03], with a potentially netting effect, hence a loss of age-variability. Another hypothesis for this reduced age-variability is that the dosage of sevoflurane is less variable than propofol across patients since it relies on standard abacus of MAC target values (minimum alveolar concentration), whereas propofol dosage is determined by a personalized TCI target value.

Drug-crossed models' poor performances indicate that sevoflurane must have a markedly different EEG signature than propofol that somehow hinders age prediction: generalization across drugs does not work well. The performance of these cross models could also be driven by age instead of drug, knowing that patients maintained under sevoflurane are often older than under propofol, yet with a smaller STD.

The drug-aware model (joint/propofol) shows at least as good performance as propofol-specific model when predicting on propofol subjects. Its observed reduced variance could show a more refined prediction although we can't rule out the statistical effect of a larger training sample size. This joint model therefore offers a promising framework that can allow pooling of patients, and may reveal itself as the model of choice when learning from larger amounts of data, leveraging what the two drugs have in common.

3.5 Discussion & future work

Our work represents the most extensive validation of a ML approach to estimate BA during GA. We presented a robust end-to-end biomarker learning strategy for EEG during anaesthesia. We demonstrated how to achieve state-of-the-art performance for EEG-based Brain age prediction during anaesthesia. We showed that the drug critically impacts BA prediction under GA and analyzed its impact through interaction learning.

We validated a potential EEG-based brain age measure against burst suppression and ASA clinical score: higher brain age is correlated with more burst suppression, whereas age has a far less consistent effect. Hence we provided evidence that BA captures patient propensity to develop BS assuming a stable and adequate GA depth (SEF95 in the range of 8–13 Hz). We showed that EEG in the OR, today only used for monitoring depth of anaesthesia (via BIS and PSI indices), could be exploited to estimate a personalized physiological age of the brain of an anaesthetized subject that can help detect a propensity to the anomaly of BS, a recognized marker of the risk of developing cognitive dysfunction within the postoperative period. One possible hypothesis consists in considering that the peroperative rate of suppression patterns and the cognitive trajectory of postoperative patients could be epiphenomena of the same symptom of an elevated brain age (and not chronological age), that can in principle be estimated before surgery. If this hypothesis is confirmed, this could guide the therapeutic intervention in the operating room, but also would in principle allow to develop preventive procedures and help to improve postoperative medical care *e.g.,* by early referring patient to a neurologist. A few open questions are nevertheless worth discussing.

One of the limitations of our study is methodological: the observation period of the features is defined using the SEF95 index, following a definition that is largely agreed in the literature. But since the SEF95 is itself related to the EEG power features we could lose some variability that could be useful to predict. One possible solution would be to learn from the data itself the regions in the data that should be trusted and considered. Our algorithm could for instance select the best interval of SEF95 in a nested cross-validation fashion and potentially discover that the [8-13]Hz interval doctors are using may not be the optimal for building biomarkers. Also, this study was restricted to one hospital. Results obtained should be replicated using data coming from other hospitals with different devices and clinical protocols.

Coming back to the unexpected experimental finding of Figure 3.5 where older and unhealthy patients seem to have a brain looking younger than their chronological age, we already discussed the potential confounding effect of age (age-related variance in both the brain age delta and BS). Besides, we have two main hypotheses

on the source of the problem. The first most obvious hypothesis is that Lariboisière subjects should not be considered 'healthy': this is a suffering population as they came to the hospital to undergo surgery. To test this hypothesis, one could use the Riemann regression model trained on TUH healthy patients to predict age (restricted to four frontal electrodes), and test if indeed older brains are more frequent in the non-healthy population, in effect sorting the ASA score in Lariboisière population in a coherent manner. Also the large number of subjects in the TUH dataset would address the sample bias we could have suffered from learning from the small sample size of Lariboisière data. The second hypothesis relates to the causal structure of the data. According to this hypothesis, we don't observe the natural relation between resting state EEG and age because there's an intervention, in the form of a treatment (the drug dosage) that biases it. Data have indeed not been collected in a controlled environment: most fragile patients have been treated, here induced, differently precisely to avoid BS. We have been confirmed that doctors changed the drug dosage according to their belief of age/health status, which then controls the EEG: most fragile patients have been administered less drugs, which increases the relative alpha power hence making them appear younger. If this is true, our age prediction may not be interpreted as brain age anymore and the relation to BS is not easily interpretable. Other interventions occurred: some patients received bolus of ketamine that is known to boost alpha and beta power for $\sim$ 20/30min, and some patients maintained with sevoflurane received bolus of propofol if necessary. Of course more data with the same data generating mechanism (the doctors) won't help this causality problem. One possible solution would be to conduct deconfounding analysis by estimating the intervention effect with the propofol target dosage (TCI). This way the treatment effect could be compensated and treatment-independent conclusions reached. Such a study could be feasible in the near future: Lariboisière hospital is currently testing a commercial solution for automated collection of hemodynamic and brain signals, synchronized with the timing and volume of administered drugs. Finally, one alternative to this purely statistical analysis, yet more demanding, would consist in doing causal inference from carefully designed probabilistic graphical models.

Regarding the performance of our brain age model, it is noteworthy that we obtained with 4-channel clinical grade EEGs of 345 subjects the same performance in age prediction (8y) than with 102-channels research-grade MEGs of 643 subjects (Cam-CAN) or 21-channels EEGs of 1000 subjects (TUH), for a comparable chance level. This result could come from several reasons but we can't rule out a positive impact of GA. According to this hypothesis, GA would be seen as a physiological stress test and the hypnotic drug would enhance the brain response, forcing it to 'speak'. With $\sim$ $10\,\mathrm{m}$ GA per year in France, the wealth of EEG recorded during GA has the potential to allow medical discoveries if taking into account the specificities of GA. In this thesis we established that Brain age under GA can be a serious candidate to be a

drug-agnostic biomarker of BS, potentially integrated in a GA monitoring tool in the future. Further studies could investigate its relation to medium and long-term outcome of patients, *e.g.,* the development of neurodegenerative disorders.

Finally, having the opportunity to work on a raw clinical dataset was a real chance to glimpse over the real-world challenges of applying AI to medical data. I learnt a few lessons along the way among which to discuss with the medical doctors and look at the data the earlier possible to assess the data collection procedure. The time needed to clean and preprocess the data should also not be underestimated, which again is very common in real-world ML projects. Exploiting data in Healthcare is not an easy journey but can be very rewarding. We finally contributed to build something unique since there's no public or even private dataset of EEGs during GA, and hopefully something useful with potential medical application on health. This study could trigger a new line of research moving forward, developing clinically-relevant biomarkers from the GA period, with potential for medical applications as a diagnostic tool, paving the way for a finer understanding of brain diseases, and a more targeted approach to medical treatment.

Conclusion

Contents

Future directions

In this thesis we established that Brain age under GA could be a candidate biomarker of complications during general anesthesia. Yet, our brain age model would still benefit from a few technical refinements.

First it would need to be further robustified. To provide a robust end-to-end biomarker learning strategy for EEG during anaesthesia, we should allow our algorithm to learn from the data itself the regions of the EEG recording that should be trusted and considered, potentially challenging the mostly acceptable definition of a SEF95 index in the [8-13]Hz range. To overcome limited spatial resolution of clinical EEG it would also benefit from capturing temporal information of EEG and complement EEG with other signals monitored during GA. To demonstrate its generalization capacity we should also probe it using data coming from other hospitals with difference devices and clinical protocols.

Then, we should disentangle the causal factors shaping GA observations, *e.g.,* the effect of drug dosage on the link between health and EEG. One possible solution is to conduct a deconfounding analysis by estimating the intervention effect with the Propofol target dosage and then compensate for it. This study will be feasible after the expected future technical upgrade of Lariboisière's data collection system, which would allow the automated and time-synchronized collection of brain signals and volume of administrated drugs. One more demanding alternative would consist in doing a proper causal inference from a carefully designed probabilistic graphical models.

Moving forward, we are convinced that this study could trigger a new line of research, developing clinically-relevant biomarkers from the GA period, with potential for

medical applications as a diagnostic tool. Nevertheless, we observe that most machine learning models trained to make medical decisions that perform at nearly the same level as human experts are not in clinical use. Indeed collecting data from one hospital, train and test the model on data from the same hospital, and showing the algorithms are comparable to human doctors in spotting certain conditions is enough to publish a research paper. But it often turns out that the same model used in a different hospital, applied to data collected with a different device, and a slightly different protocol used by the technician will show a significantly degraded performance. In contrast, the performance of any human doctor would stay the same. Indeed, there are challenges in translating a research paper into something useful in a clinical setting: these models still need a lot of work to reach production. This proof-of-concept-to-production gap between research and practice is not unique to medicine but exists throughout the machine learning world. Modeling is just one step towards production: finding the right data, deploying the model, monitoring it and showing safety are among other necessary steps.

If a proper randomized clinical study demonstrates that BA under GA is a valid biomarker for preoperative health it could be integrated into a GA monitoring tool to guide anesthetists in their perioperative decisions. If further studies show it is related to medium/long term outcome of patients, *e.g.,* development of neurodegenerative disorders, BA could even be part of a 'medical consultation' under GA: patients would receive early feedback on brain age and risk of developing neurodegenerative diseases along with advices to improve brain health. These results would pave the way for a finer understanding of brain diseases, and a more targeted approach to medical treatment. This could set new standards in biomedical research, releasing GA-based data from the operating room into neurological and cardiological consulting, in the long run of potential benefit to millions of patients

From scientific to societal impact

This thesis builds on the unique combination of expertise and associated preliminary data obtained within the APHP and Inria teams, which have been engaged in a close collaboration for over five years. Leveraging my 25y+ industry experience, my post-thesis objective is to transfer these present and future scientific findings into the socio-economic world by creating a medtech start-up company, which will lead the transformation of this project results into a product on the market.

More precisely, this startup project would aim to reduce the risk of perioperative complications due to sub-optimal anesthesia management by providing an optimal knowledge of patient's physiological state and by contributing to a more 'intelligent' administration of anesthesia drugs ('the right dose at the right time'). To that purpose, it will develop a combined alarm & decision support system – or virtual

assistant – based on a predictive digital twin that integrates all information available on a patient under anesthesia in a unified form, including the brain. The system will combine biophysical models and statistical models (our Brain Age) using the patient's physiological signals for optimized use of monitoring data. This augmented cardiovascular and cerebral monitoring will provide a global vision of the patient's condition and will help guide physicians' therapeutic choices by recommending personalized medical strategies. By doing so, we aim at significantly improve patients conditions and quality of life after surgery. Additionally the startup will propose a post-surgery Personalized Anesthesia Report that describes patients' reaction to anesthetic and surgical challenges and summarize cardiovascular and cerebral status (*e.g.,* the Brain Age of the patient). This will help anesthesia management to pass the door of the operating room and propagate information to the following health care providers.

This project will benefit to all patients who will get surgery under general anesthesia. The greater benefits will be for patients labelled as high-risk surgical patients contraindicated for surgery, namely patients older than 65 and/or with altered cardiovascular function, and/or brain frailty since they are more prone to suffer from complications. By using our virtual assistant, hospitals may offer novel opportunities to many patients previously excluded from surgery. Reducing peri-operative complications immediately translates into reducing hospital length of stay and post-hospital discharge costs, in addition to improving patient's quality of care. This approach goes beyond the state of the art in anesthesia monitoring and opens new fields in the patient pathway at the hospital.

To finance this entrepreneurial endeavour I participated to the creation of a consortium of academic (Inria), clinical (AP-HP Lariboisière) and industrial (Philips Healthcare) partners and answered two main Call for Proposals: European EIT Health 2022 and French RHU 2021 (Recherche Hospitalo Universitaire en santé). Our proposal and common aspiration is to leverage mathematical and AI new techniques to guide the medical doctors in their daily practice, with the hope to ultimately give scientific discoveries an additional social and economic impact on the world.

Synthèse en français

Les maladies neurodégénératives figurent parmi les principales causes de mortalité dans le monde. Malheureusement, leur diagnostic précoce nécessite un examen médical prescrit souvent trop tardivement et des équipements de laboratoire dédiés. Il repose aussi fréquemment sur des mesures prédictives souffrant d'un biais de sélection. Cette thèse présente une solution prometteuse à ces problèmes : une méthode robuste, directement utilisable en clinique, pour construire des biomar-

queurs prédictifs à partir des signaux cérébraux M/EEG, validés contre les troubles neurocognitifs apparaissant après une anesthésie générale.

Dans une première contribution (théorique) [Sab+19], nous avons évalué des modèles de régression capables d'apprendre des biomarqueurs à partir des matrices de covariance de signaux M/EEG. Notre analyse mathématique a identifié différents modèles garantissant une prédiction parfaite dans des circonstances idéales, lorsque la cible est une fonction (log-)linéaire en la puissance des sources cérébrales. Ces modèles, basés sur les approches mathématiques de filtrage spatial supervisé et de géométrie riemannienne, permettent une prédiction optimale sans nécessiter une coûteuse localisation des sources. Nos simulations confirment cette analyse mathématique et suggèrent que ces algorithmes de régression sont robustes à travers les mécanismes de génération de données et les violations de modèles. Cette étude suggère que les méthodes riemanniennes sont des méthodes de choix pour l'analyse automatisée à grande échelle des données M/EEG en l'absence d'IRM, condition importante pour pouvoir développer des biomarqueurs cliniques.

Dans une deuxième contribution (empirique) [Sab+20], nous avons validé nos modèles prédictifs sur plusieurs ensembles de données de neuro-imagerie et avons montré qu'ils peuvent être utilisé pour apprendre l'âge du cerveau à partir de signaux cérébraux M/EEG, sans localisation de sources, et avec un prétraitement minimal des données. De plus, la performance de notre méthode riemannienne est proche de celle des méthodes de référence nécessitant une localisation de sources et donc un traitement manuel des données, la disponibilité d'images IRM anatomiques et une expertise en modélisation de sources M/EEG. Une analyse empirique à grande échelle a ensuite permis de démontrer que l'âge du cerveau dérivé de la MEG capture des informations uniques liées à l'activité neuronale et non expliquées par l'IRM anatomique. Conformément aux simulations, ces résultats suggèrent également que l'approche riemannienne est une méthode pouvant s'appliquer dans un large éventail de situations, avec une robustesse considérable aux différents choix de prétraitement, y compris minimaliste. Les bonnes performances obtenues avec la MEG ont ensuite été répliquées avec des EEGs de qualité recherche.

Dans une troisième contribution (clinique) [Sab+21, en préparation], nous avons validé le concept d'âge cérébral directement au bloc opératoire de l'hôpital Lariboisière à Paris, à partir d'EEG de qualité clinique recueillis pendant la période de l'anesthésie générale. Nous avons évalué notre mesure de l'âge cérébral comme prédicteur de complications peropératoires liées aux dysfonctions cognitives post opération, validant ainsi l'âge du cerveau comme un biomarqueur clinique prometteur des troubles neurocognitifs. Nous avons également montré que le sédatif utilisé a un impact important sur la prédiction de l'âge du cerveau et avons démontré la robustesse de notre approche à différents types de médicaments.

Combinant des concepts précédemment étudiés séparément, notre contribution démontre la pertinence clinique de la notion d'âge du cerveau prédit à partir de l'EEG pour révéler les pathologies des fonctions cérébrales dans des situations où l'IRM ne peut pas être réalisée. Ces résultats fournissent également une première preuve que l'anesthésie générale est une période propice à la découverte de biomarqueurs cérébraux, avec un impact potentiel profond sur la médecine préventive et une influence sociale et économique durable.

Bibliography

[Aa2] *Alzheimer's Disease Facts and Figures*. Tech. rep. Alzheimer's & Dementia, 2021:17(3). Alzheimer's Association 2021 (cit. on p. 2).

[AG01] LI Aftanas and SA Golocheikine. "Human anterior and frontal midline theta and lower alpha reflect emotionally positive state and internalized attention: high-resolution EEG investigation of meditation". In: *Neuroscience letters* 310.1 (2001), pp. 57–60 (cit. on p. 15).

[Ahl+10] Seppo P Ahlfors, Jooman Han, John W Belliveau, and Matti S Hämäläinen. "Sensitivity of MEG and EEG to source orientation". In: *Brain topography* 23.3 (2010), pp. 227–232 (cit. on p. 11).

[AJWW66] HW Agnew Jr, Wilse B Webb, and Robert L Williams. "The first night effect: an EEG study of sleep". In: *Psychophysiology* 2.3 (1966), pp. 263–266 (cit. on p. 10).

[AMS09] P-A Absil, Robert Mahony, and Rodolphe Sepulchre. *Optimization algorithms on matrix manifolds*. Princeton University Press, 2009 (cit. on pp. 50, 51).

[And+15] Lau M Andersen, Michael N Pedersen, Kristian Sandberg, and Morten Overgaard. "Occipital MEG activity in the early time range (< 300 ms) predicts graded changes in perceptual consciousness". In: *Cerebral Cortex* 26.6 (2015), pp. 2677–2688 (cit. on p. 63).

[App+19] Stefan Appelhoff, Matthew Sanderson, Teon L Brooks, et al. "MNE-BIDS: Organizing electrophysiological data into the BIDS format and facilitating their analysis". In: *The Journal of Open Source Software* 4.44 (2019) (cit. on p. 117).

[Ast+08] L. Astolfi, F. Cincotti, D. Mattia, et al. "Tracking the time-varying cortical connectivity patterns by adaptive multivariate estimators". In: *IEEE Transactions on Biomedical Engineering* 55.3 (2008), pp. 902–913 (cit. on p. 43).

[Att+19] Zachi I. Attia, Paul A. Friedman, Peter A. Noseworthy, et al. "Age and Sex Estimation Using Artificial Intelligence From Standard 12-Lead ECGs". In: *Circulation: Arrhythmia and Electrophysiology* 12.9 (2019), e007284. eprint: https://www.ahajournals.org/doi/pdf/10.1161/CIRCEP.119.007284 (cit. on p. 105).

[AUT19] JJ Allaire, Kevin Ushey, and Yuan Tang. *reticulate: Interface to 'Python'*. R package version 1.11. 2019 (cit. on pp. 25, 84).

[AYH18] David B Arciniegas, Stuart C Yudofsky, and Robert E Hales. *The American Psychiatric Publishing Textbook of Neuropsychiatry and Behavioral Neuroscience*. American Psychiatric Pub, 2018 (cit. on p. 15).

[Bab+06] Claudio Babiloni, Giuliano Binetti, Andrea Cassarino, et al. “Sources of cortical rhythms in adults during physiological aging: a multicentric EEG study”. In: *Human brain mapping* 27.2 (2006), pp. 162–172 (cit. on p. 21).

[Bai17] Sylvain Baillet. “Magnetoencephalography for brain electrophysiology and imaging”. In: *Nature Neuroscience* 20 (Feb. 2017), 327 EP – (cit. on pp. 61, 62).

[Bak+75] J Dennis Baker, Bernd Gluecklich, C Wesley Watson, et al. “An evaluation of electroencephalographic monitoring for carotid study”. In: *Surgery* 78.6 (1975), pp. 787–794 (cit. on p. 10).

[Bar+06] William E Barlow, Emily White, Rachel Ballard-Barbash, et al. “Prospective breast cancer risk prediction model for women undergoing screening mammography”. In: *Journal of the National Cancer Institute* 98.17 (2006), pp. 1204–1214 (cit. on p. 32).

[Bar+11] Alexandre Barachant, Stéphane Bonnet, Marco Congedo, and Christian Jutten. “Multiclass brain–computer interface classification by Riemannian geometry”. In: *IEEE Transactions on Biomedical Engineering* 59.4 (2011), pp. 920–928 (cit. on p. 64).

[Bar+12] A. Barachant, S. Bonnet, M. Congedo, and C. Jutten. “Multiclass Brain–Computer Interface Classification by Riemannian Geometry”. In: *IEEE Transactions on Biomedical Engineering* 59.4 (2012), pp. 920–928 (cit. on pp. 49, 53, 64, 85).

[Bar+13] Alexandre Barachant, Stéphane Bonnet, Marco Congedo, and Christian Jutten. “Classification of covariance matrices using a Riemannian-based kernel for BCI applications”. In: *Neurocomputing* 112 (2013), pp. 172–178 (cit. on p. 64).

[BD04] György Buzsáki and Andreas Draguhn. “Neuronal oscillations in cortical networks”. In: *science* 304.5679 (2004), pp. 1926–1929 (cit. on p. 15).

[Bel+97] Adel Belouchrani, Karim Abed-Meraim, J-F Cardoso, and Eric Moulines. “A blind source separation technique using second-order statistics”. In: *IEEE Transactions on signal processing* 45.2 (1997), pp. 434–444 (cit. on p. 57).

[Ber29] Hans Berger. “Über das elektroenkephalogramm des menschen”. In: *Archiv für psychiatrie und nervenkrankheiten* 87.1 (1929), pp. 527–570 (cit. on p. 10).

[Bha09] Rajendra Bhatia. *Positive definite matrices*. Princeton university press, 2009 (cit. on p. 52).

[BHT21] Stephen Bates, Trevor Hastie, and Robert Tibshirani. *Cross-validation: what does it estimate and how well does it do it?* 2021. arXiv: 2104.00673 [stat.ME] (cit. on p. 38).

[BI+16] Charles H Brown IV, Laura Max, Andrew LaFlam, et al. “The association between preoperative frailty and postoperative delirium after cardiac surgery”. In: *Anesthesia and analgesia* 123.2 (2016), p. 430 (cit. on p. 129).

[BI19] Danilo Bzdok and John PA Ioannidis. “Exploration, inference, and prediction in neuroscience and biomedicine”. In: *Trends in neurosciences* (2019) (cit. on pp. 14, 107).

[Bis07] Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer, 2007 (cit. on p. 38).

[BJF10] Luc Berthouze, Leon M James, and Simon F Farmer. “Human EEG shows long-range temporal correlations of oscillation amplitude in Theta, Alpha and Beta bands across a wide age range”. In: *Clinical Neurophysiology* 121.8 (2010), pp. 1187–1197 (cit. on p. 89).

[BJL18] Rajendra Bhatia, Tanvi Jain, and Yongdo Lim. “On the Bures–Wasserstein distance between positive definite matrices”. In: *Expositiones Mathematicae* (2018) (cit. on p. 54).

[BL17] György Buzsáki and Rodolfo Llinás. “Space and time in the brain”. In: *Science* 358.6362 (2017), pp. 482–485. eprint: https://science.sciencemag.org/content/358/6362/482.full.pdf (cit. on p. 15).

[Bla+08] B. Blankertz, R. Tomioka, S. Lemm, M. Kawanabe, and K. Muller. “Optimizing Spatial filters for Robust EEG Single-Trial Analysis”. In: *IEEE Signal Processing Magazine* 25.1 (2008), pp. 41–56 (cit. on pp. 60, 64).

[BLS10] Emery N Brown, Ralph Lydic, and Nicholas D Schiff. “General anesthesia, sleep, and coma”. In: *New England Journal of Medicine* 363.27 (2010), pp. 2638–2650 (cit. on p. 113).

[BM14] György Buzsáki and Kenji Mizuseki. “The log-dynamic brain: how skewed distributions affect network operations”. In: *Nature Reviews Neuroscience* 15.4 (2014), p. 264 (cit. on pp. 60, 64, 103).

[Bou+14] N. Boumal, B. Mishra, P.-A. Absil, and R. Sepulchre. “Manopt, a Matlab Toolbox for Optimization on Manifolds”. In: *Journal of Machine Learning Research* 15 (2014), pp. 1455–1459 (cit. on p. 30).

[Bro+11] Matthew J Brookes, Mark Woolrich, Henry Luckhoo, et al. “Investigating the electrophysiological basis of resting state networks using magnetoencephalography”. In: *Proceedings of the National Academy of Sciences* 108.40 (2011), pp. 16783–16788 (cit. on pp. 19, 63).

[Bro+17] M. M. Bronstein, J. Bruna, Y. LeCun, A. Szlam, and P. Vandergheynst. “Geometric Deep Learning: Going beyond Euclidean data”. In: *IEEE Signal Processing Magazine* 34.4 (2017), pp. 18–42 (cit. on p. 29).

[Bru+03] Jörgen Bruhn, Thomas W Bouillon, Lucian Radulescu, et al. “Correlation of approximate entropy, bispectral index, and spectral edge frequency 95 (SEF95) with clinical signs of “anesthetic depth” during coadministration of propofol and remifentanil”. In: *The Journal of the American Society of Anesthesiologists* 98.3 (2003), pp. 621–627 (cit. on p. 113).

[BS09] Silvere Bonnabel and Rodolphe Sepulchre. “Riemannian metric and geometric mean for positive semidefinite matrices of fixed rank”. In: *SIAM Journal on Matrix Analysis and Applications* 31.3 (2009), pp. 1055–1070 (cit. on p. 54).

[BV00] J. . Bercher and C. Vignat. “Estimating the entropy of a signal with applications”. In: *IEEE Transactions on Signal Processing* 48.6 (2000), pp. 1687–1694 (cit. on p. 41).

[Bzd17] Danilo Bzdok. “Classical Statistics and Statistical Learning in Imaging Neuroscience”. In: *Frontiers in Neuroscience* 11.OCT (2017), pp. 1–23 (cit. on p. 15).

[Bzd+18] Danilo Bzdok, Denis Engemann, Olivier Grisel, Gaël Varoquaux, and Bertrand Thirion. “Prediction and inference diverge in biomedicine: Simulations and real-world data”. In: (2018) (cit. on p. 107).

[Car+19] Jérôme Cartailler, Pierre Parutto, Cyril Touchard, Fabrice Vallée, and David Holcman. “Alpha rhythm collapse predicts iso-electric suppressions during anesthesia”. In: *Communications biology* 2.1 (2019), pp. 1–10 (cit. on pp. 114, 129).

[CBA13] M. Congedo, A. Barachant, and A. Andreev. “A New Generation of Brain-Computer Interface Based on Riemannian Geometry”. In: *arXiv e-prints* (Oct. 2013). arXiv: 1310.8115 [cs.HC] (cit. on pp. 25, 84).

[CBB17] Marco Congedo, Alexandre Barachant, and Rajendra Bhatia. “Riemannian geometry for EEG-based brain-computer interfaces; a primer and a review”. In: *Brain-Computer Interfaces* 4.3 (2017), pp. 155–174 (cit. on p. 29).

[Cha+13] Frédéric Chazal, Leonidas J Guibas, Steve Y Oudot, and Primoz Skraba. “Persistence-based clustering in Riemannian manifolds”. In: *Journal of the ACM (JACM)* 60.6 (2013), pp. 1–38 (cit. on p. 44).

[Che+10] Yilun Chen, Ami Wiesel, Yonina C Eldar, and Alfred O Hero. “Shrinkage algorithms for MMSE covariance estimation”. In: *IEEE Transactions on Signal Processing* 58.10 (2010), pp. 5016–5029 (cit. on pp. 46, 86, 89, 102, 117).

[Cla+04] C Richard Clark, Melinda D Veltmeyer, Rebecca J Hamilton, et al. “Spontaneous alpha peak frequency predicts working memory performance across the age span”. In: *International Journal of Psychophysiology* 53.1 (2004), pp. 1–9 (cit. on p. 89).

[CMF03] Jason A Campagna, Keith W Miller, and Stuart A Forman. “Mechanisms of actions of inhaled anesthetics”. In: *New England Journal of Medicine* 348.21 (2003), pp. 2110–2124 (cit. on p. 135).

[Col+17] James H Cole, Rudra PK Poudel, Dimosthenis Tsagkrasoulis, et al. “Predicting brain age with deep learning from raw imaging data results in a reliable and heritable biomarker”. In: *NeuroImage* 163 (2017), pp. 115–124 (cit. on p. 110).

[Col+18] James H Cole, Stuart J Ritchie, Mark E Bastin, et al. “Brain age predicts mortality”. In: *Molecular psychiatry* 23.5 (2018), p. 1385 (cit. on pp. 18, 110).

[Col+19] James H Cole, Riccardo E Marioni, Sarah E Harris, and Ian J Deary. “Brain age and other bodily ‘ages’: implications for neuropsychiatry”. In: *Molecular psychiatry* 24.2 (2019), pp. 266–281 (cit. on p. 110).

[Cox+19] Simon R Cox, Donald M Lyall, Stuart J Ritchie, et al. “Associations between vascular risk factors and brain MRI indices in UK Biobank”. In: *European heart journal* 40.28 (2019), pp. 2290–2300 (cit. on p. 16).

[CP14] Alain de Cheveigné and Lucas C. Parra. “Joint decorrelation, a versatile tool for multichannel data analysis”. In: *NeuroImage* 98 (2014), pp. 487 –505 (cit. on pp. 64, 72, 102).

[CPB16] M. Congedo, R. Phlypo, and A. Barachant. “A fixed-point algorithm for estimating power means of positive definite matrices”. In: *2016 24th European Signal Processing Conference (EUSIPCO)*. 2016, pp. 2106–2110 (cit. on p. 53).

[CR+07] Delphine Cosandier-Rimélé, Jean-Michel Badier, Patrick Chauvel, and Fabrice Wendling. “A physiologically plausible spatio-temporal model for EEG signals recorded with intracerebral electrodes in human partial epilepsy”. In: *IEEE Transactions on Biomedical Engineering* 54.3 (2007), pp. 380–388 (cit. on p. 17).

[CR95] Michael GH Coles and Michael D Rugg. *Event-related brain potentials: An introduction*. Oxford University Press, 1995 (cit. on p. 14).

[CRJ19] M. Congedo, P. L. C. Rodrigues, and C. Jutten. “The Riemannian Minimum Distance to Means Field Classifier”. In: *Graz BCI Conference 2019*. 2019 (cit. on p. 44).

[Csj] *Blood flow, blood pressure or both?* Tech. rep. vol. 9, no. 6, pp. 15–18, May 2018. Clinical services journal (cit. on pp. 22, 128).

[CSM17] Erik Clarke and Scott Sherrill-Mix. *ggbeeswarm: Categorical Scatter (Violin Point) Plots*. R package version 0.6.0. 2017 (cit. on pp. 25, 84).

[Cul+17] Deborah J Culley, Devon Flaherty, Margaret C Fahey, et al. “Poor performance on a preoperative cognitive screening test predicts postoperative complications in older orthopedic surgical patients”. In: *Anesthesiology* 127.5 (2017), pp. 765–774 (cit. on p. 129).

[DA01] Peter Dayan and Laurence F Abbott. *Theoretical neuroscience: computational and mathematical modeling of neural systems*. Computational Neuroscience Series, 2001 (cit. on pp. 4, 8).

[Dad+15] Christophe Dadure, Anaïs Marie, Fabienne Seguret, and Xavier Capdevila. “One year of anaesthesia in France: A comprehensive survey based on the national medical information (PMSI) database. Part 1: In-hospital patients”. In: *Anaesthesia Critical Care & Pain Medicine* 34.4 (2015), pp. 191–197 (cit. on pp. 22, 128).

[Dad+19] Kamalaker Dadi, Mehdi Rahim, Alexandre Abraham, et al. “Benchmarking functional connectome-based predictive models for resting-state fMRI”. In: *NeuroImage* 192 (2019), pp. 115–134 (cit. on pp. 15, 63, 69).

[Dad+21] Kamalaker Dadi, Gaël Varoquaux, Josselin Houenou, et al. “Population modeling with machine learning can enhance measures of mental health”. In: *GigaScience* 10 (2021), pp. 1–16 (cit. on pp. 111, 131).

[Däh+13] Sven Dähne, Felix Bießmann, Frank C Meinecke, et al. “Integration of multivariate data streams with bandpower signals”. In: *IEEE Transactions on Multimedia* 15.5 (2013), pp. 1001–1013 (cit. on p. 64).

[Däh+14a] Sven Dähne, Frank C Meinecke, Stefan Haufe, et al. “SPoC: a novel framework for relating the amplitude of neuronal oscillations to behaviorally relevant parameters”. In: *NeuroImage* 86 (2014), pp. 111–122 (cit. on pp. 60, 64, 72, 73, 102).

[Däh+14b] Sven Dähne, Vadim V Nikulin, David Ramírez, et al. “Finding brain oscillations with power dependencies in neuroimaging data”. In: *NeuroImage* 96 (2014), pp. 334–348 (cit. on pp. 64, 107).

[DCS99] Alain Destexhe, Diego Contreras, and Mircea Steriade. “Spatiotemporal analysis of local field potentials and unit discharges in cat cerebral cortex during natural wake and sleep states”. In: *Journal of Neuroscience* 19.11 (1999), pp. 4595–4608 (cit. on p. 11).

[Del+12] Arnaud Delorme, Jason Palmer, Julie Onton, Robert Oostenveld, and Scott Makeig. “Independent EEG sources are dipolar”. In: *PloS one* 7.2 (2012), e30135 (cit. on p. 63).

[Den+21] Stijn Denissen, Denis Alexander Engemann, Alexander De Cock, et al. “Brain age as a surrogate marker for information processing speed in multiple sclerosis”. In: *medRxiv* (2021) (cit. on p. 111).

[Dmo+12] Jacek Dmochowski, Paul Sajda, Joao Dias, and Lucas Parra. “Correlated Components of Ongoing EEG Point to Emotionally Laden Attention – A Possible Marker of Engagement?” In: *Frontiers in Human Neuroscience* 6 (2012), p. 112 (cit. on p. 64).

[Dos+10] Nico UF Dosenbach, Binyam Nardos, Alexander L Cohen, et al. “Prediction of individual brain maturity using fMRI”. In: *Science* 329.5997 (2010), pp. 1358–1361 (cit. on p. 17).

[DPm99] Roberto Domingo Pascual-marqui. “Review of Methods for Solving the EEG Inverse Problem”. In: *Int. J. Biomagn.* 1 (Oct. 1999) (cit. on p. 12).

[EG15] Denis A Engemann and Alexandre Gramfort. “Automated model selection in covariance estimation and spatial whitening of MEG and EEG signals”. In: *NeuroImage* 108 (2015), pp. 328–342 (cit. on pp. 46, 86, 102).

[Eng+18] Denis A Engemann, Federico Raimondo, Jean-Rémi King, et al. “Robust EEG-based cross-site and cross-protocol classification of states of consciousness”. In: *Brain* 141.11 (Oct. 2018), pp. 3179–3192. eprint: http://oup.prod.sis.lan/brain/article-pdf/141/11/3179/26172804/awy251.pdf (cit. on pp. 15, 16, 21).

[Eng+20] Denis Alexander Engemann, Oleh Kozynets, David Sabbagh, et al. “Combining magnetoencephalography with magnetic resonance imaging enhances learning of surrogate-biomarkers”. In: *eLife* 9 (2020), e54055 (cit. on pp. 19–21, 25, 60, 98, 121, 127).

[Eng+21] Denis A Engemann, Apolline Mellot, Richard Hoechenberger, et al. “A reusable benchmark of brain-age prediction from M/EEG resting-state signals”. In: *bioRxiv* (2021) (cit. on pp. 25, 62).

[Eura] Eurostat. *[hlth_cd_acdr2] database* (cit. on pp. 2, 110).

[Eurb] Eurostat. *[proj_18ndbi] database* (cit. on pp. 2, 128).

[Eurc] Eurostat. *[proj_18np] database* (cit. on pp. 2, 128).

[Fac06] Brain Facts. “a Primer on the Brain and Nervous System”. In: *Washington, DC: Society for Neuroscience* (2006) (cit. on pp. 4, 5, 8).

[Fis12] Bruce Fischl. “FreeSurfer”. In: *NeuroImage* 62.2 (2012), pp. 774–781 (cit. on p. 91).

[FM03] Wolfgang Förstner and Boudewijn Moonen. “A metric for covariance matrices”. In: *Geodesy-The Challenge of the 3rd Millennium*. Springer, 2003, pp. 299–309 (cit. on pp. 51, 64).

[Fra+12] Katja Franke, Eileen Luders, Arne May, Marko Wilke, and Christian Gaser. “Brain maturation: predicting individual BrainAGE in children and adolescents using structural MRI”. In: *Neuroimage* 63.3 (2012), pp. 1305–1312 (cit. on p. 110).

[Fri+16] Bradley A Fritz, Philip L Kalarickal, Hannah R Maybrier, et al. “Intraoperative electroencephalogram suppression predicts postoperative delirium”. In: *Anesthesia and analgesia* 122.1 (2016), p. 234 (cit. on pp. 111, 129).

[Fri+20] Bradley A Fritz, Christopher R King, Arbi Ben Abdallah, et al. “Preoperative cognitive abnormality, intraoperative electroencephalogram suppression, and postoperative delirium: a mediation analysis”. In: *Anesthesiology* 132.6 (2020), pp. 1458–1468 (cit. on pp. 110, 128).

[Fru+17] Wolfgang Fruehwirt, Matthias Gerstgrasser, Pengfei Zhang, et al. “Riemannian tangent space mapping and elastic net regularization for cost-effective EEG markers of brain atrophy in Alzheimer’s disease”. In: *arXiv preprint arXiv:1711.08359* (2017) (cit. on pp. 14, 21, 64).

[Fuk90] Keinosuke Fukunaga. “Chapter 2 - RANDOM VECTORS AND THEIR PROPERTIES”. In: *Introduction to Statistical Pattern Recognition (Second Edition)*. Ed. by Keinosuke Fukunaga. Second Edition. Boston: Academic Press, 1990, pp. 11 –50 (cit. on p. 74).

[Gar+17] Pilar Garcés, David López-Sanz, Fernando Maestú, and Ernesto Pereda. “Choice of magnetometers and gradiometers after signal space separation”. In: *Sensors* 17.12 (2017), p. 2926 (cit. on pp. 88, 105).

[Gau+19] Sinead Gaubert, Federico Raimondo, Marion Houot, et al. “EEG evidence of compensatory mechanisms in preclinical Alzheimer?s disease”. In: *Brain* 142.7 (2019), pp. 2096–2112 (cit. on pp. 16, 19, 21).

[GB08] M. Grosse-Wentrup* and M. Buss. “Multiclass Common Spatial Patterns and Information Theoretic Feature Extraction”. In: *IEEE Transactions on Biomedical Engineering* 55.8 (2008), pp. 1991–2000 (cit. on p. 60).

[GBD09] Amir A Ghaferi, John D Birkmeyer, and Justin B Dimick. “Variation in hospital mortality associated with inpatient surgery”. In: *New England Journal of Medicine* 361.14 (2009), pp. 1368–1375 (cit. on p. 128).

[Gel+05] Andrew Gelman et al. “Analysis of variance — why it is more important than ever”. In: *The annals of statistics* 33.1 (2005), pp. 1–53 (cit. on p. 96).

[Ger+14] Wulfram Gerstner, Werner M Kistler, Richard Naud, and Liam Paninski. *Neuronal dynamics: From single neurons to networks and models of cognition*. Cambridge University Press, 2014 (cit. on pp. 4, 7).

[GHW79] Gene H. Golub, Michael Heath, and Grace Wahba. “Generalized Cross-Validation as a Method for Choosing a Good Ridge Parameter”. In: *Technometrics* 21.2 (1979), pp. 215–223 (cit. on pp. 84, 89, 121).

[Gia+17] Charles M Giattino, Jacob E Gardner, Faris M Sbahi, et al. “Intraoperative frontal alpha-band power correlates with preoperative neurocognitive function in older adults”. In: *Frontiers in systems neuroscience* 11 (2017), p. 24 (cit. on p. 129).

[Gol+02] Robin I Goldman, John M Stern, Jerome Engel Jr, and Mark S Cohen. “Simultaneous EEG and fMRI of the alpha rhythm”. In: *Neuroreport* 13.18 (2002), p. 2487 (cit. on p. 15).

[Gra+14] Alexandre Gramfort, Martin Luessi, Eric Larson, et al. “MNE software for processing MEG and EEG data”. In: *NeuroImage* 86 (2014), pp. 446–460 (cit. on pp. 25, 30, 48, 84, 85, 91, 96).

[Gro+13] Joachim Gross, Sylvain Baillet, Gareth R Barnes, et al. “Good practice for conducting and reporting MEG research”. In: *Neuroimage* 65 (2013), pp. 349–363 (cit. on p. 88).

[Häm+93] Matti Hämäläinen, Riitta Hari, Risto J Ilmoniemi, Jukka Knuutila, and Olli V Lounasmaa. “Magnetoencephalography—theory, instrumentation, and applications to noninvasive studies of the working human brain”. In: *Reviews of modern Physics* 65.2 (1993), p. 413 (cit. on pp. 10, 56, 57, 61).

[Har+14] A Harati, S Lopez, I Obeid, et al. “The TUH EEG CORPUS: A big data resource for automated EEG interpretation”. In: *2014 IEEE Signal Processing in Medicine and Biology Symposium (SPMB)*. IEEE. 2014, pp. 1–5 (cit. on pp. 25, 92, 93).

[Has+05] Trevor Hastie, Robert Tibshirani, Jerome Friedman, and James Franklin. “The elements of statistical learning: data mining, inference and prediction”. In: *The Mathematical Intelligencer* 27.2 (2005), pp. 83–85 (cit. on p. 39).

[Hau+14] Stefan Haufe, Frank Meinecke, Kai Görgen, et al. “On the interpretation of weight vectors of linear models in multivariate neuroimaging”. In: *NeuroImage* 87 (2014), pp. 96 –110 (cit. on pp. 58, 63, 95, 105).

[Haw+80] RC Hawkes, GN Holland, WS Moore, and BS Worthington. “Nuclear magnetic resonance (NMR) tomography of the brain: a preliminary clinical assessment with demonstration of pathology.” In: *Journal of Computer Assisted Tomography* 4.5 (1980), pp. 577–586 (cit. on p. 9).

[He+19] Tong He, Ru Kong, Avram J. Holmes, et al. “Deep neural networks and kernel regression achieve comparable accuracies for functional connectivity prediction of behavior and demographics”. In: *NeuroImage* (2019), p. 116276 (cit. on p. 63).

[HI84] MS Hämäläinen and RJ Ilmoniemi. *Interpreting magnetic fields of the brain: minimum norm estimates*. Tech. rep. TKK-F-A559. Helsinki University of Technology, 1984 (cit. on p. 62).

[HI94] Matti S Hämäläinen and Risto J Ilmoniemi. “Interpreting magnetic fields of the brain: minimum norm estimates”. In: *Medical & biological engineering & computing* 32.1 (1994), pp. 35–42 (cit. on p. 61).

[HIN09] Kenneth E Hild II and Srikantan S Nagarajan. “Source localization of EEG/MEG data by correlating columns of ICA and lead field matrices”. In: *IEEE Transactions on Biomedical Engineering* 56.11 (2009), pp. 2619–2626 (cit. on p. 63).

[HK70] Arthur E Hoerl and Robert W Kennard. “Ridge regression: Biased estimation for nonorthogonal problems”. In: *Technometrics* 12.1 (1970), pp. 55–67 (cit. on p. 84).

[HKO04] Aapo Hyvärinen, Juha Karhunen, and Erkki Oja. *Independent component analysis*. Vol. 46. John Wiley & Sons, 2004 (cit. on p. 99).

[HLR00] Riitta Hari, Sari Levänen, and Tommi Raij. “Timing of human cortical functions during cognition: role of MEG”. In: *Trends in cognitive sciences* 4.12 (2000), pp. 455–462 (cit. on p. 11).

[HO00] Aapo Hyvärinen and Erkki Oja. “Independent component analysis: algorithms and applications”. In: *Neural networks* 13.4-5 (2000), pp. 411–430 (cit. on pp. 57, 62).

[Hot92] Harold Hotelling. “Relations between two sets of variates”. In: *Breakthroughs in statistics*. Springer, 1992, pp. 162–190 (cit. on p. 64).

[HP17] Riitta Hari and Aina Puce. *MEG-EEG Primer*. Oxford University Press, 2017 (cit. on pp. 4, 7, 10, 12, 15, 62).

[HP18] Hanna-Leena Halme and Lauri Parkkonen. “Across-subject offline decoding of motor imagery from MEG and EEG”. In: *Scientific reports* 8.1 (2018), pp. 1–12 (cit. on p. 85).

[HS14] Olaf Hauk and Matti Stenroos. “A framework for the design of flexible cross-talk functions for spatial filtering of EEG/MEG data: DeFleCT”. In: *Human brain mapping* 35.4 (2014), pp. 1642–1653 (cit. on p. 61).

[HS15] Joerg F Hipp and Markus Siegel. “BOLD fMRI correlation reflects frequency-specific neuronal correlation”. In: *Current Biology* 25.10 (2015), pp. 1368–1374 (cit. on p. 19).

[HSH17] Mehrtash Harandi, Mathieu Salzmann, and Richard Hartley. “Dimensionality reduction on SPD manifolds: The emergence of geometry-aware methods”. In: *IEEE transactions on pattern analysis and machine intelligence* 40.1 (2017), pp. 48–62 (cit. on p. 63).

[HT08] Holger Höfling and Robert Tibshirani. “A study of pre-validation”. In: *The Annals of Applied Statistics* 2.2 (2008), pp. 643–664 (cit. on p. 127).

[HYS16] Inbal Horev, Florian Yger, and Masashi Sugiyama. “Geometry-aware principal component analysis for symmetric positive definite matrices”. In: *Machine Learning* 106 (Nov. 2016) (cit. on p. 63).

[Ino+16] Sharon K Inouye, Edward R Marcantonio, Cyrus M Kosar, et al. “The short-term and long-term relationship between delirium and cognitive trajectory in older surgical patients”. In: *Alzheimer’s & Dementia* 12.7 (2016), pp. 766–775 (cit. on p. 129).

[Jas+17] Mainak Jas, Denis A Engemann, Yousra Bekhti, Federico Raimondo, and Alexandre Gramfort. “Autoreject: Automated artifact rejection for MEG and EEG data”. In: *NeuroImage* 159 (2017), pp. 417–429 (cit. on pp. 89, 93).

[Jas+18] Mainak Jas, Eric Larson, Denis A. Engemann, et al. “A reproducible MEG/EEG group study with the MNE software: Recommendations, quality assessments, and good practices”. In: *Frontiers in Neuroscience* 12.AUG (2018), pp. 1–18 (cit. on pp. 88, 102).

[JCPPV21] C. Touchard J. Cartailler, C. Paquet P. Parutto E. Gayat, and F. Vallee. “Brain fragility among middle-aged and elderly patients from electroencephalogram during induction of anaesthesia”. In: 38 (2021), pp. 1–3 (cit. on p. 129).

[JK17] Eric Jonas and Konrad Paul Kording. “Could a neuroscientist understand a microprocessor?” In: *PLoS computational biology* 13.1 (2017), e1005268 (cit. on p. 15).

[Jón+19] Benedikt Atli Jónsson, Gyda Bjornsdottir, TE Thorgeirsson, et al. “Brain age prediction using deep learning uncovers associated sequence variants”. In: *Nature communications* 10.1 (2019), pp. 1–10 (cit. on p. 110).

[Jou+10] Michel Journée, Francis Bach, P-A Absil, and Rodolphe Sepulchre. “Low-rank optimization on the cone of positive semidefinite matrices”. In: *SIAM Journal on Optimization* 20.5 (2010), pp. 2327–2351 (cit. on p. 54).

[JVV12] Ben Jeuris, Raf Vandebril, and Bart Vandereycken. “A survey and comparison of contemporary algorithms for computing the matrix geometric mean”. In: *Electronic Transactions on Numerical Analysis* 39.ARTICLE (2012), pp. 379–402 (cit. on p. 53).

[KD14] Jean-Rémi King and Stanislas Dehaene. “Characterizing the dynamics of mental representations: the temporal generalization method”. In: *Trends in cognitive sciences* 18.4 (2014), pp. 203–210 (cit. on p. 63).

[Kha+18] Sheraz Khan, Javeria A Hashmi, Fahimeh Mamashli, et al. “Maturation trajectories of cortical resting-state networks depend on the mediating frequency band”. In: *NeuroImage* 174 (2018), pp. 57–68 (cit. on pp. 21, 62, 91, 96).

[Khu+05] Shukri F Khuri, William G Henderson, Ralph G DePalma, et al. “Determinants of long-term survival after major surgery and the adverse effect of postoperative complications”. In: *Annals of surgery* 242.3 (2005), p. 326 (cit. on p. 128).

[Kie+19] Tim C. Kietzmann, Courtney J. Spoerer, Lynn K. A. Sörensen, et al. “Recurrence is required to capture the representational dynamics of the human visual system”. In: *Proceedings of the National Academy of Sciences* (2019). eprint: https://www.pnas.org/content/early/2019/10/04/1905544116.full.pdf (cit. on p. 62).

[Kin+13] Jean-Rémi King, Frédéric Faugeras, Alexandre Gramfort, et al. “Single-trial decoding of auditory novelty responses facilitates the detection of residual consciousness”. In: *Neuroimage* 83 (2013), pp. 726–738 (cit. on p. 63).

[Kin+18] Jean-Rémi King, Laura Gwilliams, Chris Holdgraf, et al. “Encoding and decoding neuronal dynamics: Methodological framework to uncover the algorithms of cognition”. In: (2018) (cit. on p. 63).

[Kob+21] Reinmar J Kobler, Jun-Ichiro Hirayama, Lea Hehenberger, et al. “On the interpretation of linear Riemannian tangent space model parameters in M/EEG”. In: *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE. 2021, pp. 5909–5913 (cit. on p. 95).

[Koc+19] Susanne Koch, Insa Feinkohl, Sourish Chakravarty, et al. “Cognitive impairment is associated with absolute intraoperative frontal α-band power but not with baseline α-band power: a pilot study”. In: *Dementia and geriatric cognitive disorders* 48.1-2 (2019), pp. 83–92 (cit. on p. 129).

[Kol91] Zoltan Joseph Koles. “The quantitative extraction and topographic mapping of the abnormal components in the clinical EEG”. In: *Electroencephalography and clinical Neurophysiology* 79.6 (1991), pp. 440–447 (cit. on p. 64).

[Kre17] Matthias Kreuzer. “EEG based monitoring of general anesthesia: taking the next steps”. In: *Frontiers in computational neuroscience* 11 (2017), p. 56 (cit. on p. 129).

[Kum+19] D. Kumral, F. Sansal, E. Cesnaite, et al. “BOLD and EEG signal variability at rest differently relate to aging in the human brain”. In: *NeuroImage* (2019), p. 116373 (cit. on p. 19).

[Kwo+92] Kenneth K Kwong, John W Belliveau, David A Chesler, et al. “Dynamic magnetic resonance imaging of human brain activity during primary sensory stimulation.” In: *Proceedings of the National Academy of Sciences* 89.12 (1992), pp. 5675–5679 (cit. on p. 10).

[Leb05] Guy Lebanon. “Riemannian Geometry and Statistical Machine Learning”. AAI3159986. PhD thesis. Pittsburgh, PA, USA, 2005 (cit. on p. 29).

[Lie+17] Franziskus Liem, Gaël Varoquaux, Jana Kynast, et al. “Predicting brain-age from multimodal imaging data captures cognitive impairment”. In: *NeuroImage* 148 (2017), pp. 179 –188 (cit. on pp. 18, 96, 103).

[Lin+06] Fa-Hsuan Lin, John W Belliveau, Anders M Dale, and Matti S Hämäläinen. “Distributed current estimates using cortical orientation constraints”. In: *Human brain mapping* 27.1 (2006), pp. 1–13 (cit. on p. 61).

[LNH09] Christoph H Lampert, Hannes Nickisch, and Stefan Harmeling. “Learning to detect unseen object classes by between-class attribute transfer”. In: *2009 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE. 2009, pp. 951–958 (cit. on p. 32).

[Log+01] Nikos K Logothetis, Jon Pauls, Mark Augath, Torsten Trinath, and Axel Oeltermann. “Neurophysiological investigation of the basis of the fMRI signal”. In: *nature* 412.6843 (2001), pp. 150–157 (cit. on p. 10).

[Lot+07] F Lotte, M Congedo, A Lécuyer, F Lamarche, and B Arnaldi. “A review of classification algorithms for EEG-based brain–computer interfaces”. In: *Journal of Neural Engineering* 4.2 (2007), R1–R13 (cit. on p. 102).

[Lot+18] F Lotte, L Bougrain, A Cichocki, et al. “A review of classification algorithms for EEG-based brain–computer interfaces: a 10 year update”. In: *Journal of Neural Engineering* 15.3 (2018), p. 031005 (cit. on pp. 29, 102).

[LP+13] Linda J Larson-Prior, Robert Oostenveld, Stefania Della Penna, et al. “Adding dynamics to the Human Connectome Project with MEG”. In: *Neuroimage* 80 (2013), pp. 190–201 (cit. on p. 89).

[Lut07] Helmut Lutkepohl. *New introduction to multiple time series analysis*. New York City, US: Springer, 2007 (cit. on pp. 40, 42).

[MA18] Estelle Massart and Pierre-Antoine Absil. *Quotient geometry with simple geodesics for the manifold of fixed-rank positive-semidefinite matrices.* Tech. rep. preprint on webpage at http://sites.uclouvain.be/absil/2018.06. UCLouvain, 2018 (cit. on p. 54).

[Mak+96] Scott Makeig, Anthony J Bell, Tzyy-Ping Jung, and Terrence J Sejnowski. "Independent component analysis of electroencephalographic data". In: *Advances in neural information processing systems*. 1996, pp. 145–151 (cit. on p. 57).

[Mak+97] Scott Makeig, Tzyy-Ping Jung, Anthony J Bell, Dara Ghahremani, and Terrence J Sejnowski. "Blind separation of auditory event-related brain responses into independent components". In: *Proceedings of the National Academy of Sciences* 94.20 (1997), pp. 10979–10984 (cit. on p. 62).

[Mar17] Edward R Marcantonio. "Delirium in hospitalized older adults". In: *New England Journal of Medicine* 377.15 (2017), pp. 1456–1466 (cit. on p. 129).

[McR+17] Donald W McRobbie, Elizabeth A Moore, Martin J Graves, and Martin R Prince. *MRI from Picture to Proton*. Cambridge university press, 2017 (cit. on p. 9).

[MLL92] John C Mosher, Paul S Lewis, and Richard M Leahy. "Multiple dipole modeling and localization from spatio-temporal MEG data". In: *IEEE Transactions on Biomedical Engineering* 39.6 (1992), pp. 541–557 (cit. on p. 61).

[MLL99] John C Mosher, Richard M Leahy, and Paul S Lewis. "EEG and MEG: forward solutions for inverse methods". In: *IEEE Transactions on Biomedical Engineering* 46.3 (1999), pp. 245–259 (cit. on p. 61).

[Nää75] R Näätänen. "Selective attention and evoked potentials inhumans—A critical review". In: *Biological Psychology* 2.4 (1975), pp. 237–307 (cit. on p. 14).

[Nen+20] Maximilian Nentwich, Lei Ai, Jens Madsen, et al. "Functional connectivity of EEG is subject-specific, associated with phenotype, and different from fMRI". In: *NeuroImage* 218 (2020), p. 117001 (cit. on p. 19).

[Nep+19] Dmitri Nepogodiev, Janet Martin, Bruce Biccard, et al. "Global burden of postoperative death". In: *The Lancet* 393.10170 (2019), p. 401 (cit. on p. 128).

[NM93] C. L. Nikias and J. M. Mendel. "Signal processing with higher-order spectra". In: *IEEE Signal Processing Magazine* 10.3 (1993), pp. 10–37 (cit. on p. 41).

[NNC11] Vadim V Nikulin, Guido Nolte, and Gabriel Curio. "A novel method for reliable and fast extraction of neuronal EEG/MEG oscillations on the basis of spatio-spectral decomposition". In: *NeuroImage* 55.4 (2011), pp. 1528–1535 (cit. on pp. 64, 74, 107).

[Nol+06] Guido Nolte, Andreas Ziehe, Frank Meinecke, and Klaus-Robert Müller. "Analyzing Coupled Brain Sources: Distinguishing True from Spurious Interaction". In: *Advances in Neural Information Processing Systems 18*. Ed. by Y. Weiss, B. Schölkopf, and J. C. Platt. MIT Press, 2006, pp. 1027–1034 (cit. on p. 60).

[NS05] Paul L. Nunez and Ramesh Srinivasan. *Electric Fields of the Brain - The Neurophysics of EEG*. Oxford University Press, 2005 (cit. on p. 12).

[NVVH07] Rudolf Nieuwenhuys, Jan Voogd, and Christiaan Van Huijzen. *The human central nervous system: a synopsis and atlas*. Springer Science & Business Media, 2007 (cit. on pp. 4–6, 8).

[OKA14] Emanuele Olivetti, Seved Mostafa Kia, and Paolo Avesani. “MEG decoding across subjects”. In: *2014 International Workshop on Pattern Recognition in Neuroimaging*. IEEE. 2014, pp. 1–4 (cit. on p. 85).

[Oos+11] Robert Oostenveld, Pascal Fries, Eric Maris, and Jan-Mathijs Schoffelen. “FieldTrip: open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data”. In: *Computational intelligence and neuroscience* 2011 (2011), p. 1 (cit. on pp. 25, 85).

[OS94] M. Omologo and P. Svaizer. “Acoustic event localization using a crosspower-spectrum phase based technique”. In: *Proceedings of ICASSP ’94. IEEE International Conference on Acoustics, Speech and Signal Processing*. Vol. ii. 1994, II/273–II/276 vol.2 (cit. on p. 40).

[Pal06] Andre Palmini. “The concept of the epileptogenic zone: a modern look at Penfield and Jasper’s views on the role of interictal spikes”. In: *Epileptic disorders* 8.2 (2006), pp. 10–15 (cit. on p. 11).

[Par+05] Lucas C Parra, Clay D Spence, Adam D Gerson, and Paul Sajda. “Recipes for the linear analysis of EEG”. In: *Neuroimage* 28.2 (2005), pp. 326–341 (cit. on p. 63).

[Par+15] Chintan Parmar, Patrick Grossmann, Johan Bussink, Philippe Lambin, and Hugo JWL Aerts. “Machine learning methods for quantitative radiomic biomarkers”. In: *Scientific reports* 5.1 (2015), pp. 1–11 (cit. on p. 16).

[Ped+11] F. Pedregosa, G. Varoquaux, A. Gramfort, et al. “Scikit-learn: Machine Learning in Python”. In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830 (cit. on pp. 14, 25, 30, 84).

[Per+18] Cyril R Pernet, Marta Garrido, Alexandre Gramfort, et al. *Best Practices in Data Analysis and Sharing in Neuroimaging using MEEG*. 2018 (cit. on p. 88).

[Per+19] Cyril R Pernet, Stefan Appelhoff, Krzysztof J Gorgolewski, et al. “EEG-BIDS, an extension to the brain imaging data structure for electroencephalography”. In: *Scientific data* 6.1 (2019), pp. 1–5 (cit. on p. 117).

[PFA06] Xavier Pennec, Pierre Fillard, and Nicholas Ayache. “A Riemannian framework for tensor computing”. In: *International Journal of computer vision* 66.1 (2006), pp. 41–66 (cit. on p. 50).

[Pfu92] Gert Pfurtscheller. “Event-related synchronization (ERS): an electrophysiological correlate of cortical areas at rest”. In: *Electroencephalography and clinical neurophysiology* 83.1 (1992), pp. 62–69 (cit. on p. 15).

[PHV20] Russell A Poldrack, Grace Huckins, and Gael Varoquaux. “Establishment of best practices for evidence for prediction: a review”. In: *JAMA psychiatry* 77.5 (2020), pp. 534–540 (cit. on p. 17).

[PK95] John Polich and Albert Kok. “Cognitive and biological determinants of P300: an integrative review”. In: *Biological psychology* 41.2 (1995), pp. 103–146 (cit. on p. 14).

[PP02] Athanasios Papoulis and S. Unnikrishna Pillai. *Probability, Random Variables, and Stochastic Processes*. Fourth. Boston: McGraw Hill, 2002 (cit. on p. 44).

[Pri+17] Darren Price, Lorraine Komisarjevsky Tyler, R Neto Henriques, et al. “Age-related delay in visual and auditory evoked responses is mediated by white- and grey-matter differences”. In: *Nature communications* 8 (2017), p. 15671 (cit. on p. 21).

[Pri83] M. B. Priestley. *Spectral Analysis and Time Series, Two-Volume Set, Volume 1-2: Volumes I and II*. Academic Press, 1983 (cit. on pp. 42, 44).

[Pur+15a] Patrick L Purdon, KJ Pavone, O Akeju, et al. “The ageing brain: age-dependent changes in the electroencephalogram during propofol and sevoflurane general anaesthesia”. In: *British journal of anaesthesia* 115.suppl_1 (2015), pp. i46–i57 (cit. on p. 129).

[Pur+15b] Patrick L Purdon, Aaron Sampson, Kara J Pavone, and Emery N Brown. “Clinical electroencephalography for anesthesiologists: part I: background and basic signatures”. In: *Anesthesiology* 123.4 (2015), pp. 937–960 (cit. on pp. 111, 113, 123, 129, 134).

[PW93] Donald B. Percival and Andrew T. Walden. *Spectral analysis for physical applications*. Cambridge, US: Cambridge University Press, 1993 (cit. on pp. 42, 43).

[R C19] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria, 2019 (cit. on pp. 25, 84).

[RB15] Pedro L. C. Rodrigues and Luiz A. Baccala. “A new algorithm for neural connectivity estimation of EEG event related potentials”. In: *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, Aug. 2015 (cit. on p. 43).

[RBB15] James A Roberts, Tjeerd W Boonstra, and Michael Breakspear. “The heavy tail of the human brain”. In: *Current opinion in neurobiology* 31 (2015), pp. 164–172 (cit. on p. 64).

[RCJ18] Pedro Luiz Coelho Rodrigues, Marco Congedo, and Christian Jutten. “Multivariate Time-Series Analysis Via Manifold Learning”. In: *2018 IEEE Statistical Signal Processing Workshop (SSP)*. IEEE. 2018, pp. 573–577 (cit. on p. 102).

[Ric+20] Hugo Richard, Luigi Gresele, Aapo Hyvärinen, et al. “Modeling shared responses in neuroimaging studies through multiview ica”. In: *arXiv preprint arXiv:2006.06635* (2020) (cit. on p. 64).

[RJC19] P. L. C. Rodrigues, C. Jutten, and M. Congedo. “Riemannian Procrustes Analysis: Transfer Learning for Brain–Computer Interfaces”. In: *IEEE Transactions on Biomedical Engineering* 66.8 (2019), pp. 2390–2401 (cit. on p. 64).

[Rod+17] Pedro Luiz Coelho Rodrigues, Florent Bouchard, Marco Congedo, and Christian Jutten. “Dimensionality Reduction for BCI classification using Riemannian geometry”. In: *7th Graz Brain-Computer Interface Conference (BCI 2017)*. Gernot R. Müller-Putz. Graz, Austria, 2017 (cit. on p. 46).

[Roy+19] Yannick Roy, Hubert Banville, Isabela Albuquerque, et al. “Deep learning-based electroencephalography analysis: a systematic review”. In: *Journal of Neural Engineering* 16.5 (2019), p. 051001 (cit. on p. 106).

[Sab+19a] David Sabbagh, Pierre Ablin, Gael Varoquaux, Alexandre Gramfort, and Denis A. Engemann. “Manifold-regression to predict from MEG/EEG brain signals without source modeling”. In: *Advances in Neural Information Processing Systems 32*. Ed. by H. Wallach, H. Larochelle, A. Beygelzimer, et al. Curran Associates, Inc., 2019, pp. 7323–7334 (cit. on pp. 16, 21, 24, 25, 29, 44, 82, 84, 102, 110, 121).

[Sab+19b] Krishnakant V Saboo, Yogatheesan Varatharajah, Brent M Berry, et al. “Unsupervised machine-learning classification of electrophysiologically active electrodes during human cognitive task performance”. In: *Scientific reports* 9.1 (2019), pp. 1–14 (cit. on p. 32).

[Sab+20] David Sabbagh, Pierre Ablin, Gaël Varoquaux, Alexandre Gramfort, and Denis A. Engemann. “Predictive regression modeling with MEG/EEG: from source power to signals and cognitive states”. In: *NeuroImage* (2020), p. 116893 (cit. on pp. 22, 24, 25, 32, 82, 110, 121).

[Sac+12] Jane S Saczynski, Edward R Marcantonio, Lien Quach, et al. “Cognitive trajectories after postoperative delirium”. In: *New England Journal of Medicine* 367.1 (2012), pp. 30–39 (cit. on p. 129).

[SC07] Saeid Sanei and J.A. Chambers. *EEG Signal Processing*. John Wiley and Sons Ltd, 2007 (cit. on p. 40).

[Sch+03] G Schneider, AW Gelb, B Schmeller, R Tschakert, and E Kochs. “Detection of awareness in surgical patients with EEG-based indices—bispectral index and patient state index”. In: *British journal of anaesthesia* 91.3 (2003), pp. 329–335 (cit. on p. 110).

[Sch+11] Jan-Mathijs Schoffelen, Jasper Poort, Robert Oostenveld, and Pascal Fries. “Selective movement preparation is subserved by selective increases in corticomuscular gamma-band coherence”. In: *Journal of Neuroscience* 31.18 (2011), pp. 6750–6758 (cit. on p. 85).

[Sch+17] R Schirrmeister, Lukas Gemein, Katharina Eggensperger, Frank Hutter, and Tonio Ball. “Deep learning with convolutional neural networks for decoding and visualization of EEG pathology”. In: *2017 IEEE Signal Processing in Medicine and Biology Symposium (SPMB)*. IEEE. 2017, pp. 1–7 (cit. on pp. 79, 106).

[Sch+19] Marc-Andre Schulz, Thomas Yeo, Joshua Vogelstein, et al. “Deep learning for brains?: Different linear and nonlinear scaling in UK Biobank brain images vs. machine-learning datasets”. In: *bioRxiv* (2019), p. 757054 (cit. on p. 63).

[SG10] Abdulhamit Subasi and M Ismail Gursoy. “EEG signal classification using PCA, ICA, LDA and support vector machines”. In: *Expert systems with applications* 37.12 (2010), pp. 8659–8666 (cit. on p. 62).

[SGM19] Emma Strubell, Ananya Ganesh, and Andrew McCallum. “Energy and Policy Considerations for Deep Learning in NLP”. In: *arXiv preprint arXiv:1906.02243* (2019) (cit. on p. 106).

[Sha+14] Meredith A Shafto, Lorraine K Tyler, Marie Dixon, et al. “The Cambridge Centre for Ageing and Neuroscience (Cam-CAN) study protocol: a cross-sectional, lifespan, multidisciplinary examination of healthy cognitive ageing”. In: *BMC neurology* 14.1 (2014), p. 204 (cit. on pp. 48, 87, 88, 102).

[Sha+20] Yu Raymond Shao, Pegah Kahali, Timothy T Houle, et al. “Low frontal alpha power is associated with the propensity for burst suppression: an electroencephalogram phenotype for a “Vulnerable Brain””. In: *Anesthesia and analgesia* 131.5 (2020), p. 1529 (cit. on p. 129).

[Smi+19] Stephen M. Smith, Diego Vidaurre, Fidel Alfaro-Almagro, Thomas E. Nichols, and Karla L. Miller. “Estimation of brain age delta from brain imaging”. In: *NeuroImage* 200 (2019), pp. 528 –539 (cit. on p. 131).

[SN18] Stephen M Smith and Thomas E Nichols. “Statistical challenges in “big data” human neuroimaging”. In: *Neuron* 97.2 (2018), pp. 263–268 (cit. on p. 83).

[SNS14] Andrew X Stewart, Antje Nuthmann, and Guido Sanguinetti. “Single-trial classification of EEG in a visual object task using ICA and machine learning”. In: *Journal of neuroscience methods* 228 (2014), pp. 1–14 (cit. on p. 62).

[Spi16] David Spiegelhalter. “How old are you, really? Communicating chronic risk through ‘effective age’of your body and organs”. In: *BMC medical informatics and decision making* 16.1 (2016), pp. 1–6 (cit. on p. 17).

[Spr+17] J Sprung, RO Roberts, TN Weingarten, et al. “Postoperative delirium in elderly patients is associated with subsequent cognitive impairment”. In: *BJA: British Journal of Anaesthesia* 119.2 (2017), pp. 316–323 (cit. on p. 129).

[SS+10] Shai Shalev-Shwartz, Ohad Shamir, Nathan Srebro, and Karthik Sridharan. “Learnability, stability and uniform convergence”. In: *The Journal of Machine Learning Research* 11 (2010), pp. 2635–2670 (cit. on p. 37).

[SS65] Eiuc R Skov and David G Simons. “EEG electrodes for in-flight monitoring”. In: *Psychophysiology* 2.2 (1965), pp. 161–167 (cit. on p. 10).

[SSBD14] Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. New York, NY, USA: Cambridge University Press, 2014 (cit. on pp. 13, 31, 33, 37, 38).

[SWS15] Mark G Stokes, Michael J Wolff, and Eelke Spaak. “Decoding rich spatial information with high temporal resolution”. In: *Trends in cognitive sciences* 19.11 (2015), pp. 636–638 (cit. on p. 63).

[Tan+08] Michael W Tangermann, Matthias Krauledat, Konrad Grzeska, et al. “Playing pinball with non-invasive BCI”. In: *Proceedings of the 21st International Conference on Neural Information Processing Systems*. Citeseer. 2008, pp. 1641–1648 (cit. on p. 102).

[Tay+17] Jason R Taylor, Nitin Williams, Rhodri Cusack, et al. “The Cambridge Centre for Ageing and Neuroscience (Cam-CAN) data repository: structural and functional MRI, MEG, and cognitive data from a cross-sectional adult lifespan sample”. In: *Neuroimage* 144 (2017), pp. 262–269 (cit. on pp. 25, 87, 88, 102).

[TB99] Michael E Tipping and Christopher M Bishop. “Probabilistic principal component analysis”. In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 61.3 (1999), pp. 611–622 (cit. on p. 102).

[TBB99] Catherine Tallon-Baudry and Olivier Bertrand. “Oscillatory gamma activity in humans and its role in object representation”. In: *Trends in cognitive sciences* 3.4 (1999), pp. 151–162 (cit. on p. 63).

[Tha+15] Alisa T Thavikulwat, Patrick Lopez, Rafael C Caruso, and Brett G Jeffrey. "The effects of gender and age on the range of the normal human electro-oculogram". In: *Documenta Ophthalmologica* 131.3 (2015), pp. 177–188 (cit. on p. 105).

[TK05] Samu Taulu and Matti Kajola. "Presentation of electromagnetic multichannel data: the signal space separation method". In: *Journal of Applied Physics* 97.12 (2005), p. 124905 (cit. on pp. 46, 47, 88, 90, 99, 102).

[Tou+20] Cyril Touchard, Jérôme Cartailler, Charlotte Levé, et al. "Propofol Requirement and EEG Alpha Band Power During General Anesthesia Provide Complementary Views on Preoperative Cognitive Decline". In: *Frontiers in aging neuroscience* 12 (2020), p. 435 (cit. on pp. 23, 25, 129).

[Tra+00] GM Trapani, Cosimo Altomare, Enrico Sanna, Giovanni Biggio, and Gaetano Liso. "Propofol in anesthesia. Mechanism of action, structure-activity relationships, and drug delivery". In: *Current medicinal chemistry* 7.2 (2000), pp. 249–271 (cit. on p. 135).

[UI97] Mikko A Uusitalo and Risto J Ilmoniemi. "Signal-space projection method for separating MEG or EEG into components". In: *Medical and Biological Engineering and Computing* 35.2 (1997), pp. 135–140 (cit. on pp. 46, 48, 89, 99).

[Un2] *World Population Ageing 2019*. Tech. rep. ST/ESA/SER.A/444. United Nations, Department of Economic and Social Affairs, Population Division (2020) (cit. on pp. 2, 128).

[Var+10] Gaël Varoquaux, Merlin Keller, Jean-Baptiste Poline, Philippe Ciuciu, and Bertrand Thirion. "ICA-based sparse features recovery from fMRI datasets". In: *2010 IEEE International Symposium on Biomedical Imaging: From Nano to Macro*. IEEE. 2010, pp. 1177–1180 (cit. on p. 12).

[Var+17] Gaël Varoquaux, Pradeep Reddy Raamana, Denis A Engemann, et al. "Assessing and tuning brain decoders: cross-validation, caveats, and guidelines". In: *NeuroImage* 145 (2017), pp. 166–179 (cit. on pp. 15, 86).

[VAV09] Bart Vandereycken, P-A Absil, and Stefan Vandewalle. "Embedded geometry of the set of symmetric positive semidefinite matrices of fixed rank". In: *2009 IEEE/SP 15th Workshop on Statistical Signal Processing*. IEEE. 2009, pp. 389–392 (cit. on p. 54).

[VD+08] JA Van Deursen, EFPM Vuurman, FRJ Verhey, VHJM van Kranen-Mastenbroek, and WJ Riedel. "Increased EEG gamma band activity in Alzheimer's disease and mild cognitive impairment". In: *Journal of neural transmission* 115.9 (2008), pp. 1301–1311 (cit. on p. 15).

[Vij+13] Sujith Vijayan, ShiNung Ching, Patrick L Purdon, Emery N Brown, and Nancy J Kopell. "Thalamocortical mechanisms for the anteriorization of alpha rhythms during propofol-induced unconsciousness". In: *Journal of Neuroscience* 33.27 (2013), pp. 11070–11075 (cit. on p. 123).

[Vol+20] Stein Emil Vollset, Emily Goren, Chun-Wei Yuan, et al. "Fertility, mortality, migration, and population scenarios for 195 countries and territories from 2017 to 2100: a forecasting analysis for the Global Burden of Disease Study". In: *The Lancet* 396.10258 (2020), pp. 1285–1306 (cit. on pp. 2, 128).

[Voy+15] Bradley Voytek, Mark A Kramer, John Case, et al. “Age-related changes in 1/f neural electrophysiological noise”. In: *Journal of Neuroscience* 35.38 (2015), pp. 13257–13265 (cit. on pp. 21, 89).

[VS19] Marijn van Vliet and Riitta Salmelin. “Post-hoc modification of linear models: Combining machine learning with domain information to make solid inferences from noisy data”. In: *NeuroImage* (2019), p. 116221 (cit. on p. 63).

[VVB88] Barry D Van Veen and Kevin M Buckley. “Beamforming: A versatile approach to spatial filtering”. In: *IEEE assp magazine* 5.2 (1988), pp. 4–24 (cit. on p. 61).

[War+16] Susan G Wardle, Nikolaus Kriegeskorte, Tijl Grootswagers, Seyed-Mahdi Khaligh-Razavi, and Thomas A Carlson. “Perceptual similarity of visual patterns predicts dynamic neural activation patterns measured with MEG”. In: *Neuroimage* 132 (2016), pp. 59–70 (cit. on p. 63).

[Wes+18] Britta U Westner, Sarang S Dalal, Simon Hanslmayr, and Tobias Staudigl. “Across-subjects classification of stimulus modality from human MEG high frequency activity”. In: *PLoS computational biology* 14.3 (2018), e1005938 (cit. on pp. 62, 85).

[Wic16] Hadley Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016 (cit. on pp. 25, 84).

[Wil+19] Troy S Wildes, Angela M Mickle, Arbi Ben Abdallah, et al. “Effect of electroencephalography-guided anesthetic administration on postoperative delirium among older adults undergoing major surgery: the ENGAGES randomized clinical trial”. In: *Jama* 321.5 (2019), pp. 473–483 (cit. on pp. 111, 129).

[WM09] Yijun Wang and Scott Makeig. “Predicting intended movement direction using EEG from human posterior parietal cortex”. In: *International Conference on Foundations of Augmented Cognition*. Springer. 2009, pp. 437–446 (cit. on p. 62).

[Woo+11] Mark Woolrich, Laurence Hunt, Adrian Groves, and Gareth Barnes. “MEG beamforming using Bayesian PCA for adaptive data covariance matrix regularization”. In: *Neuroimage* 57.4 (2011), pp. 1466–1479 (cit. on p. 102).

[Woo+17] Choong-Wan Woo, Luke J Chang, Martin A Lindquist, and Tor D Wager. “Building better biomarkers: brain models in translational neuroimaging”. In: *Nature neuroscience* 20.3 (2017), p. 365 (cit. on pp. 13–15, 17).

[XGWJ20] Jiachen Xu, Moritz Grosse-Wentrup, and Vinay Jayaram. “Tangent space spatial filters for interpretable and efficient Riemannian classification”. In: *Journal of neural engineering* 17.2 (2020), p. 026043 (cit. on p. 95).

[YBL17] F. Yger, M. Berar, and F. Lotte. “Riemannian Approaches in Brain-Computer Interfaces: A Review”. In: *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 25.10 (2017), pp. 1753–1762 (cit. on pp. 29, 64).

[ZLJZS+19] Yi Zou, Fu-Shan Liu-Jia-Zi Shao, et al. “New nomenclature of peri-operative cognitive impairments: possible impacts on further practice and research”. In: *Chinese medical journal* 132.15 (2019), p. 1859 (cit. on p. 128).